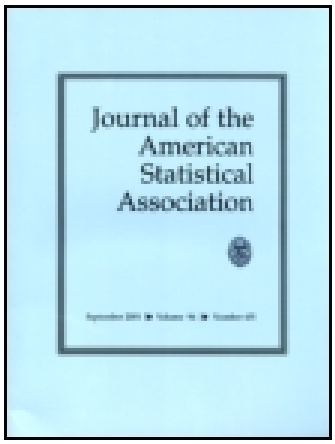

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Nonparametric Regression for Spherical Data

Marco Di Marzio, Agnese Panzera & Charles C. Taylor

Accepted author version posted online: 13 Dec 2013.Published online: 13 Jun 2014.

To cite this article: Marco Di Marzio, Agnese Panzera & Charles C. Taylor (2014) Nonparametric Regression for Spherical Data, Journal of the American Statistical Association, 109:506, 748-763, DOI: 10.1080/01621459.2013.866567

To link to this article: http://dx.doi.org/10.1080/01621459.2013.866567

Nonparametric Regression for Spherical Data

Marco Di MARZIO, Agnese PANZERA, and Charles C. TAYLOR

We develop nonparametric smoothing for regression when both the predictor and the response variables are defined on a sphere of whatever dimension. A local polynomial fitting approach is pursued, which retains all the advantages in terms of rate optimality, interpretability, and ease of implementation widely observed in the standard setting. Our estimates have a multi-output nature, meaning that each coordinate is separately estimated, within a scheme of a regression with a linear response. The main properties include linearity and rotational equivariance. This research has been motivated by the fact that very few models describe this kind of regression. Such current methods are surely not widely employable since they have a parametric nature, and also require the same dimensionality for prediction and response spaces, along with nonrandom design. Our approach does not suffer these limitations. Real-data case studies and simulation experiments are used to illustrate the effectiveness of the method.

KEY WORDS: Confidence sets; Constrained least squares; Geomagnetic field; Local smoothing; Spherical kernels; Tangent normal decomposition; Wind direction.

1. INTRODUCTION

Locations on the surface of a sphere constitute the classical case for spherical data; they are ubiquitous in Earth and planetary sciences. Consider the distribution of volcanoes on the Earth surface, or cosmic microwave background and cosmic ray data distributed on the celestial sphere. In general, the space of all directions in $\mathbb{R}^d$, $d \geq 2$, can be identified with the unit sphere $\mathbb{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}|| = 1\}$. Previously studied directional data include directions of winds, marine currents, Earth's main magnetic field, and rocket nozzle internal combustion flow. Genome sequence representations, text analysis and clustering, morphometrics, and computer vision are fields of recent interest for spherical data; see Hamsici and Martinez (2007).

Dependence when the variables have a spherical nature arises in various fields. In geology, the dependence of one tectonic plate relative to another was studied by Chang et al. (2000); in crystallography, it is of interest to relate an axis of a crystal to an axis of a standard coordinate system (Mackenzie 1957); and in the orientation of a satellite, it is necessary to study dependence between directions of stars and directions in a terrestrial coordinate system (Wahba 1965). Many applications of machine vision require the comparison of directions as detected by two different sensors. An application of spherical regression in quality assurance was given by Chapman, Chen, and Kim (1995). These authors aimed to statistically assess the spatial integrity of geometric objects described through computer-aided design and coordinate measuring machine data. Spherical regression was used in calibration experiments for an electromagnetic motion-tracking system, with the aim of tracking the orientation and position of a sensor moving in three-dimensional space in which the observed orientation is modeled as a rotationally perturbed version of the true one; see Shin, Takahara, and Murdoch (2007).

A unique family of models for pale spherical–spherical regression is available; see Chang (1986). Here, two spherical variates, both lying on $\mathbb{S}^{d-1}$, are related by a rotation. The aim is to estimate and test the unknown rotation matrix. The design variates are fixed, and, to assure uniqueness and consistency of the estimators, must include d mutually orthogonal design directions. Additionally, these models assume circular symmetry for their conditional distributions. In particular, with the constraint that these latter are von Mises–Fisher with constant concentration, a maximum likelihood estimator of the rotation matrix has been derived in a closed form. For the same model, Rivest (1989) discussed asymptotic theory defined by divergence of concentration of data for a fixed sample size. Given the variety of fields of applications, in various cases additional ad hoc hypotheses have been formulated, specific to the scientific context. For example, see the adaptation of the model studied by Chang et al. (2000) for addressing the plate tectonic problem. For the particular case of an ordinary sphere, Downs (2003) implemented spherical parametric regression through link functions based on Möbius transformations. All statistical results and calculation have been formulated in the real domain by the use of a stereographic projection.

In this article, we introduce local polynomial fitting of spherical data. The proposed smoothing is multi-output fashioned in the sense that each coordinate of the response variable—which lies on a sphere—is separately treated. Therefore, up to an asymptotically vanishing normalization task, we decompose the main task into d distinct regression problems, where the predictor lies on a sphere, and the response is linear. As a formal justification for this strategy, we prove that the *joint* estimator, which takes into account the correlation structure, has the same asymptotic efficiency as the one using the *separate* approach under reasonably mild hypotheses. An advantage of this multi-output scheme is that the prediction and response domains do not need to have the same dimensionality.

Because of its centrality in our case, a discussion of the literature on spherical–linear regression is needed. Many papers study regression where a spherical variable predicts a linear

Marco Di Marzio is Associate Professor of Statistics, DMQTE, University of Chieti-Pescara, Italy (E-mail: *mdimarzio@unich.it*). Agnese Panzera is Researcher in Statistics at DiSIA, University of Florence, Italy (E-mail: *agnesepanzera@yahoo.it*). Charles Taylor is Professor of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT (E-mail: *charles@maths.leeds.ac.uk*). We are grateful to an Associate Editor and two anonymous referees for a careful reading of the manuscript and helpful comments, which led to an improved version of this article. We also thank Isabella Fabbri for having read and discussed some parts of the article.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

DOI: 10.1080/01621459.2013.866567

one, the end very often being simple *interpolation*. Most of them come from nonstatistical fields, and the emphasis lies in the application.

A widely used technology is spherical harmonics (see, for a statistical perspective, Abrial et al. 2008), which, however, do not allow reconstruction of efficiently high-resolution signals due to their nonlocal features. On the other hand, spherical splines are a nonparametric interpolation method which initially required equispaced design. However, spherical Bernstein–Bézier splines, introduced by Alfed, Neamtu, and Schumaker (1996), do not require equispacing. A major problem of these latter, however, is that they are not able to reproduce polynomials which have degree smaller or equal to the degree of the space in which they are defined. Furthermore, splines are typically difficult to generalize to higher dimension. An alternate and interesting nonparametric method is regression via *needlets* proposed by Monnier (2011). Here, the approach has a clearly statistical nature. Unfortunately, strong limitations are due to the assumptions requiring a uniform random design and Gaussian noise. Cao et al. (2013) proposed a regularized least-squares algorithm whose output is a finite sum of spherical harmonics. They do not say how to set the tuning parameter.

By comparison with the above methods, our sphere–linear theory would appear competitive, both for providing a new method which is undoubtedly *simple and general* and for featuring a statistical approach as well. Simplicity comes from strong intuitive content, ease of implementation, and explicit formulations. Generality follows from not having serious restrictions concerning: the function to be estimated, the nature of the noise, and the type—random or fixed—of the design.

Finally, some work has been made for the case when a predictor is defined on a Riemannian manifold and the response is linear. Specifically, Pelletier (2006) defined a Nadaraya–Watson-like estimate for this setting in a fully theoretical fashion. Our local constant fit cannot be regarded as a particular case of it since the analysis is extrinsic in nature, for being carried out on tangent spaces via exponential mapping. Also, the kernels, whose argument is a generic distance, are not centered at the observations.

Local polynomial fitting also exists for *unknown* manifold-scalar regression (Bickel and Li 2007; Zhu et al. 2009; Cheng and Wu 2014), and Euclidean-*unknown* manifold regression (Aswani, Bickel, and Tomlin 2011). In principle, these approaches have the potential to be applied to our setup. Although they might be disadvantaged, they are adaptive to the unknown manifold, while the methods in this article are not. However, major differences from our model are that: (a) their resulting estimators have the common feature of not being constructed directly on the manifold, but on tangent spaces, or on the (embedding) Euclidean space; (b) a strong motivation for their use is that the manifold should be much lower dimensional than the embedding space; (c) such methods do not address the case where both the predictor and the response are defined on a manifold different from the Euclidean space; (d) the case of k predictors lying on k distinct manifolds is not treated, whereas the way in which our estimators generalize to this scenario is discussed in Remark 3.1.

In Section 2, we state a series expansion for functions defined on the sphere. In Section 3, we formulate our estimators—including the local constant and local linear fits—for a linear response, which are also used as an intermediate step for the spherical response case. Data-driven bandwidth selection transfers from standard theory to our context without big surprises, however some insights on cross-validation selectors are briefly given in Section 4. Section 5, based on the previous theory, formulates a result for the case when the response lies on a sphere of whatever dimension. We define both the above-mentioned *separate* and *joint* estimators, along with their properties. In Section 6, we prove that our regression estimators satisfy the fundamental property of *rotational equivariance*, which is a kind of robustness to the choice of the coordinate system which implies that the operations of smoothing and rotation are commutative. We next include, in Sections 7 and 8, simulation experiments for the case of sphere–linear regression and sphere–sphere regression, respectively. Finally, in Section 9 we analyze two real datasets. In both, the predictor lies on the ordinary sphere. While in the first example we have a sphere as the response space, in the second one we have a circle. In the second case study, we also construct confidence intervals based on our estimators.

2. PRELIMINARIES

The fact that we will use raw, nontransformed data involves formalizing local smoothing using function expansions and weights that are specific to the sphere. An alternate strategy for the case of Riemannian manifolds prescribes ordinary smoothing on projected data. For example, if the design space was $\mathbb{S}^2$, we could project data onto a suitable tangent plane. In these cases, distortion may arise if the data are spread over a large portion of the sphere and, further, the fitted path is not invariant under changes in the coordinate system. That is, the operations of rotating and smoothing the data are not commutative. Jupp and Kent (1987) provided a detailed critique of various strategies which use Euclidean smoothing for spherical data.

Given $\boldsymbol{x} \in \mathbb{S}^{d-1}$, consider the tangent–normal decomposition for a vector $\boldsymbol{X} \in \mathbb{S}^{d-1}$, that is,

$$\boldsymbol{X} = \boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta), \tag{1}$$

where $\theta \in (0, \pi)$ denotes the angle between $\boldsymbol{x}$ and $\boldsymbol{X}$, and $\boldsymbol{\xi}$ is a vector orthogonal to $\boldsymbol{x}$. Now, setting $\Omega_{\boldsymbol{x}} := \{\boldsymbol{\xi} \in \mathbb{S}^{d-1} : \boldsymbol{\xi} \perp \boldsymbol{x}\}$, for a real-valued function g defined on $\mathbb{S}^{d-1}$ we have

$$\int_{\mathbb{S}^{d-1}} g(\boldsymbol{u})\omega_{d-1}(d\boldsymbol{u}) = \int_0^{\pi} \sin^{d-2}(\theta)d\theta \int_{\Omega_{\boldsymbol{x}}} g\left(\cos(\theta)\boldsymbol{x} + \sin(\theta)\boldsymbol{\xi}\right)\omega_{d-2}(d\boldsymbol{\xi}) \tag{2}$$

where, for each integer $d \geq 1$, $\omega_d(d\boldsymbol{u})$ denotes the area element of $\mathbb{S}^d$, and the total mass of the measure $\omega_d(\cdot)$, interpreted as the surface area of unit sphere, is $\omega_d := \omega_d(\mathbb{S}^d) = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$. Additionally, given a function $g : \mathbb{S}^{d-1} \to \mathbb{R}$, let $\bar{g}(\boldsymbol{x}) := g(\boldsymbol{x}/||\boldsymbol{x}||)$ be the zero-degree homogeneous extension of g to $\mathbb{R}^d \setminus \{\boldsymbol{0}\}$. Provided that $\bar{g}$ has p continuous derivatives in a neighborhood of $\boldsymbol{x}$, we are able to write the pth order Taylor series expansion of g at $\boldsymbol{X}$ belonging to a neighborhood of $\boldsymbol{x}$ on the basis of decomposition (1):

$$g(\boldsymbol{X}) = g(\boldsymbol{x}) + \sum_{s=1}^{p} \frac{\theta^s}{s!}\boldsymbol{\xi}'\mathcal{D}^s_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi}^{\otimes(s-1)} + O(\theta^{p+1}), \tag{3}$$

where $\mathcal{D}^s_{\bar{g}}(\boldsymbol{x})$ is the matrix of the sth-order derivatives of $\bar{g}$ at $\boldsymbol{x}$, and, for a vector $\boldsymbol{u}$, $\boldsymbol{u}'$ denotes its transpose and $\boldsymbol{u}^{\otimes s}$ stands for its sth Kroneckerian power. The above expansion is easily interpretable considering that first-order McLaurin series expansion of $\sin(\theta)$ and $\cos(\theta)$ give $\boldsymbol{X} - \boldsymbol{x} \approx \theta\boldsymbol{\xi}$.

3. SPHERICAL–LINEAR REGRESSION

Given a random vector $(\boldsymbol{X}, Y)$ taking values in $\mathbb{S}^{d-1} \times \mathbb{R}$, assume that the conditional expectation and variance of Y at $\boldsymbol{x} \in \mathbb{S}^{d-1}$, respectively, denoted as $m(\boldsymbol{x})$ and $s^2(\boldsymbol{x})$, are both finite. For a set of independent copies $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, we suppose the model

$$Y_i = m(\boldsymbol{X}_i) + s(\boldsymbol{X}_i)\epsilon_i, \qquad i = 1, \ldots, n,$$

where the ϵ_is are iid real-valued random variables with zero mean, unit variance, and independent of the $\boldsymbol{X}_i$'s. Using the decomposition $\boldsymbol{X}_i = \boldsymbol{x}\cos(\theta_i) + \boldsymbol{\xi}_i \sin(\theta_i)$, assume that the regression m is smooth enough to be approximated through expansion (3). Similarly to the Euclidean approach, we define a pth degree local polynomial estimator of $m(\boldsymbol{x})$ as the solution for β_0 of this weighted least-squares problem

$$\underset{\{\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \sum_{j=1}^{p} \frac{\theta_i^j}{j!} \boldsymbol{\xi}_i' \boldsymbol{\beta}_j \boldsymbol{\xi}_i^{\otimes(j-1)} \right\}^2 K_\kappa(\cos(\theta_i)). \tag{4}$$

The weight (or *kernel*) K_κ is a unimodal density defined on $\mathbb{S}^{d-1}$ with rotational symmetry about its mean direction $\boldsymbol{\mu} = (0, \ldots, 0, 1)$, and concentration parameter $\kappa \in (0, \infty)$ such that $\lim_{\kappa\to\infty} \int_W K_\kappa(\boldsymbol{x}'\boldsymbol{\mu})\omega_{d-1}(d\boldsymbol{x}) = 0$, for any $W \subset \mathbb{S}^{d-1} \setminus \{\boldsymbol{\mu}\}$. Kernels of this form were used by Hall, Watson, and Cabrera (1987) for density estimation on the sphere. In our context, they implemented the locality of the method, emphasizing observations which are closer to the estimation point. For given sample data, the local feature of the estimate increases with the magnitude of κ, meaning that finer local structures arise in correspondence of an increase in concentration. Observe that here κ is typically *not* a scale factor, and this will affect most of our technical treatment. Importantly, in the subsequent asymptotic theory we will always implicitly assume that κ increases with n, whereas the bandwidth of Euclidean weights needs to decrease to obtain concentration. Finally, note that whatever the dimension is, spherical kernels are always equipped with a scalar smoothing parameter. This constitutes a remarkable difference with the Euclidean setting, where, for the case of an $\mathbb{R}^d$-valued predictor, we have the possibility to select up to $d(d+1)/2$ distinct smoothing parameters.

Remark 3.1. The case of multiple predictors requires expansions over product spaces along with product kernels to be employed in problem (4). Clearly, if some predictors are linear this scheme still holds. A case study for this latter setting, due to Jeon and Taylor (2012), models wind power in terms of wind speed, say X, and wind direction, say θ. The estimator is based on the kernel regression idea. However, to circumvent the task of addressing angular variables, they build somewhat artificial covariates given by $X\sin\theta$ and $X\cos\theta$, with a potential covariance issue. A natural alternative could be constructed by our regression scheme with a natural weight given by the product between a spherical kernel (defined on $\mathbb{S}^1$) and a Euclidean one.

Finally, we present some key quantities that are function of spherical kernels. For $j \in \mathbb{N}$, let

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

Clearly $b_j(\kappa)$ is reminiscent of the jth moment of a Euclidean kernel, and $\nu_0(\kappa)$ recalls its roughness. We will see that quantities $b_2(\kappa)$ and $\nu_0(\kappa)$ reflect, in turn, the asymptotic bias and variance of our smoothers.

