Article:
Di Marzio, M, Fensore, S, Panzera, A et al. (2019) Local binary regression with spherical predictors. Statistics and Probability Letters, 144. pp. 30-36. ISSN: 0167-7152

https://doi.org/10.1016/j.spl.2018.07.019

Local binary regression with spherical predictors

Marco Di Marzio[a,*], Stefania Fensore[a], Agnese Panzera[b], Charles C. Taylor[c]

[a] *DSFPEQ, Universitá di Chieti-Pescara, viale Pindaro 42, 65127 Pescara, Italy*
[b] *DiSIA, Universitá di Firenze, viale Morgagni 59, 50134 Florence, Italy*
[c] *Department of Statistics, University of Leeds, Leeds LS2 9JT, UK*

Abstract

We discuss some classes of local estimators for regression when the predictor lies on the d-dimensional sphere and a binary response. In particular, we adapt the theory of local polynomial regression and local likelihood estimation to deal with the problem at hand. We provide asymptotic L_2 properties for some estimators in these classes along with some simulations and a real-data application.

Keywords: Directional data, Local likelihood, Local polynomials, Spherical kernels, Tangent-normal decomposition

MSC:

1. Introduction

Data lying on the unit hypersphere embedded in $\mathbb{R}^d$, $d \geq 2$, arise in many scientific fields. They are typically referred as *directional* or *spherical* data. Classical examples, when $d = 2$, are directions of winds and marine currents, and directions of flight of birds from a point of release. Also, locations on the surface of the *ordinary* sphere ($d = 3$) are ubiquitous in Earth and planetary sciences. Fields of recent interest for directional data include genome sequence representations, text analysis and clustering, morphometrics, and computer vision, see, for example, Hamsici and Martinez (2007).

*Corresponding author
Email addresses: marco.dimarzio@unich.it (Marco Di Marzio), stefania.fensore@unich.it (Stefania Fensore), agnese.panzera@unifi.it (Agnese Panzera), charles@maths.leeds.ac.uk (Charles C. Taylor)

The non-linear nature of the hypersphere sets apart directional statistics from standard methods, which are typically designed for *linear* data. However, in the last few decades *directional statistics* has greatly evolved, and now directional counterparts of many classical statistical methods exist. Classical comprehensive accounts of directional statistics are provided by Batschelet (1981), Fisher et al. (1987), and Mardia and Jupp (2008), and more recently by Ley and Verdebout (2017, 2018).

Kernel-based methods for regression estimation when the response is a linear variable and the predictor has a directional nature have been recently studied. Indeed, the absence of a *boundary* on a spherical domain makes smoothing methods – which typically suffer from boundary bias – well-suited for analysing directional data. In particular, the local polynomial regression for linear response has been studied by Di Marzio et al. (2009) in the case of circular predictors, and by Di Marzio et al. (2014) in the case of a general d-dimensional spherical predictor, as an intermediate step in the spherical-spherical regression estimation. Then, this topic has been also studied by García-Portugués et al. (2016) in the context of goodness-of-fit tests.

Conversely, the special case of a binary response and a directional predictor by means of nonparametric regression methods seems to be unexplored, while for a parametric approach see Fernandes and Cardoso (2016) and references therein. The binary regression problem, apart from being of interest *per se*, is also useful for classification purposes. Nonparametric methods for classification of directional data, based on kernel estimation of spherical densities, have been studied by Di Marzio et al. (2018b).

In the Euclidean setting, kernel-based estimators of the binary regression with a *linear* predictor have been studied by Fan et al. (1995) and Signorini and Jones (2004), who provided asymptotic properties of various versions of the estimators. The discussed methods essentially rely on local polynomial regression and a local likelihood approach. In this paper we discuss both local polynomial and local likelihood techniques to binary regression estimation with directional predictors. A local-likelihood-

based approach has been also investigated in Di Marzio et al. (2017) in the different context of estimation of densities defined on the d-dimensional torus.

The paper is organized as follows. In Section 2 we recall a Taylor-like polynomial to approximate functions having the unit hypersphere as their domain. In Section 3 we discuss the adaptation of the theory of local polynomial regression with a directional predictor to the binary response case, while, in Section 4 we propose the nonparametric estimation using a locally weighted likelihood objective function. Finally, Section 5 collects some simulation examples and a real-data application.