3.1 Local Constant Estimator

Setting $p = 0$ in (4) leads to a local constant fit, that is,

$$\hat{m}(\boldsymbol{x}; 0) = \frac{\sum_{i=1}^n K_\kappa(\cos(\theta_i))Y_i}{\sum_{i=1}^n K_\kappa(\cos(\theta_i))}. \tag{5}$$

For the above estimator, denoting the design density as f, we get

Theorem 3.1. Given the random sample $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, taking values in $\mathbb{S}^{d-1} \times \mathbb{R}$, if

(i) $f(\boldsymbol{x}) > 0$, f, s^2 and all entries of $\mathcal{D}_{\bar{f}}$, $\mathcal{D}_{\bar{m}}$, and $\mathcal{D}^2_{\bar{m}}$ are continuous at $\boldsymbol{x} \in \mathbb{S}^{d-1}$,

(ii) K_κ is such that, for each $j \in \mathbb{Z}^+$, $\lim_{\kappa\to\infty}(1 - c_j(\kappa))/(1 - c_1(\kappa)) = j$, where $c_j(\kappa) := \omega_{d-2}\int_0^\pi K_\kappa(\cos(\theta))\cos(\theta)^j \sin^{d-2}(\theta)d\theta$,

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

(iv) $\lim_{n\to\infty} n^{-1}\nu_0(\kappa) = 0$,

then, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$

$$E[\hat{m}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] - m(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)}\left\{ \operatorname{tr}\left(\mathcal{D}^2_{\bar{m}}(\boldsymbol{x})\right) + \frac{2\mathcal{D}'_{\bar{f}}(\boldsymbol{x})\mathcal{D}_{\bar{m}}(\boldsymbol{x})}{f(\boldsymbol{x})} \right\} + o_p\left(b_2(\kappa)\right), \tag{6}$$

where $\operatorname{tr}(\mathbb{A})$ stands for the trace of the matrix $\mathbb{A}$, and

$$\operatorname{var}[\hat{m}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = \frac{s^2(\boldsymbol{x})\nu_0(\kappa)}{nf(\boldsymbol{x})} + o_p\left(\frac{\nu_0(\kappa)}{n}\right). \tag{7}$$

Moreover, it holds that

$$\left\{ \frac{\hat{m}(\boldsymbol{x}; 0) - m(\boldsymbol{x})}{\psi^{1/2}(\boldsymbol{x})} \right\} \xrightarrow{d} N(\tau(\boldsymbol{x}; 0), 1),$$

where $\tau(\boldsymbol{x}; 0)$ and $\psi(\boldsymbol{x})$ are the leading terms of the right-hand side in (6) and (7), respectively.

Proof. See the Appendix. □

After stating that, differently from the standard case, higher order terms in expansions involved in asymptotic bias and variance calculations do not necessarily decrease with their order, in the following remark we will explain the idea behind all the approximations appearing in our asymptotic theory.

Remark 3.2. First, consider that only small values of θ will be relevant for our calculations because, if κ increases with n,

then the kernel concentrates around its mean direction. Now, since for θ approaching to 0, $\theta^j \sim 2^{j/2}\{1-\cos(\theta)\}^{j/2}$, $j \in \mathbb{N}$, a simple approximation requiring a big enough κ and even j is

$$b_j(\kappa) \sim 2^{j/2}\left\{1+\sum_{s=1}^{j/2}(-1)^s\binom{j/2}{s}c_s(\kappa)\right\}. \tag{8}$$

Also consider that in the expansion of convolutions we will encounter, the quantity $b_j(\kappa)$ turns out to be multiplied by $\int_{\Omega_x}\boldsymbol{\xi}\boldsymbol{\xi}'^{\otimes(j-1)}\omega_{d-2}(d\boldsymbol{\xi})$, which is null for odd j. Now approximation (8) makes apparent that assumption (ii) of Theorem 3.1 implies that even-order terms vanish faster than the second one. This reproduces a Euclidean-like scenario, where terms of even order $j > 2$ are $o(h^2)$, and odd terms vanish by the symmetry of the kernel, and assures that leading terms can be identified and used for approximations.

An optimal smoothing degree would minimize the conditional asymptotic mean-squared error of $\hat{m}(\boldsymbol{x};0)$, say AMSE$[\hat{m}(\boldsymbol{x};0) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]$, which is the sum of the leading terms of the conditional squared bias and conditional variance. Note that the dependence of conditional bias and variance on the smoothing factor cannot be generalized with respect to the kernel, because it is not a scale family. For the important case of a Langevin kernel, which can be regarded as the spherical counterpart of the Gaussian kernel, and, on $\mathbb{S}^{d-1}$, is defined by $\kappa^{d/2-1}\{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa)\}^{-1}e^{(\kappa\cos(\theta))}$, where $\mathcal{I}_u(\cdot)$ stands for the modified Bessel function of the first kind and order u, it holds that for κ big enough, and $j \in \mathbb{Z}^+$

$$b_j(\kappa) \sim \frac{2^{j/2}\Gamma\left((d+j-1)/2\right)}{\kappa^{j/2}\Gamma\left((d-1)/2\right)}, \quad \text{and} \quad \nu_0(\kappa) \sim \frac{\kappa^{(d-1)/2}}{2^{d-1}\pi^{(d-1)/2}}. \tag{9}$$

Hence, this kernel satisfies condition (ii) in Theorem 3.1, whereas assumptions (iii) and (iv) imply that, as n diverges, $\kappa \to \infty$ and $n^{-1}\kappa^{(d-1)/2} \to 0$, respectively. Therefore, when the Langevin kernel is used

$$\begin{aligned}&\text{AMSE}[\hat{m}(\boldsymbol{x};0) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]\\ &\quad=\frac{1}{4\kappa^2}\left\{\text{tr}\left(\mathcal{D}^2_{\tilde{m}}(\boldsymbol{x})\right)+\frac{2\mathcal{D}'_{\tilde{f}}(\boldsymbol{x})\mathcal{D}_{\tilde{m}}(\boldsymbol{x})}{f(\boldsymbol{x})}\right\}^2\\ &\qquad+\frac{\kappa^{(d-1)/2}s^2(\boldsymbol{x})}{2^{d-1}\pi^{(d-1)/2}nf(\boldsymbol{x})},\end{aligned}$$

and, thus, the value of κ minimizing AMSE$[\hat{m}(\boldsymbol{x};0) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]$ is

$$\left\{\frac{2^{d-1}\pi^{(d-1)/2}nf(\boldsymbol{x})\{\text{tr}(\mathcal{D}^2_{\tilde{m}}(\boldsymbol{x}))+2\mathcal{D}'_{\tilde{f}}(\boldsymbol{x})\mathcal{D}_{\tilde{m}}(\boldsymbol{x})f(\boldsymbol{x})^{-1}\}^2}{(d-1)s^2(\boldsymbol{x})}\right\}^{2/(d+3)},$$

and $\hat{m}(\boldsymbol{x};0)$ enjoys the convergence rate $n^{-4/(3+d)}$, which is the same as the Nadaraya–Watson one when a second-order kernel is used and m has domain in $\mathbb{R}^{d-1}$.

3.2 Local Linear Estimator

Let $\boldsymbol{Y} := [Y_1 \ldots Y_n]'$, $\mathbb{W} := \text{diag}[K_\kappa(\cos(\theta_1)),\ldots,K_\kappa(\cos(\theta_n))]$, $\boldsymbol{\beta} := [\beta_0 \;\; \boldsymbol{\beta}_1']'$, and set

$$\mathbb{X} := \begin{bmatrix}1 & \theta_1\boldsymbol{\xi}_1'\\ \vdots & \vdots\\ 1 & \theta_n\boldsymbol{\xi}_n'\end{bmatrix}.$$

Then, the loss in problem (4), for $p=1$, can be rewritten as $||\mathbb{W}^{1/2}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})||^2$, and its minimization over $\boldsymbol{\beta}$ admits a unique solution if and only if $\mathbb{X}'\mathbb{W}\mathbb{X}$ is nonsingular. Unfortunately, in our setting the Euclidean condition for invertibility, that is, that at least $p+1$ weights are positive at the estimation point, is not sufficient. In fact, our least squares solution will not be unique since $\boldsymbol{\xi}_i \perp \boldsymbol{x}$, for $i \in \{1,\ldots,n\}$, and so $\mathbb{X}\boldsymbol{Q}_1 = \boldsymbol{0}_n$, where $\boldsymbol{Q}_1 = [0 \;\; \boldsymbol{x}']'$, and, for a positive integer u, $\boldsymbol{0}_u$ stands for a $u \times 1$ zero vector. Letting $\mathbb{A} := \mathbb{X}'\mathbb{W}\mathbb{X}$, denote by $\mathcal{R}(\mathbb{A}')$ and $\mathcal{N}(\mathbb{A})$ the space spanned by the columns of $\mathbb{A}'$ and the null space of $\mathbb{A}$ respectively, with $\mathcal{R}(\mathbb{A}') \perp \mathcal{N}(\mathbb{A})$. We see that $\mathcal{N}(\mathbb{A})$ is determined by the vector $\boldsymbol{Q}_1$. The weighted least squares solution is determined by the set $\{\mathbb{A}^+\mathbb{X}'\mathbb{W}\boldsymbol{Y}+\boldsymbol{y}, \boldsymbol{y} \in \mathcal{N}(\mathbb{A})\} = \{\mathbb{A}^+\mathbb{X}'\mathbb{W}\boldsymbol{Y}+c\boldsymbol{Q}_1, c \in \mathbb{R}\}$, where $\mathbb{A}^+$ denotes the Moore–Penrose pseudoinverse of $\mathbb{A}$. However, since $\boldsymbol{x}'\mathcal{D}_{\tilde{m}}(\boldsymbol{x}) = 0$, we have a further requirement for the solution to satisfy $\boldsymbol{Q}_1'\boldsymbol{\beta} = 0$. Hence, we obtain a unique solution belonging to $\mathcal{R}(\mathbb{A}')$, that is, $\hat{\boldsymbol{\beta}} = \mathbb{A}^+\mathbb{X}'\mathbb{W}\boldsymbol{Y}$.

We now obtain an explicit form of this solution, using constrained least squares, to derive its properties. Define a local linear estimator for m at $\boldsymbol{x}$ as the first entry of the solution for $\boldsymbol{\beta}$ of

$$\min_{\boldsymbol{\beta}} ||\mathbb{W}^{1/2}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})||^2 \quad \text{subject to} \quad \boldsymbol{Q}_1'\boldsymbol{\beta} = 0. \tag{10}$$

Now, let $\boldsymbol{Q}_2$ be a $(d+1)\times d$ matrix such that $\boldsymbol{Q}_2'\boldsymbol{Q}_1 = \boldsymbol{0}_d$, and the matrix $[\boldsymbol{Q}_1 \;\; \boldsymbol{Q}_2]$ is nonsingular. Then, letting $\boldsymbol{Q} := [\boldsymbol{Q}_1 \;\; \boldsymbol{Q}_2]'$, $\boldsymbol{A}_i := \mathbb{X}\boldsymbol{Q}_i(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}, i \in \{1,2\}$, and $\boldsymbol{z} := \boldsymbol{Q}_2'\boldsymbol{\beta}$, we have

$$\mathbb{X}\boldsymbol{\beta} = (\mathbb{X}\boldsymbol{Q}^{-1})(\boldsymbol{Q}\boldsymbol{\beta}) = [\boldsymbol{A}_1 \;\; \boldsymbol{A}_2]\begin{bmatrix}\boldsymbol{Q}_1'\boldsymbol{\beta}\\ \boldsymbol{z}\end{bmatrix} = \boldsymbol{A}_2\boldsymbol{z}. \tag{11}$$

Then, problem (10) reduces to $\min_{\boldsymbol{z}} ||\mathbb{W}^{1/2}(\boldsymbol{Y}-\boldsymbol{A}_2\boldsymbol{z})||^2$, which gives $\hat{\boldsymbol{z}} = (\boldsymbol{A}_2'\mathbb{W}\boldsymbol{A}_2)^{-1}\boldsymbol{A}_2'\mathbb{W}\boldsymbol{Y}$, and

$$\begin{aligned}\hat{m}(\boldsymbol{x};1) := \boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\boldsymbol{Q}_2)^{-1}\hat{\boldsymbol{z}} = \boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\\ \times \boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\boldsymbol{Y},\end{aligned} \tag{12}$$

where $\boldsymbol{e}_1 := [1 \;\; \boldsymbol{0}_d']'$. Note that to estimate m we constrain its derivatives, and this could appear somewhat artificial, given that m is separately estimated from its derivatives in local polynomial fitting. In principle, various settings of $\boldsymbol{Q}_2$ are possible in $\boldsymbol{Q}_2(\boldsymbol{Q}_2'\boldsymbol{Q}_2)^{-1}\hat{\boldsymbol{z}}$. We see this also by noting that the solution for $\hat{m}(\boldsymbol{x};1)$ uses only the first element of $\boldsymbol{\beta}$, and this will not be affected by alternative choices of generalized inverse, since the first element of $\boldsymbol{Q}_1$ is zero. The same conclusion can be algebraically obtained observing that, since for a nonsingular matrix $\boldsymbol{A}$, $\boldsymbol{A}^{-1} = \boldsymbol{A}^+$, and for a full-rank factorization $\boldsymbol{A} = \boldsymbol{F}\boldsymbol{G}$,

$\boldsymbol{A}^+ = \boldsymbol{G}^+\boldsymbol{F}^+$, it results

$$\begin{aligned}\boldsymbol{Q}_2(\boldsymbol{Q}_2'\boldsymbol{Q}_2)^{-1}\hat{\boldsymbol{z}} &= \boldsymbol{Q}_2(\mathbb{W}^{1/2}\mathbb{X}\boldsymbol{Q}_2)^+\mathbb{W}^{1/2}\boldsymbol{Y}\\ &= \boldsymbol{Q}_2\boldsymbol{Q}_2^+\mathbb{A}^+\mathbb{X}'\mathbb{W}\boldsymbol{Y},\end{aligned}$$

where $\boldsymbol{Q}_2\boldsymbol{Q}_2^+$ defines the orthogonal projector onto the space of the vectors orthogonal to $\boldsymbol{Q}_1$, that is, onto $\mathcal{R}(\mathbb{A}')$, and $\mathbb{A}^+\mathbb{X}'\mathbb{W}\boldsymbol{Y} \in \mathcal{R}(\mathbb{A}')$. Hence, from now on, we set, without loss of generality, $\boldsymbol{Q}_2 := [\boldsymbol{x} \quad \boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}']'$, where, for a positive integer u, $\boldsymbol{I}_u$ stands for the identity matrix of order u.

Linearity is easily seen. By the choice of $\boldsymbol{Q}_2$, using the orthogonality between $\boldsymbol{x}$ and the $\boldsymbol{\xi}_i$s, we have $\mathbb{X}\boldsymbol{Q}_2 = [\boldsymbol{x} + \theta_1\boldsymbol{\xi}_1 \ldots \boldsymbol{x} + \theta_n\boldsymbol{\xi}_n]'$, and then estimator (12) can be written as $\sum_{i=1}^n W_i Y_i$, where

$$W_i = \boldsymbol{x}'\left\{\sum_{j=1}^n K_\kappa(\cos(\theta_j))(\boldsymbol{x}+\theta_j\boldsymbol{\xi}_j)(\boldsymbol{x}'+\theta_j\boldsymbol{\xi}_j')\right\}^{-1} \times (\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)K_\kappa(\cos(\theta_i)).$$

Now, for estimator (12) we obtain the following.

Theorem 3.2. Let $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$ be a random sample taking values in $\mathbb{S}^{d-1} \times \mathbb{R}$. If assumptions (i)–(iv) of Theorem 3.1 hold then, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$

$$\begin{aligned}&E[\hat{m}(\boldsymbol{x};1) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] - m(\boldsymbol{x})\\ &\quad = \frac{b_2(\kappa)\mathrm{tr}(\mathcal{D}^2_{\bar{m}}(\boldsymbol{x}))}{2(d-1)} + o_p(b_2(\kappa)), \qquad (13)\\ &\mathrm{var}[\hat{m}(\boldsymbol{x};1) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]\\ &\quad = \frac{\nu_0(\kappa)s^2(\boldsymbol{x})}{nf(\boldsymbol{x})} + o_p\left(\frac{\nu_0(\kappa)}{n}\right). \qquad (14)\end{aligned}$$

Additionally, it holds that

$$\left\{\frac{\hat{m}(\boldsymbol{x};1) - m(\boldsymbol{x})}{\psi^{1/2}(\boldsymbol{x})}\right\} \xrightarrow{d} N(\tau(\boldsymbol{x};1), 1),$$

where $\tau(\boldsymbol{x};1)$ and $\psi(\boldsymbol{x})$ stand for the leading terms of the right-hand side in (13) and (14), respectively.

Proof. See the Appendix. □

The accuracy quantities in Theorem 3.2 have the same structure as the Euclidean ones, and consequently: (a) the bias of the local linear fit improves on the local-constant one for being both design adaptive (it does not depend on the design density), and unaffected by boundary bias (it does not depend on the first derivative); (b) for the Langevin kernel the optimal convergence rate is $n^{-4/(3+d)}$ (to see this use Theorem 3.2 along with approximations (9)); (c) high minimax efficiency among linear smoothers as defined by Fan (1993).

For the special case when the predictor is defined on the *circle*, a distinct nonparametric regression was proposed by Di Marzio, Panzera, and Taylor (2009). They used a Taylor series-like expansion which is different from our tangential–normal one. Consequently, their estimator does not have the same expression as ours in formula (12), and is only generalizable to the torus case, not to the sphere. The two estimators, however, have the same rates of convergence of both asymptotic bias and variance.

The solution for $\boldsymbol{\beta}_1$ of problem (10) leads to a local estimator for partial derivatives $\mathcal{D}_{\bar{m}}(\boldsymbol{x})$, that is,

$$\widehat{\mathcal{D}}_{\bar{m}}(\boldsymbol{x}) := \boldsymbol{e}_2'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\boldsymbol{Y}, \qquad (15)$$

where $\boldsymbol{e}_2 := [\boldsymbol{0}_d' \quad \boldsymbol{I}_d]'$. Hence, under the assumptions of Theorem 3.2, and assuming that all entries of $\mathcal{D}^3_{\bar{m}}$ are continuous at $\boldsymbol{x}$, similar arguments as those used in the proof of Theorem 3.2 yield, for $\ell \in \{1, \ldots, d\}$,

$$\begin{aligned}&\mathrm{AMSE}\big[\widehat{\mathcal{D}}^{(\ell)}_{\bar{m}}(\boldsymbol{x}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\big]\\ &= \left\{\frac{(d-1)^2 b_4(\kappa)\boldsymbol{t}_2^{(\ell)}(\boldsymbol{x}) - b_2^2(\kappa)\boldsymbol{t}_1^{(\ell)}(\boldsymbol{x})}{2(d-1)b_2(\kappa)f(\boldsymbol{x})} + \frac{(d-1)b_4(\kappa)\boldsymbol{t}_3^{(\ell)}(\boldsymbol{x})}{3!b_2(\kappa)}\right\}^2\\ &\quad + \{1 - (\boldsymbol{x}^{(\ell)})^2\}\frac{\nu_2(\kappa)(d-1)s^2(\boldsymbol{x})}{nb_2^2(\kappa)f(\boldsymbol{x})},\end{aligned}$$

where $\boldsymbol{a}^{(\ell)}$ stands for the ℓth entry of the vector $\boldsymbol{a}$, and

$$\begin{aligned}\boldsymbol{t}_1(\boldsymbol{x}) &:= \mathrm{tr}\left(\mathcal{D}^2_{\bar{m}}(\boldsymbol{x})\right)\mathcal{D}_{\bar{f}}(\boldsymbol{x}),\\ \boldsymbol{t}_2(\boldsymbol{x}) &:= \int_{\Omega_x} \boldsymbol{\xi}\boldsymbol{\xi}'\mathcal{D}^2_{\bar{m}}(\boldsymbol{x})\boldsymbol{\xi}\boldsymbol{\xi}'\mathcal{D}_{\bar{f}}(\boldsymbol{x})\omega_{d-2}(d\boldsymbol{\xi}),\\ \boldsymbol{t}_3(\boldsymbol{x}) &:= \int_{\Omega_x} \boldsymbol{\xi}\boldsymbol{\xi}'\mathcal{D}^3_{\bar{m}}(\boldsymbol{x})\boldsymbol{\xi}\boldsymbol{\xi}'\omega_{d-2}(d\boldsymbol{\xi}).\end{aligned}$$

Hence, for the case of a Langevin kernel, using approximation (9), along with the asymptotic approximation $\nu_2(\kappa) \sim k^{(d-3)/2}(d-1)\{\pi^{(d-1)/2}2^d\}^{-1}$, we see that the value of κ minimizing $\mathrm{AMSE}[\widehat{\mathcal{D}}^{(\ell)}_{\bar{m}}(\boldsymbol{x}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$ attains the optimal rate of $n^{-2/(5+d)}$, which is the same as that achieved when a local linear estimator, with a second-order kernel, is used to estimate the derivative of a regression function with domain in $\mathbb{R}^{d-1}$.

3.3 Data From Mixing Processes

When the sampled data are generated by stationary mixing processes, under suitable conditions, the rate of convergence of our estimators is the same as in the iid case as stated in the following.

Theorem 3.3. Let $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$ be a random sequence from the stationary process $\{(\boldsymbol{X}_i, Y_i), i \in \mathbb{Z}^+\}$. Under assumptions (i)–(iv) of Theorem 3.1, and assuming that

(i) for $\lambda \geq 2$, $\gamma_d \in \mathbb{R}^+$, and any integer $d > 1$, $\int_0^\pi \omega_{d-2}|K_\kappa(\cos(\theta))|^\lambda \sin^{d-2}(\theta)d\theta = O(\kappa^{\gamma_d(\lambda-1)})$;
(ii) for real constants C_1 and C_2, and all $\ell > 1$, the joint density of $\boldsymbol{X}_1$ and $\boldsymbol{X}_\ell$, say $g_{\boldsymbol{X}_1,\boldsymbol{X}_\ell}$, satisfies $g_{\boldsymbol{X}_1,\boldsymbol{X}_\ell}(u, v) \leq C_1$, and $E[Y_1^2 + Y_\ell^2 | \boldsymbol{X}_1, \boldsymbol{X}_\ell] \leq C_2$;
(iii) the process $(\boldsymbol{X}_i, Y_i)$, is either α-mixing with $\sum_\ell \ell^a[\alpha(\ell)]^{1-2/\lambda} < \infty$ and $E[|Y_i|^\lambda \mid \boldsymbol{X}_i] \leq C_3 < \infty$ for $\lambda > 2$, $a > 1 - 2/\lambda$, or ρ-mixing with $\sum_\ell \rho(\ell) < \infty$;

then, at $\boldsymbol{x} \in \mathbb{S}^{d-1}$, the asymptotic bias of $\hat{m}(\boldsymbol{x}; p)$, $p \in \{0, 1\}$, is the same as the iid case, and

$$\mathrm{var}[\hat{m}(\boldsymbol{x};p) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = \frac{\nu_0(\kappa)s^2(\boldsymbol{x})}{nf(\boldsymbol{x})} + o\left(\frac{\kappa^{\gamma_d}}{n}\right).$$

Proof. See the Appendix. □

This result, when $Y_i = \boldsymbol{X}_{i+1}^{(\ell)}$, allows the use of our estimators in autoregression estimation.

4. CROSS-VALIDATION SMOOTHING

Implementing cross-validation for smoothing selection is straightforward. Letting $\hat{m}^{(-i)}$ denote the estimator using the sample with $(\boldsymbol{X}_i, Y_i)$ left out, then κ can be set to minimize $\mathrm{CV}(\kappa) := 1/n \sum_{i=1}^{n} \{Y_i - \hat{m}^{(-i)}(X_i; p)\}^2$, where $p \in \{0, 1\}$. In our case, we have the same as in the Euclidean setting, that is:

$$E[\mathrm{CV}(\kappa)] = \int_{\mathbb{S}^{d-1}} \{\hat{m}(\boldsymbol{x}; p) - m(\boldsymbol{x})\}^2 f(\boldsymbol{x}) d\boldsymbol{x} + \int_{\mathbb{S}^{d-1}} s^2(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x},$$

consequently, minimizing $\mathrm{CV}(\kappa)$ is expected to be equivalent to minimizing the L_2 risk.

Concerning asymptotic properties, Härdle and Marron (1985, Theorem 1) proved that cross-validation is asymptotically optimal with respect to various L_2 norms. Their proof applies also in our case, provided that their assumption (A.1) is replaced by $\kappa > Cn^{\delta}$, where C and δ are positive constants.

5. SPHERICAL–SPHERICAL REGRESSION

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a $\mathbb{S}^{d-1} \times \mathbb{S}^{q-1}$-valued random vector, and let $Y^{(\ell)}$ be the ℓth Cartesian coordinate of $\boldsymbol{Y}$. Setting $m_\ell(\boldsymbol{x}) := E[Y^{(\ell)} \mid \boldsymbol{X} = \boldsymbol{x}]$, the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ could be modeled by the function $\boldsymbol{m} : \mathbb{S}^{d-1} \to \mathbb{S}^{q-1}$ minimizing the risk $E[||\boldsymbol{Y} - m(\boldsymbol{X})||^2 \mid \boldsymbol{X}]$ subject to $||m(\boldsymbol{X})|| = 1$, which, at $\boldsymbol{X} = \boldsymbol{x}$, is given by

$$\boldsymbol{m}(\boldsymbol{x}) := [m_1(\boldsymbol{x}) \dots m_q(\boldsymbol{x})]' ||[m_1(\boldsymbol{x}) \dots m_q(\boldsymbol{x})]||^{-1}.$$

Given the random sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \dots, n\}$, we assume the model

$$\boldsymbol{Y}_i = \boldsymbol{m}(\boldsymbol{X}_i) + \boldsymbol{\epsilon}_i, \qquad i = 1, \dots, n,$$

where the errors $\boldsymbol{\epsilon}_i := [\epsilon_i^{(1)} \dots \epsilon_i^{(q)}]'$, conditioned on the $\boldsymbol{X}_i$'s, are independent with $E[\boldsymbol{\epsilon}_i \mid \boldsymbol{X}_i] = \boldsymbol{0}_q$ and $\mathrm{var}[\boldsymbol{\epsilon}_i \mid \boldsymbol{X}_i] = \mathbb{\Sigma}(\boldsymbol{X}_i)$, where $\mathbb{\Sigma}(\boldsymbol{X}_i)$ stands for the matrix of order q having $s_\ell^2(\boldsymbol{X}_i) := \mathrm{var}[\epsilon_i^{(\ell)} \mid \boldsymbol{X}_i] < \infty$, as (ℓ, ℓ)th entry and $s_{\ell,j}(\boldsymbol{X}_i) := \mathrm{cov}[\epsilon_i^{(\ell)}, \epsilon_i^{(j)} \mid \boldsymbol{X}_i] < \infty$ as (ℓ, j)th for $(\ell, j) \in \{1, \dots, q\} \times \{1, \dots, q\}$, and $\ell \neq j$. The above model can be also written as q separate (but correlated) regression models, that is,

$$Y_i^{(\ell)} = m_\ell(\boldsymbol{X}_i) + \boldsymbol{\epsilon}_i^{(\ell)}, \quad \ell = 1, \dots, q, \quad i = 1, \dots, n.$$

5.1 Local Constant Estimation

If θ_i is the angle between $\boldsymbol{x}$ and $\boldsymbol{X}_i$, a local constant estimator of $\boldsymbol{m}(\boldsymbol{x})$, say $\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)$, is

$$\arg\min_{\boldsymbol{\beta}_0} \sum_{i=1}^{n} ||\boldsymbol{Y}_i - \boldsymbol{\beta}_0||^2 K_\kappa(\cos(\theta_i)), \quad \text{subject to} \quad ||\boldsymbol{\beta}_0|| = 1.$$

Specifically, letting $\hat{m}_\ell(\boldsymbol{x}; 0) := \{\sum_{i=1}^{n} K_\kappa(\cos(\theta_i))\}^{-1} \sum_{i=1}^{n} K_\kappa(\cos(\theta_i)) Y_i^{(\ell)}$, we have

$$\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) = ||[\hat{m}_1(\boldsymbol{x}; 0) \dots \hat{m}_q(\boldsymbol{x}; 0)]||^{-1} [\hat{m}_1(\boldsymbol{x}; 0) \dots \hat{m}_q(\boldsymbol{x}; 0)]'. \tag{16}$$

This estimator implements the idea of separately smoothing each component of $\boldsymbol{Y}$ with a common smoothing parameter, although selecting different levels of smoothing is straightforward. In the next section, the possibility to incorporate the correlation structure between the $\boldsymbol{Y}^{(\ell)}$s into our local estimators will be discussed for the local linear fit.

For $d = q = 2$, the smoother $\arctan(\hat{m}_2/\hat{m}_1)$, which estimates dependence between two angles, can be compared with the circular–circular regression smoother of Di Marzio, Panzera, and Taylor (2013). Their approach is simpler for working in only one dimension, additionally, it does not exhibit singularity issues and features the same rates for bias and variance. However, it has a different structure from ours and is not generalizable to the sphere case for the same reasons seen in the link to previous work discussed in Section 3.2.

Let $\boldsymbol{1}_u$ be a $u \times 1$ vector of ones. We have

Theorem 5.1. Given the random sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \dots, n\}$, taking values in $\mathbb{S}^{d-1} \times \mathbb{S}^{q-1}$, if assumption (i), with m_ℓ (all entries of $\mathbb{\Sigma}$ respectively) in place of m (s^2, resp.), and assumptions (ii)–(iv) of Theorem 3.1 hold, then, for estimator (16) at $\boldsymbol{x} \in \mathbb{S}^{d-1}$

$$E[\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] - \boldsymbol{m}(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)} \begin{bmatrix} \mathrm{tr}\big(\mathcal{D}^2_{\tilde{m}_1}(\boldsymbol{x})\big) + 2f(\boldsymbol{x})^{-1} \mathcal{D}'_{\tilde{f}}(\boldsymbol{x}) \mathcal{D}_{\tilde{m}_1}(\boldsymbol{x}) \\ \vdots \\ \mathrm{tr}\big(\mathcal{D}^2_{\tilde{m}_q}(\boldsymbol{x})\big) + 2f(\boldsymbol{x})^{-1} \mathcal{D}'_{\tilde{f}}(\boldsymbol{x}) \mathcal{D}_{\tilde{m}_q}(\boldsymbol{x}) \end{bmatrix} + o_p(b_2(\kappa) \boldsymbol{1}_q),$$

$$\mathrm{var}[\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] = \frac{\nu_0(\kappa)}{n f(\boldsymbol{x})} \mathbb{\Sigma}(\boldsymbol{x}) + o_p\left(\frac{\nu_0(\kappa)}{n} \boldsymbol{I}_q\right).$$

Proof. See the Appendix. □

The accuracy of (16) could be measured by the function

$$\mathcal{L}[\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)] := E[2(1 - \hat{\boldsymbol{m}}(\boldsymbol{x}; 0)' \boldsymbol{m}(\boldsymbol{x})) \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n], \tag{17}$$

where, since $||\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)|| = ||\boldsymbol{m}(\boldsymbol{x})|| = 1$, $\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)' \boldsymbol{m}(\boldsymbol{x})$ is the cosine of the angle between $\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)$ and $\boldsymbol{m}(\boldsymbol{x})$. This loss corresponds to $E[||\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) m(\boldsymbol{x})||^2 \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n]$, and can then be decomposed into the sum of $E[||\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) - E[\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)]||^2 \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n]$ and $||E[\hat{\boldsymbol{m}}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] - \boldsymbol{m}(\boldsymbol{x})||^2$, and regarded as the spherical counterpart of the conditional mean-squared error, since its summands are the conditional spherical variance and the conditional squared bias of $\hat{\boldsymbol{m}}(\boldsymbol{x}; 0)$, respectively.

Recalling approximations (9), from Theorem 5.1 it results that for estimator (16) with the Langevin kernel, the value of κ which minimizes the asymptotic version of (17) is

$$\left\{ \frac{2^{d-1} \pi^{(d-1)/2} n f(\boldsymbol{x}) \sum_{\ell=1}^{q} J_\ell^2(\boldsymbol{x})}{(d-1) \sum_{\ell=1}^{q} s_\ell^2(\boldsymbol{x})} \right\}^{2/(d+3)},$$

where $J_\ell(\boldsymbol{x}) := \mathrm{tr}(\mathcal{D}^2_{\tilde{m}_\ell}(\boldsymbol{x})) + 2\mathcal{D}'_{\tilde{f}}(\boldsymbol{x}) \mathcal{D}_{\tilde{m}_\ell}(\boldsymbol{x}) f^{-1}(\boldsymbol{x})$, $\ell \in \{1, \dots, q\}$. Note that the rate achieved by using κ which minimizes the leading term of loss (17) depends only on the dimension (d) of the input space.

5.2 Local Linear Estimator

Let $\mathbb{A} \otimes \mathbb{B}$ denote the Kronecker product between matrices $\mathbb{A}$ and $\mathbb{B}$, and let $\tilde{\mathbb{A}} := \boldsymbol{I}_q \otimes \mathbb{A}$. Due to decomposition (3), a local linear estimator of $\boldsymbol{m}(\boldsymbol{x})$ could be based on this first-order

expansion

$$\boldsymbol{m}(\boldsymbol{X}_i) \approx \boldsymbol{m}(\boldsymbol{x}) + \theta_i \tilde{\boldsymbol{\xi}}_i{}' \boldsymbol{D}_{\bar{\boldsymbol{m}}}(\boldsymbol{x}), \tag{18}$$

where $\bar{\boldsymbol{m}}(\boldsymbol{x}) := [\bar{m}_1(\boldsymbol{x}) \dots \bar{m}_q(\boldsymbol{x})]'$, and $\boldsymbol{D}_{\bar{\boldsymbol{m}}}(\boldsymbol{x}) := [\mathcal{D}'_{\bar{m}_1}(\boldsymbol{x}) \dots \mathcal{D}'_{\bar{m}_q}(\boldsymbol{x})]'$. Now, let $\boldsymbol{\mathcal{B}} := [\boldsymbol{\mathcal{B}}'_1 \dots \boldsymbol{\mathcal{B}}'_q]'$, with $\boldsymbol{\mathcal{B}}_\ell := [\beta_0^{(\ell)} \ \boldsymbol{\beta}_1'^{(\ell)}]'$, $\ell \in \{1, \dots, q\}$, and set $\mathbb{Y} := [\boldsymbol{Y}'^{(1)} \dots \boldsymbol{Y}'^{(q)}]'$, with $\boldsymbol{Y}^{(\ell)}$ being the $n \times 1$ vector having $Y_i^{(\ell)}$ as ith entry. Then, by the same arguments used for estimator (12) as applied to each entry of $\boldsymbol{m}(\boldsymbol{x})$, a local linear version of (16) could be defined as the solution for $[\beta_0^{(1)} \dots \beta_0^{(q)}]'$ of

$$\min_{\boldsymbol{\mathcal{B}}} ||\tilde{\mathbb{W}}^{1/2}(\mathbb{Y} - \tilde{\mathbb{X}}\boldsymbol{\mathcal{B}})||^2 \quad \text{subject to} \quad \tilde{\boldsymbol{Q}}_1{}' \boldsymbol{\mathcal{B}} = \boldsymbol{0}_q, \tag{19}$$

which, with the additional constraint $|||[\beta_0^{(1)} \dots \beta_0^{(q)}]|| = 1$, leads to

$$\hat{\boldsymbol{m}}(\boldsymbol{x}; 1) = |||[\hat{m}_1(\boldsymbol{x}; 1) \dots \hat{m}_q(\boldsymbol{x}; 1)]||^{-1} [\hat{m}_1(\boldsymbol{x}; 1) \dots \hat{m}_q(\boldsymbol{x}; 1)]', \tag{20}$$

where $\hat{m}_\ell(\boldsymbol{x}; 1) := \boldsymbol{e}_1' \boldsymbol{Q}_2 (\boldsymbol{Q}_2' \mathbb{X}' \mathbb{W} \mathbb{X} \boldsymbol{Q}_2)^{-1} \boldsymbol{Q}_2' \mathbb{X}' \mathbb{W} \boldsymbol{Y}^{(\ell)}$. As a matter of interpretation, if we see the sphere as a manifold, we observe that this estimator and the next one interestingly combine elements of both intrinsic and extrinsic analyses. While its mean structure belongs to an extrinsic scheme, the local approximation on which they are based has an intrinsic nature, for using expansion (18), which is equipped with a geodesic distance. For such an estimator, we obtain the following.