2. Series expansion for functions on the sphere

Let $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}|| = 1\}$ denote the unit hypersphere embedded in $\mathbb{R}^d$, $d \geq 2$. The *tangent-normal* decomposition provides a possible parametrization of a point on $\mathbb{S}^{d-1}$. Specifically, for fixed $\boldsymbol{x} \in \mathbb{S}^{d-1}$, according to the tangent-normal decomposition, any vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$ can be expressed as

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

where θ is the angle between $\boldsymbol{u}$ and $\boldsymbol{x}$, and $\boldsymbol{\xi}$ is a unit vector orthogonal to $\boldsymbol{x}$. Now, letting μ_d denote the Lebesgue measure of $\mathbb{S}^d$, with

$$\omega_d = \mu_d\left(\mathbb{S}^d\right) = \frac{2\pi^{(d+1)/2}}{\Gamma((d+1)/2)},$$

and setting $\mathbb{T}_{\boldsymbol{x}} = \{\boldsymbol{\xi} \in \mathbb{S}^{d-1} : \boldsymbol{\xi} \perp \boldsymbol{x}\}$, for a real-valued function g defined on $\mathbb{S}^{d-1}$, the integration formula corresponding to the above parametrization is

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

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

$\{\mathbf{0}_d\}$, with $\mathbf{0}_d$ being the d-dimensional zero vector, we have that

$$\frac{\partial^\ell}{\partial\theta^\ell} g(\boldsymbol{u}(\boldsymbol{\xi},\theta))\bigg|_{\theta=0} = \mathscr{D}^{(\ell)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}),$$

where $\mathscr{D}^{(\ell)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x})$ is the *directional derivative* of order ℓ of $\bar{g}$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{\xi}$. Clearly $\mathscr{D}^{(0)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) = g(\boldsymbol{x})$, while, letting $\nabla^\ell_{\bar{g}}(\boldsymbol{x})$ be the matrix of the derivatives of total order ℓ of $\bar{g}$ at $\boldsymbol{x}$, one has

$$\mathscr{D}^{(\ell)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}'\nabla^\ell_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi}^{\otimes(\ell-1)},$$

where $\boldsymbol{a}^{\otimes\ell}$ stands for the Kroneckerian power of order ℓ of a vector $\boldsymbol{a}$. Then, for example, we have $\mathscr{D}^{(1)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}'\nabla_{\bar{g}}(\boldsymbol{x})$ and $\mathscr{D}^{(2)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}'\nabla^2_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi}$, with $\nabla^1_{\bar{g}}(\boldsymbol{x})$ and $\nabla^2_{\bar{g}}(\boldsymbol{x})$ respectively being the gradient vector and the Hessian matrix of $\bar{g}$ at $\boldsymbol{x}$, while $\mathscr{D}^{(3)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}'\nabla^3_{\bar{g}}(\boldsymbol{x})\boldsymbol{\xi}\otimes\boldsymbol{\xi}$, with $\boldsymbol{a}\otimes\boldsymbol{a}$ being the Kroneckerian product of the vector $\boldsymbol{a}$ by itself.

Now, under suitable continuity assumptions, a Taylor-like expansion of a real valued function g defined on $\mathbb{S}^{d-1}$ can be provided. Specifically, by assuming the continuity of $\nabla^\ell_{\bar{g}}(\boldsymbol{x})$, $\boldsymbol{x}\in\mathbb{S}^{d-1}$, for $\ell\in(1,\ldots,p)$, a pth-order series expansion of g around $\boldsymbol{x}$ yields

$$\begin{aligned} g(\boldsymbol{u}) &\approx g(\boldsymbol{x}) + \sum_{\ell=1}^{p}\frac{\theta^\ell}{\ell!}\mathscr{D}^{(\ell)}_{\boldsymbol{\xi}}\bar{g}(\boldsymbol{x}) \\ &= g(\boldsymbol{x}) + \sum_{\ell=1}^{p}\frac{\theta^\ell}{\ell!}\boldsymbol{\xi}'\nabla^\ell_{\bar{g}(\boldsymbol{x})}\boldsymbol{\xi}^{\otimes(\ell-1)}. \end{aligned} \tag{2}$$

The above expansion has been employed for deriving the asymptotic properties of kernel estimators for spherical densities by Hall et al. (1987) and Klemela (2000), to obtain a component-wise local approximation of spherical-spherical regression by Di Marzio et al. (2014), and to approximate the entries of skew-symmetric matrices and define rotations for spherical regression by Di Marzio et al. (2018a).