Theorem 5.2. Given the random sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \dots, n\}$ taking values in $\mathbb{S}^{d-1} \times \mathbb{S}^{q-1}$, if assumption (i), with m_ℓ (all entries of $\mathbb{\Sigma}$ respectively) in place of m (s^2, resp.), and assumptions (ii)–(iv) of Theorem 3.1 hold, then, for estimator (20)

$$\begin{aligned} &E[\hat{\boldsymbol{m}}(\boldsymbol{x}; 1) \mid \boldsymbol{X}_1, \dots \boldsymbol{X}_n] - \boldsymbol{m}(\boldsymbol{x}) \\ &\quad = \frac{b_2(\kappa)}{2(d-1)} \begin{bmatrix} \operatorname{tr}\left(\mathcal{D}^2_{\bar{m}_1}(\boldsymbol{x})\right) \\ \vdots \\ \operatorname{tr}\left(\mathcal{D}^2_{\bar{m}_q}(\boldsymbol{x})\right) \end{bmatrix} + o_p(b_2(\kappa)\mathbf{1}_q), \\ &\operatorname{var}[\hat{\boldsymbol{m}}(\boldsymbol{x}; 1) \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] \\ &\quad = \frac{\nu_0(\kappa)}{nf(\boldsymbol{x})} \mathbb{\Sigma}(\boldsymbol{x}) + o_p\left(\frac{\nu_0(\kappa)}{n} \boldsymbol{I}_q\right). \end{aligned}$$

Proof. See the Appendix. □

Now, for $\ell \in \{1, \dots, q\}$, $j \in \{1, \dots, q\}$, and $\ell \neq j$, denote by $\boldsymbol{S}_{\ell,\ell}$ the diagonal matrix of order n having $s_\ell^2(\boldsymbol{X}_i)$ as (i, i)th entry, and by $\boldsymbol{S}_{\ell,j}$ the diagonal matrix of order n having $s_{\ell,j}(\boldsymbol{X}_i)$ as (i, i)th entry. Furthermore, let $\mathbb{V}$ denote a block matrix with (i, j)th entry $\boldsymbol{S}_{i,j}$, $(i, j) \in \{1, \dots, q\} \times \{1, \dots, q\}$. In what follows, we will treat $\mathbb{V}$ as known. When $\mathbb{V}$ is unknown we could replace it by a consistent estimate. Then, to take into account the correlation structure between the components of $\boldsymbol{Y}$, we could define our estimator as the solution for $[\beta_0^{(1)} \cdots \beta_0^{(q)}]'$ of the problem

$$\min_{\boldsymbol{\mathcal{B}}} ||\mathbb{V}^{-1/2} \tilde{\mathbb{W}}^{1/2}(\mathbb{Y} - \tilde{\mathbb{X}}\boldsymbol{\mathcal{B}})||^2 \quad \text{subject to} \quad \tilde{\boldsymbol{Q}}_1{}' \boldsymbol{\mathcal{B}} = \boldsymbol{0}_q. \tag{21}$$

However, starting from (21), the resulting estimator is defined as

$$\begin{aligned} \hat{\boldsymbol{m}}^*(\boldsymbol{x}; 1) := \ &\|\tilde{\boldsymbol{e}}_1{}' \tilde{\boldsymbol{Q}}_2 (\tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \tilde{\mathbb{X}} \tilde{\boldsymbol{Q}}_2)^{-1} \tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \mathbb{Y}\|^{-1} \\ &\times \tilde{\boldsymbol{e}}_1{}' \tilde{\boldsymbol{Q}}_2 (\tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \tilde{\mathbb{X}} \tilde{\boldsymbol{Q}}_2)^{-1} \tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \mathbb{Y}, \end{aligned} \tag{22}$$

and, we get:

Theorem 5.3. Let $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \dots, n\}$ be a random sample taking values in $\mathbb{S}^{d-1} \times \mathbb{S}^{q-1}$. If assumption (i), with m_ℓ in place of m, and assumptions (ii)–(iv) of Theorem 3.1 hold, and for each $(\ell, j) \in \{1, \dots, q\} \times \{1, \dots, q\}$, the matrices $\boldsymbol{S}_{\ell,j}$s are nonsingular, and the gradients of the extensions to $\mathbb{R}^d$ of their entries exist in a neighborhood of $\boldsymbol{x}$, then, the asymptotic conditional bias and variance of estimator (22) are the same of those obtained by componentwise smoothing of the $\boldsymbol{Y}_i$'s.

Proof. See the Appendix. □

Observe that this theorem could be adapted also for use in the multivariate Euclidean setting, whereas the univariate one has been inspected by Welsh and Yeeb (2006).

6. ROTATIONAL EQUIVARIANCE

An important property when dealing with spherical data is *rotational equivariance*, which is the analog of equivariance under ordinary translation of linear data. Here, we treat this from a spherical-linear perspective, the extension to the case with a spherical response being trivial.

Definition 6.1. Let $\mathbb{G}_\alpha$ denote the matrix of order d which performs rotations of vectors in $\mathbb{S}^{d-1}$ about the x-axis by an angle $\alpha \in (0, 2\pi)$. Assume that the regression function $m : \mathbb{S}^{d-1} \to \mathbb{R}$ is the target. We say that an estimator $\hat{m}$ of m is rotationally equivariant if and only if for whatever location $\boldsymbol{x} \in \mathbb{S}^{d-1}$ we have $\hat{m}(\boldsymbol{x}) = \hat{m}_G(\mathbb{G}_\alpha \boldsymbol{x})$, where $\hat{m}$ is the estimator using the sample $\{(\boldsymbol{X}_i, Y_i), i = 1, \dots, n\}$, and $\hat{m}_G$ is the estimator using the sample $\{(\mathbb{G}_\alpha \boldsymbol{X}_i, Y_i), i = 1, \dots, n\}$.

This property makes the inference robust with respect to the coordinate system, since *any* data rotation produces an estimate which has a certain relationship with the one based on nonrotated data. This can be regarded as a data reduction property because distinct samples constitute an equivalence class.

Rotational equivariance of our estimators is easily seen. First, consider that rotational symmetry properties of our kernels give straightforwardly rotational equivariance of them. This gives, in turn, equivariance of local constant estimation. Concerning the local linear estimator, first recall that, by the choice of $\boldsymbol{Q}_2 := \begin{bmatrix} \boldsymbol{x} & \boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}' \end{bmatrix}'$, we have $\boldsymbol{A}_2 = \mathbb{X}\boldsymbol{Q}_2 = [\boldsymbol{x} + \theta_1 \boldsymbol{\xi}_1 \dots \boldsymbol{x} + \theta_n \boldsymbol{\xi}_n]'$. Hence performing rotations by the angle α of both the vectors $\boldsymbol{x}$ and $\boldsymbol{\xi}_i$, $i \in \{1, \dots, n\}$, the matrix $\boldsymbol{A}_2$ becomes $\boldsymbol{A}_2 \mathbb{G}_\alpha'$, which leads to

$$\begin{aligned} &(\mathbb{G}_\alpha \boldsymbol{A}_2' \mathbb{W} \boldsymbol{A}_2 \mathbb{G}_\alpha')^{-1} \mathbb{G}_\alpha \boldsymbol{A}_2' \mathbb{W} \boldsymbol{Y} \\ &\quad = (\mathbb{G}_\alpha')^{-1} (\boldsymbol{A}_2' \mathbb{W} \boldsymbol{A}_2)^{-1} \mathbb{G}_\alpha^{-1} \mathbb{G}_\alpha \boldsymbol{A}_2' \mathbb{W} \boldsymbol{Y} \\ &\quad = \mathbb{G}_\alpha (\boldsymbol{A}_2' \mathbb{W} \boldsymbol{A}_2)^{-1} \boldsymbol{A}_2' \mathbb{W} \boldsymbol{Y} \end{aligned} \tag{23}$$

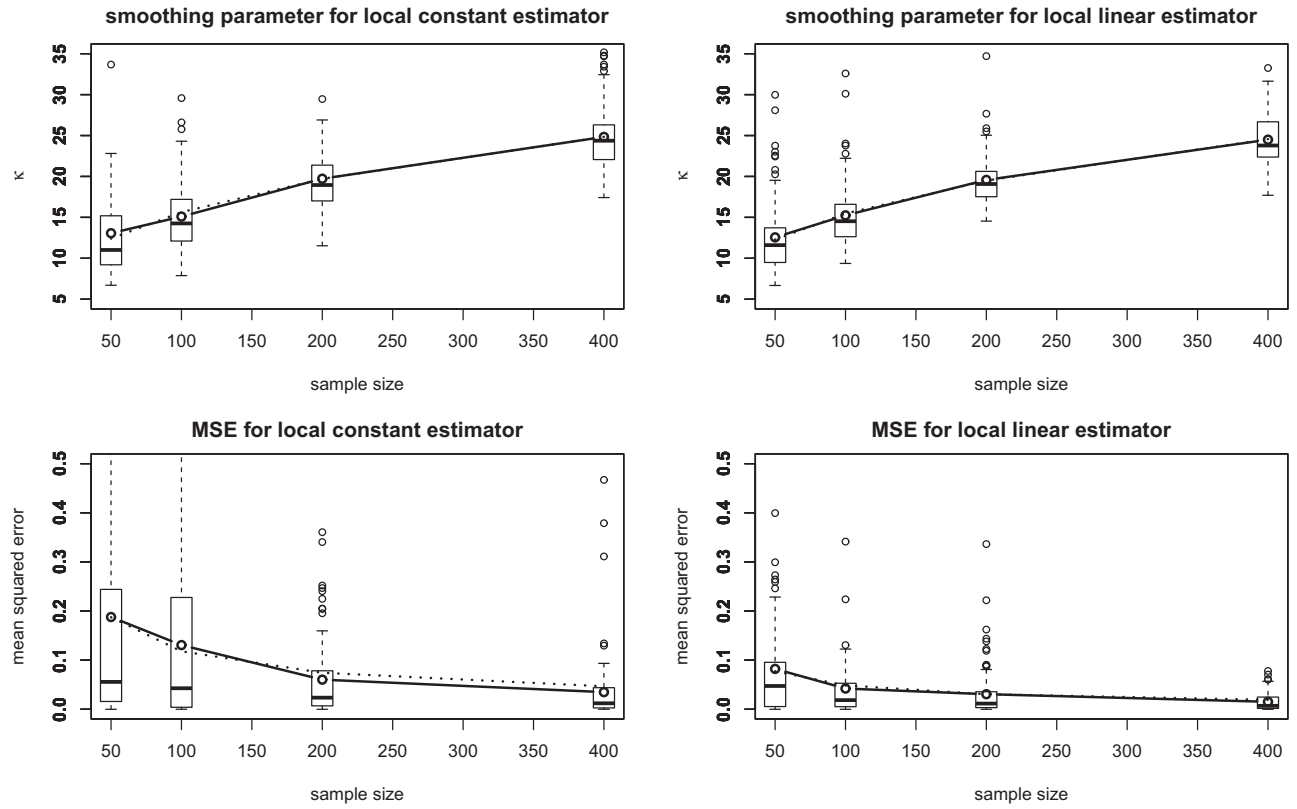

Figure 1. Top row: values of CV-selected smoothing parameters for each of 100 samples (in which $\boldsymbol{x}$ is uniform random over the sphere, and y is generated from model (24)) for various sample sizes, summarized by a boxplot, with mean values connected by a continuous line, and corresponding theoretical values connected by a dotted line. Bottom row: corresponding mean-squared prediction error evaluated at the point $(1, 0, 0)$. Left column: local constant estimator; right column: local linear estimator.

where the last identity holds since $\mathbb{G}_\alpha^{-1} = \mathbb{G}_\alpha'$. Moreover, $\boldsymbol{Q}_2$ becomes $[\mathbb{G}_\alpha \boldsymbol{x} \quad \boldsymbol{I}_d - \mathbb{G}_\alpha \boldsymbol{x}\boldsymbol{x}'\mathbb{G}_\alpha']'$, which leads to

$$\boldsymbol{e}_1' \begin{bmatrix} \boldsymbol{x}'\mathbb{G}_\alpha' \\ \boldsymbol{I}_d - \mathbb{G}_\alpha \boldsymbol{x}\boldsymbol{x}'\mathbb{G}_\alpha' \end{bmatrix} = \boldsymbol{e}_1' \boldsymbol{Q}_2 \mathbb{G}_\alpha'.$$

This, along with (23), and using again $\mathbb{G}_\alpha' = \mathbb{G}_\alpha^{-1}$ yields the result.

For estimator (15), formula (23) turns out to be premultiplied by $\mathbb{G}_\alpha(\boldsymbol{I}_d - \boldsymbol{x}'\boldsymbol{x})\mathbb{G}_\alpha'$, not giving equivariance.

7. SPHERE–LINEAR REGRESSION SIMULATIONS

In this section, we use simulated data to illustrate the dependence on the sample size, and to compare our methods with others, in the case of $\boldsymbol{x} \in \mathbb{S}^2$ and response variable $y \in \mathbb{R}$. We initially investigate the effect of sample size on the smoothing parameter and the resulting mean-squared error. Then, we briefly describe two other methods that we will use as competitors. Our choice reflects the main and easily implementable "off-the-shelf" methods: spherical harmonics and splines. Concerning the weight function, in all of our experiments we will use the Langevin kernel, in the next sections as well.

7.1 Optimal Smoothing and Sample Size

With $\boldsymbol{x} = (x_1, x_2, x_3)$ we consider the model

$$y = \exp(x_1 + x_2 + x_3) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \tag{24}$$

with $\sigma = 0.25$. Since our theoretical results are mostly related to mean-squared error for prediction at a specific point, we will estimate y at the point $\boldsymbol{x} = (1, 0, 0)$. For each value of $n \in \{50, 100, 200, 400\}$, we take 100 samples in which $\boldsymbol{X}$ is uniformly distributed over the sphere. For each sample, we use cross-validation to find the smoothing parameter, and with this choice of κ we compute the squared error $(\hat{y} - \exp(1))^2$, where $\hat{y}$ is the predicted value at $(1, 0, 0)$. The results are shown in Figure 1 in which we have shown the boxplots of the smoothing parameters, and the squared prediction errors, together with the sample means, and a fitted line obtained from the theoretical rates ($\kappa = O(n^{1/3})$ and AMSE $= O(n^{-2/3})$). It appears that cross-validation has the potential to agree with such optimal decays, the well-known instability issue being less problematic for the local linear fit.

7.2 Other Methods

Kernel ridge regression was recently applied to spherical design spaces of whatever dimension by Cao et al. (2013). They used a spherical harmonic kernel given by

$$K(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{\pi} \sum_{j=1}^{5} \cos(j \cos^{-1}(\boldsymbol{u}'\boldsymbol{v}))$$

and then estimate coefficients $\boldsymbol{\beta}$ to minimize $||\boldsymbol{y} - \mathbb{K}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||^2$ where $\mathbb{K}$ is a matrix of order n having $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as its (i, j)th entry, and $\lambda \geq 0$ is a regularization parameter. In our implementation, we used the R ridge

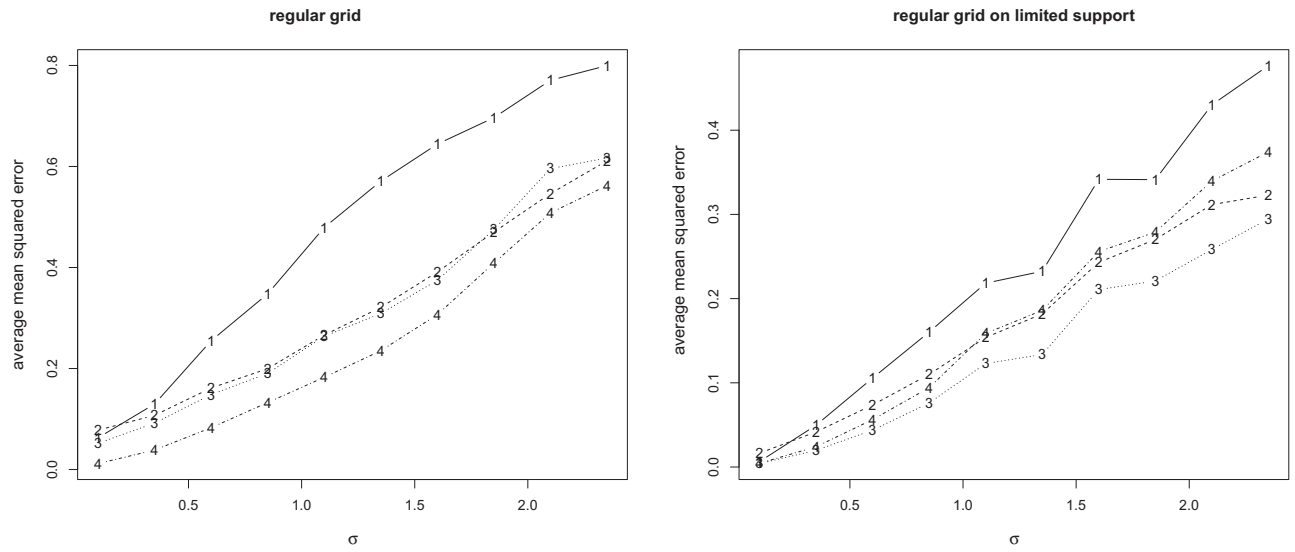

Figure 2. Average mean-squared prediction error as a function of σ for each of four methods, where the explanatory data are uniformly distributed over the whole sphere (left) and restricted to latitude less than 0.5 rad (right). The functions are evaluated on a regular grid of points, compared using the average of the mean-squared errors. Lines correspond to spherical harmonic kernel with ridge regression (1), local constant (2), local linear (3), and splines on the sphere (4).

regression function `lm.ridge` and selected λ by generalized cross-validation, although the authors, in fact, do not specify any strategy for selecting this parameter. After estimating coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$, we predict y at a new location $\boldsymbol{x}$ by using $\hat{y} = \sum K(\boldsymbol{x}_i, \boldsymbol{x})\hat{\beta}_i$.

A further method is to use reduced rank splines on the *ordinary* sphere (Wahba 1981, 1982; Wood 2003) in which we have used the R functions described in `smooth.construct.sos.smooth.spec` of the `mgcv` library (Wood 2013). In our implementation, we used the second derivative penalty. We experimented with selecting the basis dimension by leave-one-out cross-validation, but this revealed to be a costly computational burden for which the results were little better than the default value of 50, and so the results are reported for this value. Finally recall that, in the standard setting, splines are close to kernel estimators, these latter being superior in the minimax sense, as defined by Jennen-Steinmetz and Gasser (1988).

7.3 Results

We consider two examples for which we will use the model

$$y = m(\boldsymbol{x}) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \tag{25}$$

where $m(\boldsymbol{x}) = 3 + x_2\{x_1^2 + x_2^2\}^{-1/2}(\cos(1.3)\sqrt{1 - x_3^2} + x_3 \sin(1.3))$ and $\sigma \in (0.10, 2.35)$. In the first example, we take $\boldsymbol{x}$ to be uniform over the sphere, and in the second example we restrict the support to those values with the absolute value of the latitude less than 0.5. For each σ, we take 100 samples of 100 observations as the training set. Although our theory mostly presents results for mean-squared error for a specific point $\boldsymbol{x}$ we here consider an average over many points as this is more akin to the usage of the methods in applications. So for the test set we use a regular grid of points: 1212 over the full sphere, of which 580 lie in the limited support. The results are summarized in Figure 2 where we show the average (over 100 samples) of the "out-of-sample" mean-squared prediction error—given by $n^{-1}\sum_i (m(\boldsymbol{x}_i) - \hat{y}_i)^2$—for each of the methods.

It can be seen that the spherical harmonic kernel with ridge regression method is worst in both settings. The spline method is best for the case when the support of $\boldsymbol{x}$ is the whole sphere, but this method is not as good as the kernel smoothers for the case of limited support. Both of the smoothers perform similarly for the full sphere, but the local linear performs somewhat better in the case of limited support.

8. SPHERE–SPHERE REGRESSION SIMULATIONS

In this section, we consider the case of $\boldsymbol{x} \in \mathbb{S}^2$ and response variable $\boldsymbol{y} \in \mathbb{S}^2$. The only obvious competitor here is Chang's method (Chang et al. 2000), though it is also possible to apply any spherical–linear methos to each component of the vector, and then to combine the results. When both variables lie on the sphere, that is, we have data $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{S}^2 \times \mathbb{S}^2, i \in \{1, \ldots, n\}$, then it is straightforward to estimate a 3×3 rotation matrix $\mathbb{A}$ to minimize $\sum_{i=1}^n ||\boldsymbol{y}_i - \boldsymbol{x}_i'\mathbb{A}||$. Let $\mathbb{Y}$ be the matrix whose ith row is $\boldsymbol{x}_i'$, and $\mathbb{A}$ the matrix whose ith row is $\boldsymbol{y}_i'$. Then (Chang 1986), the rotation matrix $\mathbb{A}$ is estimated by $\hat{\mathbb{A}} = \mathbb{U}\mathbb{V}'$, where $\mathbb{U}$ and $\mathbb{V}$ are determined by the singular value decomposition of $\mathbb{Y}'\mathbb{A} = \mathbb{U}\mathbb{\Lambda}\mathbb{V}'$, where $\mathbb{U}, \mathbb{V} \in SO(3)$, and $\mathbb{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$. This *rigid* transformation of the data, used by Chang et al. (2000), is computationally very simple and fast.

8.1 Results

In general, we suppose that, conditioned on $\boldsymbol{X} = \boldsymbol{x}$, $\boldsymbol{Y}$ has a von-Mises–Fisher distribution with concentration parameter κ and mean direction $\boldsymbol{m}(\boldsymbol{x})$, that is, $\boldsymbol{Y} \mid \boldsymbol{x} \sim vM(\boldsymbol{m}(\boldsymbol{x}), \kappa)$. In the first case, we choose $\boldsymbol{m}$ to be a rigid rotation of the data, and in the second case we consider a nonrigid transformation in which

$$m(\boldsymbol{x}) \propto (\sin(\psi)\cos(\phi), 2\sin(\psi)\cos(\phi - 0.3), \\ 3\sin(\psi)\cos(\phi - 1)), \tag{26}$$

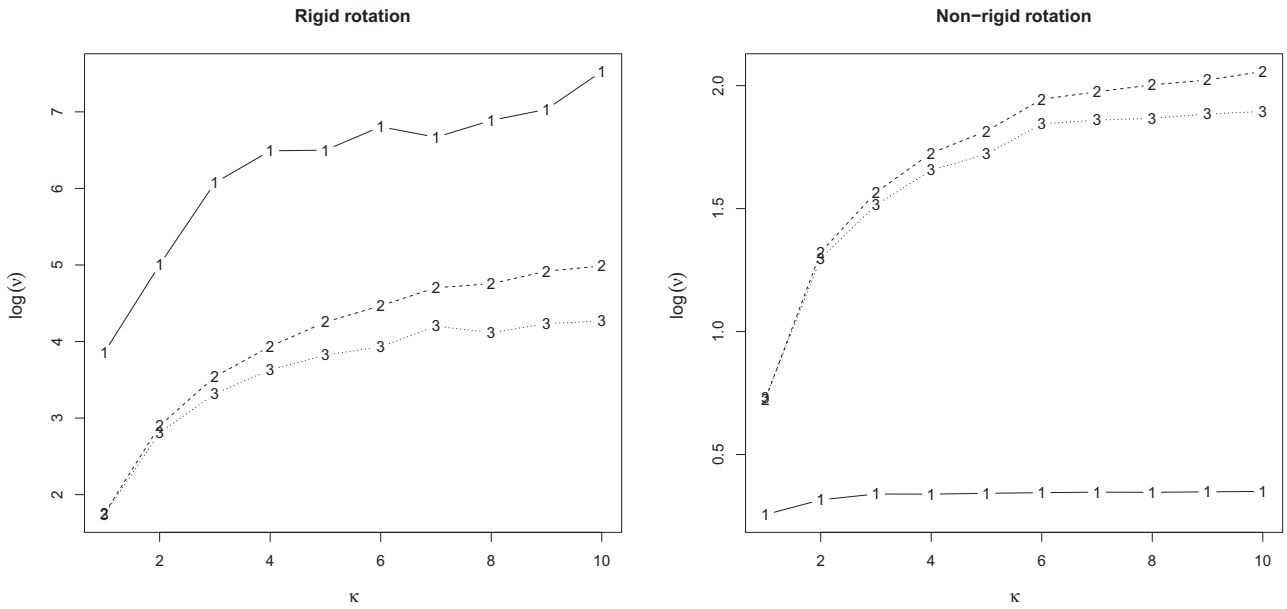

Figure 3. For 1212 design points regularly placed on the sphere the estimated concentration parameter ν of the errors in correspondence to the noise in the observations, as measured by concentration parameter κ. One hundred simulations of 100 observations in the training data. Left: rigid rotation model; right: nonrigid transformation using model given by regression function (26). Line numbers: (1) rotation matrix; (2) local linear; (3) local constant.

where (ψ, ϕ) is the longitude, latitude representation of $\boldsymbol{x}$, and $m(\boldsymbol{x})$ is normalized so that $||m(\boldsymbol{x})|| = 1$. In both rigid and nonrigid transformations, we consider various values of κ, in which 100 observations are simulated with $\boldsymbol{X}$ uniformly distributed on the sphere, and $\boldsymbol{Y}$ is generated according to the stated model. The observations are used to estimate the required functions, using cross-validation where appropriate. We then use the functions to obtain predicted values $\hat{\boldsymbol{y}}$ at our regular grid of 1212 values of $\boldsymbol{x}$, and this is repeated 100 times for each κ. To summarize the quality of the estimates, we consider $\theta_i = \cos^{-1}(\hat{\boldsymbol{y}}_i' \boldsymbol{m}(\boldsymbol{x}_i))$, $i = 1, \ldots, 1212$. These angles are summarized using the maximum likelihood estimate of the concentration parameter when the angles are assumed to have a von Mises distribution, that is, the solution of ν to the equation $\bar{C} = \coth(\nu) - 1/\nu$ where $\bar{C} = n^{-1} \sum_i \cos\theta_i$. The average (over 100 simulations) values of ν are shown in Figure 3 for both the rigid and nonrigid transformations. It can be seen that the rotation model performs best for the rigid transformation (as expected), but performs very poorly for the nonrigid transformation. It should be remembered that *large* values of ν and κ (the concentration parameters) indicate *smaller* errors, so the overall monotonic pattern is as expected. In this case, the local linear estimate performs better than the local constant estimate for all except high levels of noise (corresponding to small concentration κ).

9. REAL-DATA EXAMPLES

We consider two datasets which have been chosen to illustrate the applicability of our methods. In both cases, there are well-established alternative nonstatistical approaches, though there may be potential for our methods to contribute to these fields. The first dataset concerns the orientation of Earth's magnetic field, as measured from a satellite, and the second concerns prediction of wind directions at locations on the Earth surface. In both cases, we used a common smoothing parameter for each component.

9.1 Geomagnetic Field

Earth's magnetic field extends from the inner core into the atmosphere and beyond, and protects the Earth from solar wind which emanates from the sun. Since about 1980, various satellites have been launched to measure this field, using three-axis magnetometers to probe the three-dimensional structure. Over the last few decades, the results from these remote sensors have been combined to produce ever more accurate *world magnetic models* which are used for navigation and heading referencing systems using the geomagnetic field. One of the first satellites was NASA's MAGSAT spacecraft, which orbited the earth every 88 min for about 7 months at around 400 km altitude. Data, available during 2/11/79–6/5/80, are recorded every half second and can be downloaded from NASA's National Space Science Data Center.[1]

We illustrate our methods using a sample of the available MAGSAT data. First, from each day we sampled 22 equally spaced observations, and then a random sample of 1000 was taken from this combined set. We used the geocentric latitude and longitude of the spacecraft (but not the time/day of the observation) as the explanatory $(\boldsymbol{x})$ variables, and the north, east, and vertical components of the magnetic field vector were converted to polar coordinates and used as the response $(\boldsymbol{y})$ variables. The relationships between the four (polar coordinate) variables are shown in Figure 4.

For the purposes of comparison, we split the 1000 observations (randomly) into train and test sets. The training set is used to select smoothing parameters (by cross-validation) and the test sets are used to provide a measure of performance.

[1] *nssdcftp.gsfc.nasa.gov/spacecraft_data/magsat*

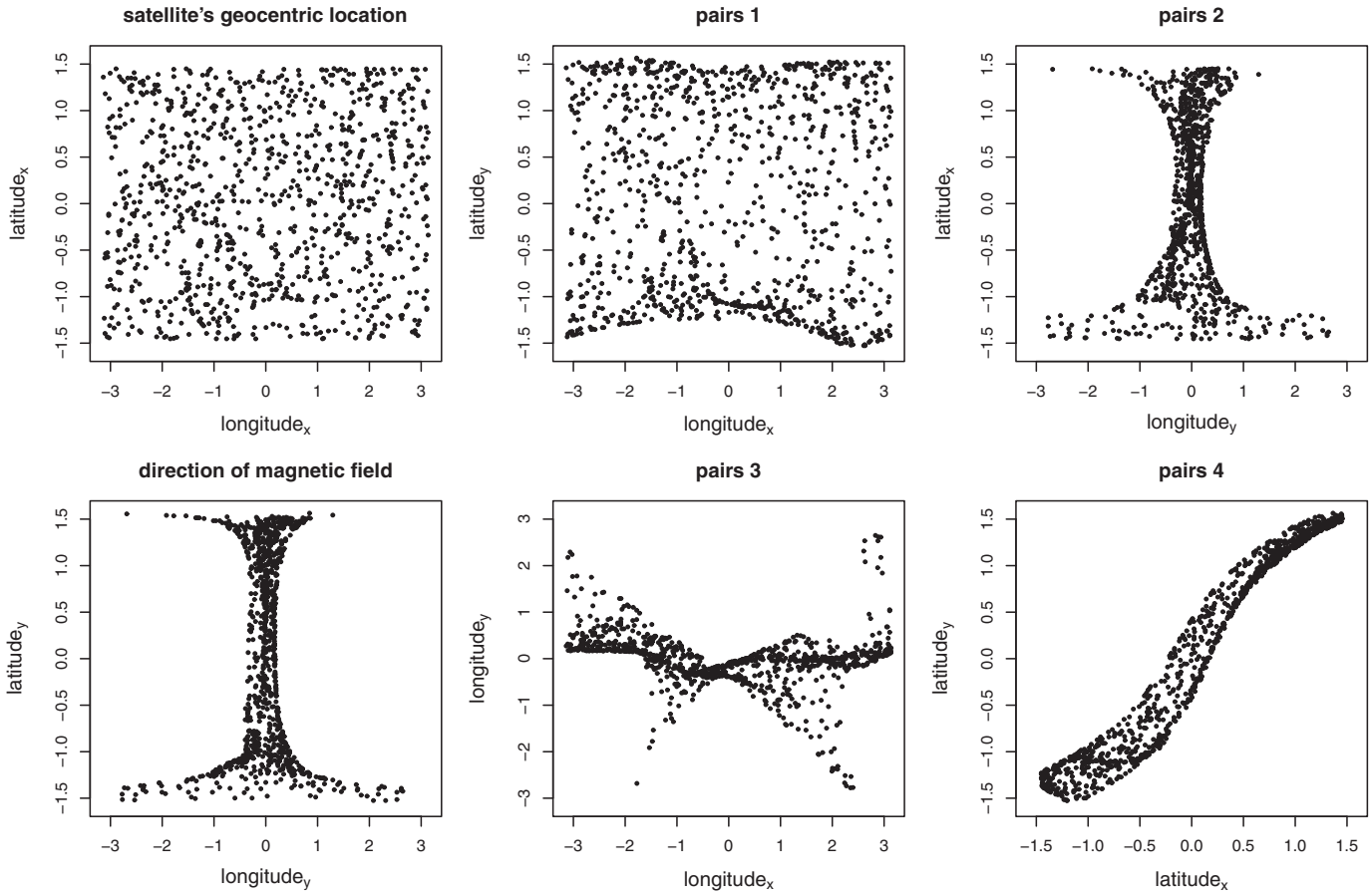

Figure 4. Pairwise plots for each combination of the four—latitude and longitude for the response (y) and explanatory (x)—variables of the magnetic field data. Left column shows longitude–latitude plots for the satellite location (top) and the direction of the magnetic field (bottom). Other plots (pairs 1–4) show relationships between variables as indicated on the axis labels.

The cross-validation selected smoothing parameter (the same for each component) was found to be $\kappa = 402.2$ for the local linear estimate, and $\kappa = 75.4$ for the local constant estimate.

For target values $\boldsymbol{y}_i$ and estimated ones $\hat{\boldsymbol{y}}_i$ we measure the accuracy by considering the angles $\theta_i = \cos^{-1}(\hat{\boldsymbol{y}}_i' \boldsymbol{y}_i)$. These angles are summarized as in the simulations above.

In Table 1, we report the concentration parameters for the test data for the methods we consider: local constant, local linear, splines on the sphere, spherical harmonic kernel with ridge regression, and a rotation matrix (Chang et al. 2000). It can be seen that the nonparametric methods easily outperform the inflexible rotation model, with the local linear estimate performing somewhat better than the others. This is consistent with the results of the simulations for low noise.

Table 1. Maximum likelihood estimates ($\hat{\nu}$) of concentration parameters (using a Fisher model) for angular errors based on estimates of remotely sensed magnetic field orientation. Smoothing parameters (when required) were chosen using leave-one-out cross-validation for the training set

	Method				
Error measure	Local constant	Local linear	Splines on sphere	Harmonic kernel	Rigid rotation
$\hat{\nu}$	703	8,340	3,998	1,062	2.5

9.2 Prediction of Wind Directions

Wind direction modeling is a difficult task. The problem is that direction at a point has traditionally been predicted through directional time series, although data registered at a single location have the potential to be strongly erratic due to phenomena like turbulence, microbursts, and gusts. From a statistical point of view, we say that wind direction data are considerably nonlinear and non-Gaussian. Hirata et al. (2008) thoroughly discussed this (see also the references therein), and proposed, as a possible solution, a parametric nonlinear model taking into account the observations arising from multiple observation points. By following this reasoning, we think that local smoothing of all directions registered nearby the prediction point could be useful, at least for predicting surface trends of the directions in a context of spatial data analysis. Moreover, it is arguable that the erratic feature of the data would require interval estimates as an additional tool. This perspective would motivate using our spherical–linear regression fit, as a generalization of their idea, both for point and interval estimation, as detailed in the following.