3. Local polynomial binary regression

Let $(\boldsymbol{X}, Y)$ be a $\mathbb{S}^{d-1} \times \{0,1\}$-valued random variable, and set $\lambda(\boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. If independent copies $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ of $(\boldsymbol{X}, Y)$ are observed, by ignoring the binary nature of $\boldsymbol{Y}$, a naive nonparametric estimation of $\lambda(\boldsymbol{x})$ can be performed by using the local polynomial estimators with real-valued response and spherical predictor, which have been studied by Di Marzio et al. (2014).

In particular, following this approach, the regression function at $\boldsymbol{X}_i$ is approximated by a suitable pth degree polynomial around $\boldsymbol{x} \in \mathbb{S}^{d-1}$, and a local estimator of $\lambda(\boldsymbol{x})$ is defined as the solution (for the zero order coefficient) of the minimization of a weighted L_2 distance between the Y_is and the approximating polynomial. Different values of p give different estimators. Formally, by using expansion (2), a pth degree local polynomial estimator of λ at $\boldsymbol{x} \in \mathbb{S}^{d-1}$, say $\hat{\lambda}(\boldsymbol{x}; p)$, can be defined as the solution for β_0 of

$$\underset{\{\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \sum_{\ell=1}^{p} \frac{\theta_i^\ell}{\ell} \boldsymbol{\xi}_i' \boldsymbol{\beta}_\ell \boldsymbol{\xi}_i^{\otimes(\ell-1)} \right\}^2 K_\kappa(\boldsymbol{x}' \boldsymbol{X}_i), \tag{3}$$

where θ_i is the angle between $\boldsymbol{X}_i$ and $\boldsymbol{x}$, and the weight K_κ is a *spherical kernel*. A spherical kernel can be essentially defined as a unimodal density having $\mathbb{S}^{d-1}$ as its support, with rotational symmetry about its mean direction $\boldsymbol{\mu} = (0, \ldots, 0, 1)$, and concentration parameter $\kappa > 0$ such that as κ increases K_κ concentrates around $\boldsymbol{\mu}$. In equation (3) the weight function emphasizes the contribution of the observations $\boldsymbol{X}_i$s which are *closer* to the estimation point $\boldsymbol{x}$. Kernels of this form have been used by Hall et al. (1987) for density estimation on the sphere and by Di Marzio et al. (2014) and Di Marzio et al. (2018a) for spherical-spherical regression estimation.

Now, when $p = 0$, the solution for β_0 leads to the *local constant* estimator

$$\hat{\lambda}(\boldsymbol{x}; 0) = \frac{\sum_{i=1}^{n} Y_i K_\kappa(\boldsymbol{x}' \boldsymbol{X}_i)}{\sum_{i=1}^{n} K_\kappa(\boldsymbol{x}' \boldsymbol{X}_i)}, \tag{4}$$

while, when $p = 1$, the unique solution for β_0 of the above least squares problem under

a suitable constraint (see Di Marzio et al. (2014) for details) can be expressed as

$$\hat{\lambda}(\boldsymbol{x};1) = \sum_{i=1}^{n} W_{\kappa}(\boldsymbol{x}'\boldsymbol{X}_i)Y_i,$$

where

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

Now, in order to discuss the asymptotic properties of the estimators, we need to recall the *spherical counterparts* of the jth moment, $j \in \mathbb{N}$, and the roughness of a Euclidean kernel, which, for a kernel K_{κ}, respectively are

$$b_j(\kappa) = \omega_{d-2} \int_0^{\pi} K_{\kappa}(\cos(\theta))\theta^j \sin^{d-2}(\theta) d\theta,$$

and

$$v_0(\kappa) = \omega_{d-2} \int_0^{\pi} K_{\kappa}^2(\cos(\theta)) \sin^{d-2}(\theta) d\theta.$$

Let $\mathrm{Tr}(\boldsymbol{A})$ denote the trace of the matrix $\boldsymbol{A}$, and use f to denote the common density of the $\boldsymbol{X}_i$s. Then, for the cases $p = 0$ and $p = 1$, by respectively using results in Theorem 1 and Theorem 2 in Di Marzio et al. (2014), we obtain the following

Result 1. *Given the $\mathbb{S}^{d-1} \times \{0,1\}$-valued random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, consider estimator $\hat{\lambda}(\boldsymbol{x}; p)$, $\boldsymbol{x} \in \mathbb{S}^{d-1}$. If*

i) K_{κ} is a spherical kernel such that as n increases $b_2(\kappa)$ and $v_0(\kappa)/n$ both go to 0, and for $j > 2$, $b_j(K_{\kappa}) = o(b_2(\kappa))$;

ii) $f(\boldsymbol{x}) > 0$ and all the entries of $\nabla_{\bar{f}}(\boldsymbol{x})$, $\nabla_{\bar{\lambda}}(\boldsymbol{x})$, and $\nabla^2_{\bar{\lambda}}(\boldsymbol{x})$ are continuous,

then

$$\mathsf{E}[\hat{\lambda}(\boldsymbol{x};0)] - \lambda(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)} \left(\mathsf{Tr}\left\{ \nabla^2_{\bar{\lambda}}(\boldsymbol{x}) \right\} + \frac{2\nabla'_{\bar{\lambda}}(\boldsymbol{x})\nabla_{\bar{f}}(\boldsymbol{x})}{f(\boldsymbol{x})} \right) + o\left(b_2(\kappa)\right),$$

$$\mathsf{E}[\hat{\lambda}(\boldsymbol{x};1)] - \lambda(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)} \mathsf{Tr}\left\{ \nabla^2_{\bar{\lambda}}(\boldsymbol{x}) \right\} + o\left(b_2(\kappa)\right),$$

and, for both $p = 0$ and $p = 1$,

$$\mathsf{Var}[\hat{\lambda}(\boldsymbol{x};p)] = \frac{\nu_0(\kappa)}{n}\frac{\lambda(\boldsymbol{x})(1-\lambda(\boldsymbol{x}))}{f(\boldsymbol{x})} + o\left(\frac{\nu_0(\kappa)}{n}\right).$$

Remark 1. *Recently, García-Portugués et al. (2016) proposed a different series expansion of the regression function with* linear *response and directional predictor, which generalizes the proposal of Di Marzio et al. (2009) in the circular case when $p = 1$. The optimization of the corresponding L_2 loss leads to a projected local linear estimator which shares the asymptotic properties of the local linear estimator of Di Marzio et al. (2014).*

An optimal smoothing degree would minimize the asymptotic mean-squared error of $\hat{\lambda}(\boldsymbol{x};p)$, which is the sum of the leading terms of the asymptotic squared bias and the asymptotic variance. Notice that the dependence of asymptotic bias and variance on the concentration parameter cannot be generalized with respect to the kernel, because it is not a scale factor.

For the important case of a von Mises-Fisher kernel (which can be regarded as the spherical counterpart of the Gaussian kernel), and is defined on $\mathbb{S}^{d-1}$ as

$$K_\kappa(\boldsymbol{x}'\boldsymbol{\mu}) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathscr{I}_{d/2-1}(\kappa)}\exp(\kappa\boldsymbol{x}'\boldsymbol{\mu}),$$

with $\mathscr{I}_u(\cdot)$ being the modified Bessel function of the first kind and order u, when κ is big enough, and $j \in \mathbb{N}$, the following approximations of $b_j(\kappa)$ and $\nu_0(\kappa)$ hold

$$b_j(\kappa) \sim \frac{2^{j/2}\Gamma((d+j-1)/2)}{\kappa^{j/2}\Gamma((d-1)/2)}, \quad \text{and} \quad \nu_0(\kappa) \sim \frac{\kappa^{(d-1)/2}}{2^{d-1}\pi^{(d-1)/2}}. \tag{5}$$