Wind directions are automatically recorded by the U.S. NOAA's *National Data Buoy Center* at many locations every 15 min. Most of these sites are close to the seaboard of the USA, but several are in the Caribbean, and some are elsewhere around the world. We have selected a single time point (noon on 15 July 2011) and extracted the wind direction from the annual historical datasets at each of 422 locations. The locations (which are taken

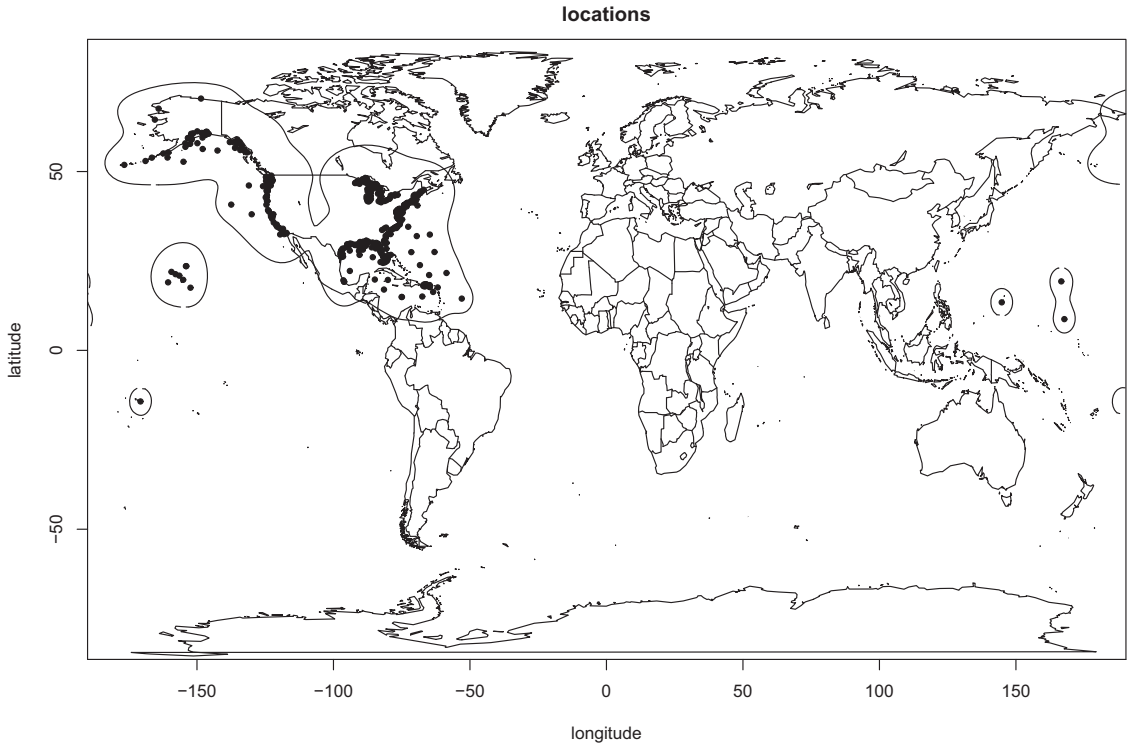

Figure 5. Locations of wind direction recording sites, as collated by NDBC, on 15th July 2011, together with a kernel density estimate at square root scale. Only one contour level is shown for clarity.

as the explanatory variables) are shown in Figure 5, together with a kernel density estimate of the locations. We consider estimation of the wind direction at each location, using the information from all the other locations. This sphere–circle regression problem is approached using both the local constant, and the local linear fits along with leave-one-out cross-validated smoothing.

At each location we have a predicted wind direction, which can be compared to the actual wind direction. Fisher et al. (1996) used pivotal methods to obtain confidence regions for directional data, but their approach seems hard to adapt in this setting of conditional estimation. So, an approximate confidence interval for each prediction is calculated as follows. Using Equations (7) and (14) we need to estimate $f(\boldsymbol{x})$, $s^2(\boldsymbol{x})$, and compute $\nu_0(\kappa)$ for the selected κ. We estimate $f(\boldsymbol{x})$ by spherical kernel density estimate of Hall, Watson, and Cabrera (1987), with likelihood cross-validated smoothing. We assume stationarity and isotropy

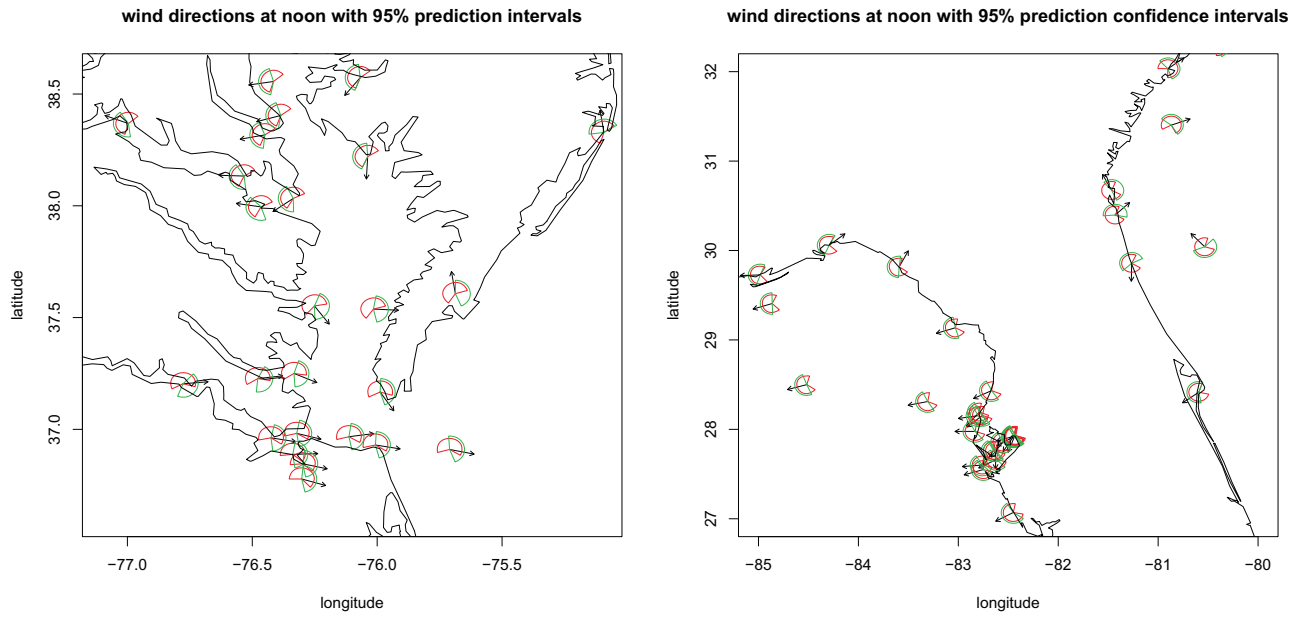

Figure 6. Wind predictions for two sample regions (Left: Chesapeake Bay; Right: northern Florida). The circular segments show 95% confidence intervals, and the arrows show actual wind directions at each location. The larger (radius) segments correspond to the local linear estimates, and the smaller radius segments correspond to the local constant estimates.

of the random errors, and use the (leave-one-out) residuals to estimate $s^2(\boldsymbol{x}) = s^2$. Then, we assume a von Mises distribution for the estimate of the wind direction, in which the concentration parameter is taken to be $1/\text{var}(\hat{m}(\boldsymbol{x}; p), p \in \{0, 1\})$. This can then used to find an interval, (l, u) say, for which $\int_l^u g(\theta)\text{d}\theta = 1 - \alpha$ where $g(\cdot)$ is a von Mises density with mean $\hat{m}(\boldsymbol{x}; p)$, $p \in \{0, 1\}$, and concentration given by the inverse of the estimated variance.

For each site, we can obtain an interval corresponding to the prediction. When compared to the actual values, 16 (3.8%) of the local constant prediction intervals are not contained in the 95% confidence intervals, and 15 (3.6%) of the local linear estimators are not contained in their respective confidence intervals. The width of the intervals ranges from 3.7 to 6.0 (median 5.0) rad for the local constant estimator, and from 2.6 to 6.0 (median 4.0) for the local linear estimator. It should be noted that a width of 6.0 is almost consistent with a uniform distribution such that there is no preferred direction for the estimate; this occurs only for the isolated observation in the South Pacific $(-14, -171)$. These widths, together with the greater coverage, indicate that the local linear estimator appears to be performing better overall.

Figure 6 shows a sample of results for two parts of the Eastern seaboard of USA. We have noted that the results look reasonable, even though the confidence intervals are quite wide.

APPENDIX

Proof of Theorem 3.1. Using the fact that $n^{-1}\sum K_\kappa(\cos(\theta_i)) = f(\boldsymbol{x}) + o_p(1)$, and expansion (3) for $m(\boldsymbol{X}_i)$, $i \in \{1, \ldots, n\}$, in a neighborhood of $\boldsymbol{x}$, we have

$$E[\hat{m}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \approx \{f(\boldsymbol{x}) + o_p(1)\}^{-1} n^{-1} \left\{\sum_{i=1}^n K_\kappa(\cos(\theta_i)) \left(m(\boldsymbol{x}) + \boldsymbol{\xi}_i' \mathcal{D}_{\tilde{m}}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\xi}_i' \mathcal{D}^2_{\tilde{m}}(\boldsymbol{x})\boldsymbol{\xi}_i\right)\right\}.$$

Now, because $\int_{\Omega_x} \boldsymbol{\xi}\omega_{d-2}(d\boldsymbol{\xi}) = \boldsymbol{0}_d$, $\int_{\Omega_x} \boldsymbol{\xi}\boldsymbol{\xi}'\omega_{d-2}(d\boldsymbol{\xi}) = \omega_{d-2}(d-1)^{-1}(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')$, and $\boldsymbol{x}'\mathcal{D}_{\bar{f}}(\boldsymbol{x}) = 0$, in virtue of assumptions (i)–(iii), we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i\boldsymbol{\xi}_i \\
&\quad = \int_0^\pi K_\kappa(\cos(\theta))\theta \sin^{d-2}(\theta)d\theta \\
&\qquad \times \int_{\Omega_x} \boldsymbol{\xi} f(\boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta))\omega_{d-2}(d\boldsymbol{\xi}) + o_p(1) \\
&\quad = \int_0^\pi K_\kappa(\cos(\theta))\theta^2 \sin^{d-2}(\theta)d\theta \\
&\qquad \times \int_{\Omega_x} \boldsymbol{\xi}\boldsymbol{\xi}'\mathcal{D}_{\bar{f}}(\boldsymbol{x})\omega_{d-2}(d\boldsymbol{\xi}) + o_p(1) \\
&\quad = b_2(\kappa)(d-1)^{-1}\mathcal{D}_{\bar{f}}(\boldsymbol{x}) + o_p(\boldsymbol{1}_d b_2(\kappa)), \qquad \text{(A.1)} \\
&\frac{1}{n}\sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i' \\
&\quad = \int_0^\pi K_\kappa(\cos(\theta))\theta^2 \sin^{d-2}(\theta)d\theta \\
&\qquad \times \int_{\Omega_x} \boldsymbol{\xi}\boldsymbol{\xi}' f(\boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta))\omega_{d-2}(d\boldsymbol{\xi}) + o_p(1) \\
&\quad = b_2(\kappa)(d-1)^{-1}(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')f(\boldsymbol{x}) + o_p(\boldsymbol{I}_d b_2(\kappa)). \qquad \text{(A.2)}
\end{aligned}$$

The above approximations, in conjunction with the fact that for a function g defined on $\mathbb{S}^d$, $\boldsymbol{\xi}'\mathcal{D}^2_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi} = \text{tr}(\mathcal{D}^2_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi}\boldsymbol{\xi}')$, and $\boldsymbol{x}'\mathcal{D}^2_{\bar{g}}(\boldsymbol{x})\boldsymbol{x} = 0$, yield

$$E[\hat{m}(\boldsymbol{x}; 0) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \approx \{f(\boldsymbol{x})\}^{-1}\big\{f(\boldsymbol{x})m(\boldsymbol{x}) + (d-1)^{-1}b_2(\kappa) \times \big[\mathcal{D}'_{\bar{f}}(\boldsymbol{x})\mathcal{D}_{\tilde{m}}(\boldsymbol{x}) + \frac{1}{2}\text{tr}(\mathcal{D}^2_{\tilde{m}}(\boldsymbol{x}))f(\boldsymbol{x})\big]\big\}.$$

Now, using assumptions (i) and (iv), the asymptotic variance can be calculated starting from

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n K^2_\kappa(\cos(\theta_i))s^2(\boldsymbol{X}_i) \\
&\quad = \int_0^\pi K^2_\kappa(\cos(\theta))\sin^{d-1}(\theta)d\theta \\
&\qquad \times \int_{\Omega_x} s^2(\boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta))f(\boldsymbol{x}\cos(\theta) \\
&\qquad + \boldsymbol{\xi}\sin(\theta))\omega_{d-2}(d\boldsymbol{\xi}) + o_p(1) \\
&\quad = \nu_0(\kappa)f(\boldsymbol{x})s^2(\boldsymbol{x}) + o_p(\nu_0(\kappa)). \qquad \text{(A.3)}
\end{aligned}$$

Finally, the asymptotic distribution of the estimator comes from its linearity and bias–variance results. □

Proof of Theorem 3.2. Letting $\boldsymbol{M}$ be the $n \times 1$ vector having $m(\boldsymbol{X}_i)$ as its ith entry, we get

$$E[\hat{m}(\boldsymbol{x}; 1) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = \boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\boldsymbol{M},$$

and, using expansion (3) for $m(\boldsymbol{X}_i)$ in a neighborhood of $\boldsymbol{x}$, $i \in \{1, \ldots, n\}$, we have

$$\boldsymbol{M} \approx \mathbb{X}\begin{bmatrix} m(\boldsymbol{x}) \\ \mathcal{D}_{\tilde{m}}(\boldsymbol{x}) \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \theta_1^2\boldsymbol{\xi}_1'\mathcal{D}^2_{\tilde{m}}(\boldsymbol{x})\boldsymbol{\xi}_1 \\ \vdots \\ \theta_n^2\boldsymbol{\xi}_n'\mathcal{D}^2_{\tilde{m}}(\boldsymbol{x})\boldsymbol{\xi}_n \end{bmatrix},$$

which leads to

$$\begin{aligned}
&E[\hat{m}(\boldsymbol{x}; 1) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \\
&\quad \approx \boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\begin{bmatrix} m(\boldsymbol{x}) \\ \mathcal{D}_{\tilde{m}}(\boldsymbol{x}) \end{bmatrix} \\
&\qquad + \frac{1}{2}\boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\begin{bmatrix} \theta_1^2\boldsymbol{\xi}_1'\mathcal{D}^2_m(\boldsymbol{x})\boldsymbol{\xi}_1 \\ \vdots \\ \theta_n^2\boldsymbol{\xi}_n'\mathcal{D}^2_m(\boldsymbol{x})\boldsymbol{\xi}_n \end{bmatrix}. \qquad \text{(A.4)}
\end{aligned}$$

In virtue of (11), the first term in (A.4) is $m(\boldsymbol{x})$ whereas, recalling that

$$\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2 = \sum_{i=1}^n K_\kappa(\cos(\theta_i))\big\{\boldsymbol{x}\boldsymbol{x}' + \theta_i\boldsymbol{\xi}_i\boldsymbol{x}' + \theta_i\boldsymbol{x}\boldsymbol{\xi}_i' + \theta_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i'\big\},$$

approximations given in the proof of Theorem 3.1 for the summands in the right-hand side of the above equation lead to

$$\begin{aligned}
n^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2 &\approx \boldsymbol{x}\boldsymbol{x}'f(\boldsymbol{x}) + (d-1)^{-1}b_2(\kappa) \\
&\quad \times \big[\mathcal{D}_{\bar{f}}(\boldsymbol{x})\boldsymbol{x}' + \boldsymbol{x}\mathcal{D}'_{\bar{f}}(\boldsymbol{x}) + f(\boldsymbol{x})(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')\big], \qquad \text{(A.5)}
\end{aligned}$$

$$\begin{aligned}
(n^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1} &\approx f(\boldsymbol{x})^{-1}\big[\boldsymbol{x}\boldsymbol{x}' - f(\boldsymbol{x})^{-1}\boldsymbol{x}\mathcal{D}'_{\bar{f}}(\boldsymbol{x}) - f(\boldsymbol{x})^{-1} \\
&\quad \times \mathcal{D}_{\bar{f}}(\boldsymbol{x})\boldsymbol{x}' + (d-1)b_2(\kappa)^{-1}(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')\big],
\end{aligned}$$

and

$$\boldsymbol{Q}_2(n^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2' \approx \begin{bmatrix} f(\boldsymbol{x})^{-1} & -\mathcal{D}'_{\bar{f}}(\boldsymbol{x})f(\boldsymbol{x})^{-2} \\ -\mathcal{D}_{\bar{f}}(\boldsymbol{x})f(\boldsymbol{x})^{-2} & (d-1)\{b_2(\kappa)f(\boldsymbol{x})\}^{-1}(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}') \end{bmatrix}. \qquad \text{(A.6)}$$

Additionally, we see that

$$\mathbb{X}'\mathbb{W}\begin{bmatrix}\theta_1^2\boldsymbol{\xi}_1'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_1\\ \vdots\\ \theta_n^2\boldsymbol{\xi}_n'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_n\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^2\boldsymbol{\xi}_i'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_i\\ \sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^3\boldsymbol{\xi}_i\{\boldsymbol{\xi}_i'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_i\}\end{bmatrix},$$

and, after approximating similarly to before, along with $\boldsymbol{x}'\mathcal{D}_{\tilde{f}}^2(\boldsymbol{x})\boldsymbol{x}=0$, and assumptions (i)–(iii) we get

$$n^{-1}\mathbb{X}'\mathbb{W}\begin{bmatrix}\theta_1^2\boldsymbol{\xi}_1'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_1\\ \vdots\\ \theta_n^2\boldsymbol{\xi}_n'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_n\end{bmatrix}=\begin{bmatrix}b_2(\kappa)(d-1)^{-1}f(\boldsymbol{x})\mathrm{tr}\big(\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\big)+o_p(b_2(\kappa))\\ O_p\left(\mathbf{1}_d b_4(\kappa)\right)\end{bmatrix},\tag{A.7}$$

and plugging (A.6) and (A.7) in (13) yields the bias. Concerning the variance, we have

$$\mathrm{var}[\hat{m}(\boldsymbol{x};1)\mid\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]=\boldsymbol{e}_1'\boldsymbol{Q}_2(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\boldsymbol{S}\mathbb{W}\mathbb{X}\boldsymbol{Q}_2\times(\boldsymbol{Q}_2'\mathbb{X}'\mathbb{W}\mathbb{X}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\boldsymbol{e}_1,\tag{A.8}$$

where $\boldsymbol{S}$ denotes a diagonal matrix of order n having $s^2(\boldsymbol{X}_i)$ as (i,i)th entry. Now, since

$$n^{-1}\mathbb{X}'\mathbb{W}\boldsymbol{S}\mathbb{X}\mathbb{W}=\begin{bmatrix}n^{-1}\sum_{i=1}^n K_\kappa^2(\cos(\theta_i))s^2(\boldsymbol{X}_i) & & & \\ & n^{-1}\sum_{i=1}^n K_\kappa^2(\cos(\theta_i))\theta_i s^2(\boldsymbol{X}_i)\boldsymbol{\xi}_i' & & \\ n^{-1}\sum_{i=1}^n K_\kappa^2(\cos(\theta_i))\theta_i s^2(\boldsymbol{X}_i)\boldsymbol{\xi}_i & & & \\ & & n^{-1}\sum_{i=1}^n K_\kappa^2(\cos(\theta_i))\theta_i^2 s^2(\boldsymbol{X}_i)\boldsymbol{\xi}_i\boldsymbol{\xi}_i'\end{bmatrix},\tag{A.9}$$

approximations similar to those used above leads to

$$n^{-1}\mathbb{X}'\mathbb{W}\boldsymbol{S}\mathbb{X}\mathbb{W}\approx\begin{bmatrix}\nu_0(\kappa)s^2(\boldsymbol{x})f(\boldsymbol{x}) & & \\ & \nu_2(\kappa)\{f(\boldsymbol{x})\mathcal{D}_{\tilde{s}^2}'(\boldsymbol{x})+s^2(\boldsymbol{x})\mathcal{D}_{\tilde{f}}'(\boldsymbol{x})\} & \\ \nu_2(\kappa)\{f(\boldsymbol{x})\mathcal{D}_{\tilde{s}^2}(\boldsymbol{x})+s^2(\boldsymbol{x})\mathcal{D}_{\tilde{f}}(\boldsymbol{x})\} & & \\ & & (d-1)^{-1}\nu_2(\kappa)s^2(\boldsymbol{x})f(\boldsymbol{x})(\boldsymbol{I}_d-\boldsymbol{x}\boldsymbol{x}')\end{bmatrix},\tag{A.10}$$

and assumptions (i), (ii), and (iv) of Theorem 3.1, the above approximation, together with (A.6), give the variance. Finally, the linearity property along with the bias–variance results lead to its asymptotic distribution. □

Proof of Theorem 3.3. We employ the same idea as in the proof of Theorem 2 in Masry and Fan (1997). In particular, for $\hat{m}(\boldsymbol{x};0)$, letting $w_i:=\{\sum_{s=1}^n K_\kappa(\cos(\theta_i))\}^{-1}K_\kappa(\cos(\theta_i))$, and $Z_i:=w_i\{Y_i-m(\boldsymbol{X}_i)\}$, in virtue of expansion (3) for $m(\boldsymbol{X}_i)$ around $\boldsymbol{x}$, we obtain $\sum_i Z_i\approx\hat{m}(\boldsymbol{x};0)-\sum_i w_i\{m(\boldsymbol{x})+\theta_i\boldsymbol{\xi}_i'\mathcal{D}_{\tilde{m}}(\boldsymbol{x})+\frac{1}{2}\theta_i^2\boldsymbol{\xi}_i'\mathcal{D}_{\tilde{m}}^2(\boldsymbol{x})\boldsymbol{\xi}_i\}$. Then, the expectation of $\hat{m}(\boldsymbol{x};0)$ follows by using the fact that $n^{-1}\sum_{i=1}^n K_\kappa(\cos(\theta_i))\xrightarrow{p}f(\boldsymbol{x})$, along with approximations (A.1) and (A.2). For the variance, by stationarity,

$$\mathrm{var}\left[\frac{1}{n}\sum_{i=1}^n Z_i\right]=\frac{1}{n}\mathrm{var}[Z_1]+\frac{2}{n}\sum_{\ell=1}^{n-1}\left(1-\frac{\ell}{n}\right)\mathrm{cov}[Z_1,Z_{1+\ell}].$$

Thus, recalling the variance result in Theorem 3.1, and noting that by assumption (i) with $\lambda=2$, $\nu_0(\kappa)=O(\kappa^{\gamma_d})$ and $\mathrm{var}[Z_1]=O(\kappa^{\gamma_d})$. Now, choose a sequence of integers u_n satisfying $u_n\kappa^{-\gamma_d}\to 0$ as $u_n\to\infty$, and set

$$J_1:=\sum_{\ell=1}^{u_n-1}|\mathrm{cov}(Z_1,Z_{\ell+1})|\qquad\text{and}\qquad J_2:=\sum_{\ell=u_n}^{n-1}|\mathrm{cov}(Z_1,Z_{\ell+1})|.$$

Hence, reasoning as in Masry and Fan (1997), by assumption (ii), and the choice of u_n, we get $J_1=o(\kappa^{\gamma_d})$. Now, for ρ-mixing processes we have $J_2\le\mathrm{var}[Z_1]\sum_{\ell=u_n}^{\infty}\rho(\ell)=o(\kappa^{\gamma_d})$, while for strongly mixing processes, using the Davydov inequality, along with assumptions (i)–(iii), we obtain

$$J_2\le 8D\kappa^{2\gamma_d(\lambda-1)/\lambda}u_n^{-a}\sum_{i=u_n}^{\infty}i^a[\alpha(i)]^{1-2/\lambda},$$

with $D\in\mathbb{R}$. So $u_n=\kappa^{\gamma_d(1-2/\lambda)/a}$ yields $J_2\le o(\kappa^{\gamma_d})$. The bias and variance of $\hat{m}(\boldsymbol{x};1)$ similarly follow if

$$w_i=\boldsymbol{x}'\left(\sum_{j=1}^n K_\kappa(\cos(\theta_j))\{\boldsymbol{x}\boldsymbol{x}'+\theta_j\boldsymbol{x}\boldsymbol{\xi}_j'+\theta_j\boldsymbol{\xi}_j\boldsymbol{x}'+\theta_j^2\boldsymbol{\xi}_j\boldsymbol{\xi}_j'\}\right)^{-1}\times(\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)K_\kappa(\cos(\theta_i)).$$

□

Proof of Theorem 5.1. First, observe that results of Theorem 3.1 apply for each $\hat{m}_\ell$, $\ell\in\{1,\ldots,q\}$, and asymptotic bias directly follows since $||[\hat{m}_1\ldots\hat{m}_q]||\xrightarrow{p}1$. For the variance, approximation (A.3) applies for each entry of $\mathrm{var}[[\hat{m}_1\ldots\hat{m}_q]']$ with s^2 replaced by s_ℓ^2 for the diagonal terms and by $s_{j,\ell}$ for the off-diagonal ones. □

Proof of Theorem 5.2. The bias result is calculated by observing that the arguments used, in the proof of Theorem 3.2, for the asymptotic conditional expectation of $\hat{m}(\boldsymbol{x};1)$, hold for each $\hat{m}_\ell$, $\ell\in\{1,\ldots,q\}$, and that $||[\hat{m}_1\ldots\hat{m}_q]||\xrightarrow{p}1$. For the variance, we consider just the proof for the case $q=2$. In particular, we have

$$\mathrm{var}\,[\hat{\boldsymbol{m}}(\boldsymbol{x};1)\,|\,\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]=\tilde{\boldsymbol{e}}_1'\tilde{\boldsymbol{Q}}_2(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}\tilde{\mathbb{W}}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2\times(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\boldsymbol{e}}_1,$$

where

$$\mathbb{V}=\begin{bmatrix}\boldsymbol{S}_{1,1} & \boldsymbol{S}_{1,2}\\ \boldsymbol{S}_{1,2} & \boldsymbol{S}_{2,2}\end{bmatrix}.$$

Furthermore,

$$n^{-1}\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}\tilde{\mathbb{W}}\tilde{\mathbb{X}}=\begin{bmatrix}\boldsymbol{V}_{1,2} & \boldsymbol{V}_{1,2}\\ \boldsymbol{V}_{1,2} & \boldsymbol{V}_{2,2}\end{bmatrix},$$

where for $(j,\ell)\in\{1,2\}$, $\boldsymbol{V}_{j,\ell}$ corresponds to the matrix (A.9) with s^2 replaced by s_ℓ^2 for $\ell=j$, and by $s_{j,\ell}$ when $j\neq\ell$, and then can be approximated using (A.10). These approximations, along the approximation in (A.6) applied to each block of the matrix $\tilde{\boldsymbol{Q}}_2(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'$, and $||[\hat{m}_1\ldots\hat{m}_q]||\xrightarrow{p}1$, yield the variance. □

Proof of Theorem 5.3. For ease of the presentation, we refer to $q=2$. Recalling $\boldsymbol{Q}_2$, we have

$$E[\hat{\boldsymbol{m}}^*(\boldsymbol{x};1)\mid\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]=\tilde{\boldsymbol{e}}_1'\tilde{\boldsymbol{Q}}_2(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\begin{bmatrix}\boldsymbol{M}_1\\ \boldsymbol{M}_2\end{bmatrix},$$

where, for $\ell\in\{1,2\}$, $\boldsymbol{M}_\ell:=[m_\ell(\boldsymbol{X}_1)\ldots m_\ell(\boldsymbol{X}_n)]'$. Then, the expansion

$$\begin{bmatrix}\boldsymbol{M}_1\\ \boldsymbol{M}_2\end{bmatrix}\approx\tilde{\mathbb{X}}\boldsymbol{\mathcal{B}}+\frac{1}{2}\begin{bmatrix}\boldsymbol{L}_1\\ \boldsymbol{L}_2\end{bmatrix},$$

where $\boldsymbol{L}_\ell := [\theta_1^2 \boldsymbol{\xi}_1' \mathcal{D}^2_{\tilde{m}_\ell}(\boldsymbol{x})\boldsymbol{\xi}_1 \ldots \theta_n^2 \boldsymbol{\xi}_n' \mathcal{D}^2_{\tilde{m}_\ell}(\boldsymbol{x})\boldsymbol{\xi}_n]'$, and similar arguments as those used in the proof of Theorem 3.2, imply that the first term in the expansion of the conditional expectation is $\boldsymbol{m}(\boldsymbol{x})$, and

$$E[\hat{\boldsymbol{m}}^*(\boldsymbol{x};1) - \boldsymbol{m}(\boldsymbol{x}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = \frac{1}{2}\tilde{\boldsymbol{e}}_1' \tilde{\boldsymbol{Q}}_2 (n^{-1}\tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \tilde{\mathbb{X}} \tilde{\boldsymbol{Q}}_2)^{-1} n^{-1} \tilde{\boldsymbol{Q}}_2' \tilde{\mathbb{X}}' \tilde{\mathbb{W}} \mathbb{V}^{-1} \begin{bmatrix} \boldsymbol{L}_1 \\ \boldsymbol{L}_2 \end{bmatrix}.$$

Now, since

$$\mathbb{\Sigma}(\boldsymbol{X}_i) = \begin{bmatrix} s_1^2(\boldsymbol{X}_i) & s_{1,2}(\boldsymbol{X}_i) \\ s_{1,2}(\boldsymbol{X}_i) & s_2^2(\boldsymbol{X}_i) \end{bmatrix},$$

setting $\varsigma(\boldsymbol{X}_i) := s_2^2(\boldsymbol{X}_i)\{s_1^2(\boldsymbol{X}_i)s_2^2(\boldsymbol{X}_i) - s_{1,2}^2(\boldsymbol{X}_i)\}^{-1}$, $\varrho(\boldsymbol{X}_i) := -s_{1,2}(\boldsymbol{X}_i)\{s_1^2(\boldsymbol{X}_i)s_2^2(\boldsymbol{X}_i) - s_{1,2}^2(\boldsymbol{X}_i)\}^{-1}$, and $\vartheta(\boldsymbol{X}_i) := s_1^2(\boldsymbol{X}_i)\{s_1^2(\boldsymbol{X}_i)s_2^2(\boldsymbol{X}_i) - s_{1,2}^2(\boldsymbol{X}_i)\}^{-1}$, it results

$$\mathbb{\Sigma}^{-1}(\boldsymbol{X}_i) = \begin{bmatrix} \varsigma(\boldsymbol{X}_i) & \varrho(\boldsymbol{X}_i) \\ \varrho(\boldsymbol{X}_i) & \vartheta(\boldsymbol{X}_i) \end{bmatrix} \quad \text{and} \quad \mathbb{V}^{-1} = \begin{bmatrix} \boldsymbol{U}_\varsigma & \boldsymbol{U}_\varrho \\ \boldsymbol{U}_\varrho & \boldsymbol{U}_\vartheta \end{bmatrix},$$

where, for a function h defined on $\mathbb{S}^{d-1}$, $\boldsymbol{U}_h$ stands for a diagonal matrix of order n having $h(\boldsymbol{X}_i)$ as its (i,i)th entry, $i \in \{1, \ldots, n\}$. Then, denoting $K_i = K_\kappa(\cos(\theta_i))$, we have

$$\tilde{\boldsymbol{Q}}_2{}'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2 = \begin{bmatrix} \sum_{i=1}^n (\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)(\boldsymbol{x}'+\theta_i\boldsymbol{\xi}_i')K_i\varsigma(\boldsymbol{X}_i) & \sum_{i=1}^n (\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)(\boldsymbol{x}'+\theta_i\boldsymbol{\xi}_i')K_i\varrho(\boldsymbol{X}_i) \\ \sum_{i=1}^n (\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)(\boldsymbol{x}'+\theta_i\boldsymbol{\xi}_i')K_i\varrho(\boldsymbol{X}_i) & \sum_{i=1}^n (\boldsymbol{x}+\theta_i\boldsymbol{\xi}_i)(\boldsymbol{x}'+\theta_i\boldsymbol{\xi}_i')K_i\vartheta(\boldsymbol{X}_i) \end{bmatrix},$$

and, the same approximations as those used in the proof of Theorem 3.2 to derive (A.5), lead to

$$n^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2 \approx \begin{bmatrix} \boldsymbol{P}_\varsigma & \boldsymbol{P}_\varrho \\ \boldsymbol{P}_\varrho & \boldsymbol{P}_\vartheta \end{bmatrix}$$

where, for a function h defined on $\mathbb{S}^{d-1}$,

$$\boldsymbol{P}_h := \boldsymbol{x}\boldsymbol{x}' f(\boldsymbol{x})h(\boldsymbol{x}) + (d-1)^{-1}b_2(\kappa)\big[\{h(\boldsymbol{x})\mathcal{D}_{\tilde{f}}(\boldsymbol{x}) + f(\boldsymbol{x})\mathcal{D}_{\tilde{h}}(\boldsymbol{x})\}\boldsymbol{x}' + \boldsymbol{x}\{h(\boldsymbol{x})\mathcal{D}'_{\tilde{f}}(\boldsymbol{x}) + f(\boldsymbol{x})\mathcal{D}'_{\tilde{h}}(\boldsymbol{x})\} + f(\boldsymbol{x})h(\boldsymbol{x})(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')\big],$$

and, using the inversion formula for a symmetric block matrix

$$\tilde{\boldsymbol{Q}}_2(n^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2' \approx \begin{bmatrix} \boldsymbol{Q}_2\big(\boldsymbol{P}_\varsigma - \boldsymbol{P}_\varrho\boldsymbol{P}_\vartheta^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{Q}_2' & -\boldsymbol{Q}_2\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho\big(\boldsymbol{P}_\vartheta - \boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{Q}_2' \\ -\boldsymbol{Q}_2\big(\boldsymbol{P}_\vartheta - \boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{Q}_2' & \boldsymbol{Q}_2\big(\boldsymbol{P}_\vartheta - \boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{Q}_2' \end{bmatrix}.$$

The first block of the above matrix, with $\boldsymbol{x}$ dropped as the argument of all involved functions, is

$$\boldsymbol{Q}_2\big(\boldsymbol{P}_\varsigma - \boldsymbol{P}_\varrho\boldsymbol{P}_\vartheta^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{Q}_2' = \frac{1}{|\mathbb{\Sigma}^{-1}|} \times \begin{bmatrix} \vartheta/f & \dfrac{-\{\mathcal{D}'_{\tilde{f}}\vartheta + f(\mathcal{D}'_{\tilde{\varsigma}}\vartheta^2 + \mathcal{D}'_{\tilde{\vartheta}}\varrho^2 - 2\mathcal{D}'_{\tilde{\varrho}}\vartheta\varrho)/|\mathbb{\Sigma}^{-1}|\}}{f^2} \\ \dfrac{-\{\mathcal{D}_{\tilde{f}}\vartheta + f(\mathcal{D}_{\tilde{\varsigma}}\vartheta^2 + \mathcal{D}_{\tilde{\vartheta}}\varrho^2 - 2\mathcal{D}_{\tilde{\varrho}}\vartheta\varrho)/|\mathbb{\Sigma}^{-1}|\}}{f^2} & (\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)\vartheta/(b_2(\kappa)f) \end{bmatrix},$$

while, the lower-right block is

$$\boldsymbol{Q}_2\big(\boldsymbol{P}_\vartheta - \boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho\big)^{-1}\boldsymbol{Q}_2' = \frac{1}{|\mathbb{\Sigma}^{-1}|} \times \begin{bmatrix} \varsigma/f & \dfrac{-\{\mathcal{D}'_{\tilde{f}}\varsigma + f(\mathcal{D}'_{\tilde{\vartheta}}\varsigma^2 + \mathcal{D}'_{\tilde{\varsigma}}\varrho^2 - 2\mathcal{D}'_{\tilde{\varrho}}\varsigma\varrho)/|\mathbb{\Sigma}^{-1}|\}}{f^2} \\ \dfrac{-\{\mathcal{D}_{\tilde{f}}\varsigma + f(\mathcal{D}_{\tilde{\vartheta}}\varsigma^2 + \mathcal{D}_{\tilde{\varsigma}}\varrho^2 - 2\mathcal{D}_{\tilde{\varrho}}\varsigma\varrho)/|\mathbb{\Sigma}^{-1}|\}}{f^2} & (\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)\varsigma/(b_2(\kappa)f) \end{bmatrix},$$

and for the off-diagonal blocks, we have

$$\boldsymbol{Q}_2\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho(\boldsymbol{P}_\vartheta - \boldsymbol{P}_\varrho\boldsymbol{P}_\varsigma^{-1}\boldsymbol{P}_\varrho)^{-1}\boldsymbol{Q}_2' = \frac{1}{|\mathbb{\Sigma}^{-1}|} \times \begin{bmatrix} \varrho/f & \dfrac{-\{\mathcal{D}'_{\tilde{f}}\varrho - f(\mathcal{D}'_{\tilde{\vartheta}}\varsigma\varrho + \mathcal{D}'_{\tilde{\varsigma}}\vartheta\varrho - \mathcal{D}'_{\tilde{\varrho}}\varsigma\vartheta - \mathcal{D}'_{\tilde{\varrho}}\varrho^2)/|\mathbb{\Sigma}^{-1}|\}}{f^2} \\ \dfrac{-\{\mathcal{D}_{\tilde{f}}\varrho - f(\mathcal{D}_{\tilde{\vartheta}}\varsigma\varrho + \mathcal{D}_{\tilde{\varsigma}}\vartheta\varrho - \mathcal{D}_{\tilde{\varrho}}\varsigma\vartheta - \mathcal{D}_{\tilde{\varrho}}\varrho^2)/|\mathbb{\Sigma}^{-1}|\}}{f^2} & (\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)\varrho/(b_2(\kappa)f) \end{bmatrix}.$$

Thus, noting that

$$\frac{1}{|\mathbb{\Sigma}^{-1}(\boldsymbol{x})|}\begin{bmatrix} \vartheta(\boldsymbol{x}) & -\varrho(\boldsymbol{x}) \\ -\varrho(\boldsymbol{x}) & \varsigma(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} s_1^2(\boldsymbol{x}) & s_{1,2}(\boldsymbol{x}) \\ s_{1,2}(\boldsymbol{x}) & s_2^2(\boldsymbol{x}) \end{bmatrix},$$

$$\begin{aligned} \mathcal{D}_{\tilde{s}_1^2}(\boldsymbol{x}) &= -|\mathbb{\Sigma}^{-1}(\boldsymbol{x})|^{-2}(\mathcal{D}_{\tilde{\varsigma}}(\boldsymbol{x})\vartheta^2(\boldsymbol{x}) + \mathcal{D}_{\tilde{\vartheta}}(\boldsymbol{x})\varrho^2(\boldsymbol{x}) \\ &\quad - 2\mathcal{D}_{\tilde{\varrho}}(\boldsymbol{x})\vartheta(\boldsymbol{x})\varrho(\boldsymbol{x})), \\ \mathcal{D}_{\tilde{s}_2^2}(\boldsymbol{x}) &= -|\mathbb{\Sigma}^{-1}(\boldsymbol{x})|^{-2}(\mathcal{D}_{\tilde{\vartheta}}(\boldsymbol{x})\varsigma^2(\boldsymbol{x}) + \mathcal{D}_{\tilde{\varsigma}}(\boldsymbol{x})\varrho^2(\boldsymbol{x}) \\ &\quad - 2\mathcal{D}_{\tilde{\varrho}}(\boldsymbol{x})\varsigma(\boldsymbol{x})\varrho(\boldsymbol{x})), \\ \mathcal{D}_{\tilde{s}_{1,2}}(\boldsymbol{x}) &= |\mathbb{\Sigma}^{-1}(\boldsymbol{x})|^{-2}(\mathcal{D}_{\tilde{\vartheta}}(\boldsymbol{x})\varsigma(\boldsymbol{x})\varrho(\boldsymbol{x}) + \mathcal{D}_{\tilde{\varsigma}}(\boldsymbol{x})\vartheta(\boldsymbol{x})\varrho(\boldsymbol{x}) \\ &\quad - \mathcal{D}_{\tilde{\varrho}}(\boldsymbol{x})\varsigma(\boldsymbol{x})\vartheta(\boldsymbol{x}) - \mathcal{D}_{\tilde{\varrho}}(\boldsymbol{x})\varrho^2(\boldsymbol{x})), \end{aligned}$$

we obtain

$$\tilde{\boldsymbol{Q}}_2(n^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2' \approx \begin{bmatrix} s_1^2/f & -(\mathcal{D}'_{\tilde{f}}s_1^2 - \mathcal{D}'_{\tilde{s}_1^2}f)/f^2 & s_{1,2}/f & -(\mathcal{D}'_{\tilde{f}}s_{1,2} + \mathcal{D}'_{\tilde{s}_{1,2}}f)/f^2 \\ -(\mathcal{D}_{\tilde{f}}s_1^2 - \mathcal{D}_{\tilde{s}_1^2}f)/f^2 & \frac{(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)s_1^2}{b_2(\kappa)f} & -(\mathcal{D}_{\tilde{f}}s_{1,2} + \mathcal{D}_{\tilde{s}_{1,2}}f)/f^2 & \frac{(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)s_{1,2}}{b_2(\kappa)f} \\ s_{1,2}/f & -(\mathcal{D}'_{\tilde{f}}s_{1,2} + \mathcal{D}'_{\tilde{s}_{1,2}}f)/f^2 & s_2^2/f & -(\mathcal{D}'_{\tilde{f}}s_2^2 - \mathcal{D}'_{\tilde{s}_2^2}f)/f^2 \\ -(\mathcal{D}_{\tilde{f}}s_{1,2} + \mathcal{D}_{\tilde{s}_{1,2}}f)/f^2 & \frac{(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)s_{1,2}}{b_2(\kappa)f} & -(\mathcal{D}_{\tilde{f}}s_2^2 - \mathcal{D}_{\tilde{s}_2^2}f)/f^2 & \frac{(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}')(d-1)s_2^2}{b_2(\kappa)f} \end{bmatrix}. \tag{A.11}$$

Additionally, since

$$\mathbb{X}'\mathbb{W}\mathbb{V}^{-1}\begin{bmatrix} \boldsymbol{L}_1 \\ \boldsymbol{L}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^2\{\varsigma(\boldsymbol{X}_i)\boldsymbol{\xi}_i'\mathcal{D}^2_{\tilde{m}_1}(\boldsymbol{x})\boldsymbol{\xi}_i + \varrho(\boldsymbol{X}_i)\boldsymbol{\xi}_i'\mathcal{D}^2_{\tilde{m}_2}(\boldsymbol{x})\boldsymbol{\xi}_i\} \\ \sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^3\{\varsigma(X_i)\boldsymbol{\xi}_i'\mathcal{D}^3_{\tilde{m}_1}(\boldsymbol{x})\boldsymbol{\xi}_i^{\otimes 2} + \varrho(X_i)\boldsymbol{\xi}_i'\mathcal{D}^3_{\tilde{m}_2}(\boldsymbol{x})\boldsymbol{\xi}_i^{\otimes 2}\} \\ \sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^2\{\varrho(\boldsymbol{X}_i)\boldsymbol{\xi}_i'\mathcal{D}^2_{\tilde{m}_1}(\boldsymbol{x})\boldsymbol{\xi}_i + \vartheta(\boldsymbol{X}_i)\boldsymbol{\xi}_i'\mathcal{D}^2_{\tilde{m}_2}(\boldsymbol{x})\boldsymbol{\xi}_i\} \\ \sum_{i=1}^n K_\kappa(\cos(\theta_i))\theta_i^3\{\varrho(X_i)\{\boldsymbol{\xi}_i'\mathcal{D}^3_{\tilde{m}_1}(\boldsymbol{x})\boldsymbol{\xi}_i^{\otimes 2} + \vartheta(X_i)\boldsymbol{\xi}_i'\mathcal{D}^3_{\tilde{m}_2}(\boldsymbol{x})\boldsymbol{\xi}_i^{\otimes 2}\} \end{bmatrix},$$

similar approximations as those used in the proof of Theorem 3.2 to derive (A.7), yield

$$n^{-1}\mathbb{X}'\mathbb{W}\mathbb{V}^{-1}\begin{bmatrix} \boldsymbol{L}_1 \\ \boldsymbol{L}_2 \end{bmatrix} \approx \begin{bmatrix} (d-1)^{-1}b_2(\kappa)f(\boldsymbol{x})\{\varsigma(\boldsymbol{x})\mathrm{Tr}(\mathcal{D}^2_{\tilde{m}_1}(\boldsymbol{x})) + \varrho(\boldsymbol{x})\mathrm{Tr}(\mathcal{D}^2_{\tilde{m}_2}(\boldsymbol{x}))\} \\ O_p(b_4(\kappa)\mathbf{1}_d) \\ (d-1)^{-1}b_2(\kappa)f(\boldsymbol{x})\{\varrho(\boldsymbol{x})\mathrm{Tr}(\mathcal{D}^2_{\tilde{m}_1}(\boldsymbol{x})) + \vartheta(\boldsymbol{x})\mathrm{Tr}(\mathcal{D}^2_{\tilde{m}_2}(\boldsymbol{x}))\} \\ O_p(b_4(\kappa)\mathbf{1}_d) \end{bmatrix},$$

which, combined with the approximation for $\tilde{\boldsymbol{Q}}_2(n^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'$, after a little algebra, gives the bias. For the variance, we have that

$$\begin{aligned}&\operatorname{var}[\hat{\boldsymbol{m}}^*(\boldsymbol{x};1)\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]\\&\quad=\tilde{\boldsymbol{e}}_1\tilde{\boldsymbol{Q}}_2(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{W}}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2\\&\qquad\times(\tilde{\boldsymbol{Q}}_2'\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{X}}\tilde{\boldsymbol{Q}}_2)^{-1}\tilde{\boldsymbol{Q}}_2'\tilde{\boldsymbol{e}}_1,\end{aligned}$$

with

$$n^{-1}\tilde{\mathbb{X}}'\tilde{\mathbb{W}}\mathbb{V}^{-1}\tilde{\mathbb{W}}\tilde{\mathbb{X}}=\begin{bmatrix}\boldsymbol{\Gamma}_\varsigma & \boldsymbol{\Gamma}_\varrho\\ \boldsymbol{\Gamma}_\varrho & \boldsymbol{\Gamma}_\vartheta\end{bmatrix},$$

where for a function h defined on $\mathbb{S}^{d-1}$,

$$\boldsymbol{\Gamma}_h:=\frac{1}{n}\begin{bmatrix}\sum K_\kappa^2(\cos(\theta_i))h(\boldsymbol{X}_i) & \sum K_\kappa^2(\cos(\theta_i))\theta_i h(\boldsymbol{X}_i)\boldsymbol{\xi}_i'\\ \sum K_\kappa^2(\cos(\theta_i))\theta_i h(\boldsymbol{X}_i)\boldsymbol{\xi}_i & \sum K_\kappa^2(\cos(\theta_i))\theta_i^2 h(\boldsymbol{X}_i)\boldsymbol{\xi}_i\boldsymbol{\xi}_i'\end{bmatrix}.$$

Finally approximate each entry of the above matrix as in the proof of Theorem 3.2 and use (A.11). □

[Received December 2012. Revised September 2013.]

REFERENCES

Abrial, P., Moudden, Y., Starck, J.-L., Fadili, J., Delabrouille, J., and Nguyen, M. K. (2008), "CMB Data Analysis and Sparsity," *Statistical Methodology*, 5, 289–298. [749]

Alfed, P., Neamtu, M., and Schumaker, L. L. (1996), "Fitting Scattered Data on Sphere-Like Surfaces Using Spherical Splines," *Journal of Computational and Applied Mathematics*, 73, 5–43. [749]

Aswani, A., Bickel, P., and Tomlin, C. (2011), "Regression on Manifolds: Estimation of the Exterior Derivative," *The Annals of Statistics*, 39, 48–81. [749]

Bickel, P., and Li, B. (2007), "Local Polynomial Regression on Unknown Manifolds," *Lecture Notes—Monograph Series*, 54, 177–186. [749]

Cao, F., Lin, S., Chang, X., and Xu, Z. (2013), "Learning Rates of Regularized Regression on the Unit Sphere," *Science China*, 56, 861–876. [749,755]

Chang, T. (1986), "Spherical Regression," *The Annals of Statistics*, 14, 907–924. [748,756]

Chang, T., Ko, D., Royer, J.-Y., and Lu, J. (2000), "Regression Techniques in Plate Tectonics," *Statistical Science*, 15, 342–356. [748,756,758]

Chapman, G. R., Chen, C., and Kim, P. T. (1995), "Assessing Geometric Integrity Through Spherical Regression Techniques," *Statistica Sinica*, 5, 173–220. [748]

Cheng, M.-Y., and Wu, T. (2014), "Local Linear Regression on Manifolds and Its Geometric Interpretation," *Journal of the American Statistical Association*, 108, 1421–1434. [749]

Di Marzio, M., Panzera, A., and Taylor, C. C. (2009), "Local Polynomial Regression for Circular Predictors," *Statistics & Probability Letters*, 79, 2066–2075. [752]

——— (2013), "Nonparametric Regression for Circular Responses," *Scandinavian Journal of Statistics*, 40, 238–255. [753]

Downs, T. D. (2003), "Spherical Regression," *Biometrika*, 90, 655–668. [748]

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiency," *The Annals of Statistics*, 21, 196–216. [752]

Fisher, N. I., Hall, P., Jing, B.-Y., and Wood, A. T. A. (1996), "Improved Pivotal Methods for Constructing Confidence Regions With Directional Data," *Journal of the American Statistical Association*, 91, 1062–1070. [759]

Hall, P., Watson, G. S., and Cabrera, J. (1987), "Kernel Density Estimation With Spherical Data," *Biometrika*, 74, 751–762. [750,759]

Hamsici, O. C., and Martinez, A. M. (2007), "Spherical-Homoscedastic Distributions: The Equivalency of Spherical and Normal Distributions in Classification," *The Journal of Machine Learning Research*, 8, 1583–1623. [748]

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13, 1465–1481. [753]

Hirata, Y., Mandic, D. P., Suzuki, H., Aihara, K. (2008), "Wind Direction Modelling Using Multiple Observation Points," *Philosophical Transactions of the Royal Society A: Mathematics, Physical and Engineering Sciences*, 366, 591–607. [758]

Jennen-Steinmetz, C., and Gasser, T. (1988), "A Unifying Approach to Nonparametric Regression Estimation," *Journal of the American Statistical Association*, 83, 1084–1089. [756]

Jeon, J., and Taylor, J. W. (2012), "Using Conditional Kernel Density Estimation for Wind Power Density Forecasting," *Journal of the American Statistical Association*, 107, 66–79. [750]

Jupp, P. E., and Kent, J. T. (1987), "Fitting Smooth Paths to Spherical Data," *Journal of the Royal Statistical Society,* Series C, 36, 34–46. [749]

Mackenzie, J. K. (1957), "The Estimation of an Orientation Relationship," *Acta Crystallographica*, 20, 61–62. [748]

Masry, E., and Fan, J. (1997), "Local Polynomial Estimation of Regression Functions for Mixing Processes," *Scandinavian Journal of Statistics*, 24, 165–179. [761]

Monnier, J.-B. (2011), "Nonparametric Regression on the Hyper-Sphere With Uniform Design," *Test*, 20, 412–446. [749]

Pelletier, B. (2006), "Non-Parametric Regression Estimation on Closed Riemannian Manifolds," *Journal of Nonparametric Statistics*, 18, 57–67. [749]

Rivest, L.-P. (1989), "Spherical Regression for Concentrated Fisher-Von Mises Distributions," *The Annals of Statistics*, 17, 307–317. [748]

Shin, H. H., Takahara, G. K., and Murdoch, D. J. (2007), "Optimal Designs for Calibration of Orientations," *Canadian Journal of Statistics*, 35, 365–380. [748]

Wahba, G. (1965), "Section on Problems and Solutions: A Least Squares Estimate of Satellite Attitude," *SIAM Review*, 8, 384–385. [748]

——— (1981), "Spline Interpolation and Smoothing on the Sphere," *SIAM Journal on Scientific and Statistical Computing*, 2, 5–16. [756]

——— (1982), "Erratum," *SIAM Journal on Scientific and Statistical Computing*, 3, 385–386. [756]

Welsh, A. H., and Yeeb, T. W. (2006), "Local Regression for Vector Responses," *Journal of Statistical Planning and Inference*, 136, 3007–3031. [754]

Wood, S. N. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society,* Series B, 65, 95–114. [756]

——— (2013), "mgcv: Mixed GAM Computation Vehicle With GCV/AIC/REML Smoothness," R package version 1.7.22. Available at *http://CRAN.R-project.org/package=mgcv* [756]

Zhu, H., Chen, Y., Ibrahim, J. G., Li, Y., and Lin, W. (2009), "Intrinsic Regression Models for Positive-Definite Matrices With Applications to Diffusion Tensor Imaging," *Journal of the American Statistical Association*, 104, 1203–1212. [749]