As a consequence, when K_κ is a von Mises-Fisher kernel, the asymptotic bias and the asymptotic variance, for both $p = 0$ and $p = 1$, are

$$\mathsf{E}[\hat{\lambda}(\boldsymbol{x};p)] - \lambda(\boldsymbol{x}) = O\left(\frac{1}{\kappa}\right), \quad \text{and} \quad \mathsf{Var}[\hat{\lambda}(\boldsymbol{x};p)] = O\left(\frac{\kappa^{(d-1)/2}}{n}\right).$$

Then, in the case of a von Mises-Fisher kernel, for both local constant and local linear estimators, the value of κ which minimizes the asymptotic mean squared error is

$O(n^{2/(d+3)})$ and gives a convergence rate of magnitude $O(n^{-4/(d+3)})$. This is the same rate attained by single bandwidth local constant and local linear estimators of a real-valued regression function defined on $\mathbb{R}^{d-1}$, when a second-order kernel is employed.

4. Local logistic regression via likelihood

The approach discussed in the previous section does not produce *bona-fide* estimates when the polynomial degree is greater than 0. Despite the fact that a truncation could be used for exploratory data analysis, the subsequent lack of differentiability may be a serious issue. To take into account the binary nature of the response, one should consider the estimator as the optimiser of a more suited objective function, such as the log-likelihood one, instead of the least squares in (3).

Specifically, given the $\mathbb{S}^{d-1} \times \{0,1\}$-valued random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, the log-likelihood connected with the binary regression is

$$\sum_{i=1}^{n} \{Y_i \log(\lambda(\boldsymbol{X}_i)) + (1 - Y_i) \log(1 - \lambda(\boldsymbol{X}_i))\}.$$

The locally weighted version, at $\boldsymbol{x} \in \mathbb{S}^{d-1}$, of the above log-likelihood can be expressed as

$$\sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{\lambda(\boldsymbol{X}_i)}{1 - \lambda(\boldsymbol{X}_i)} \right) + \log(1 - \lambda(\boldsymbol{X}_i)) \right\} K_\kappa(\boldsymbol{x}'\boldsymbol{X}_i),$$

where $K_\kappa(\boldsymbol{x}'\boldsymbol{X}_i)$ is a spherical kernel with mean direction $\boldsymbol{X}_i$, and evaluated at $\boldsymbol{x}$. Setting $\delta = \log(\lambda/(1-\lambda))$, the above expression can be re-written as

$$\sum_{i=1}^{n} \{Y_i \delta(\boldsymbol{X}_i) - \log(1 + \exp(\delta(\boldsymbol{X}_i)))\} K_\kappa(\boldsymbol{x}'\boldsymbol{X}_i),$$

and, approximating $\delta(\boldsymbol{X}_i)$ around $\boldsymbol{x}$ in the local log-likelihood function by using expansion (2), a class of nonparametric estimators for $\lambda(\boldsymbol{x})$ can be obtained. Specifically, let $\beta_0 = \delta(\boldsymbol{x})$, and let $\boldsymbol{\beta}_\ell$ be the matrix of the derivatives of total order $\ell \in (1, \ldots, p)$

of $\bar{\delta}$ at $\boldsymbol{x}$. Then, for fixed $\boldsymbol{x} \in \mathbb{S}^{d-1}$, by expressing $\boldsymbol{X}_i$ according to the tangent normal decomposition, we define

$$q_p(\boldsymbol{X}_i; \beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) = \beta_0 + \sum_{\ell=1}^{p} \frac{\theta_i^\ell}{\ell!} \boldsymbol{\xi}_i' \boldsymbol{\beta}_\ell \boldsymbol{\xi}_i^{\otimes(\ell-1)}.$$

Hence, under suitable smoothness assumptions, the p-degree expansion of the log-likelihood can be expressed as

$$\sum_{i=1}^{n} \left\{ Y_i q_p(\boldsymbol{X}_i; \beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) - \log(1 + \exp(q_p(\boldsymbol{X}_i; \beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p))) \right\} K_\kappa(\boldsymbol{X}_i' \boldsymbol{x}). \quad (6)$$

It is interesting to note that, when κ goes to 0, the kernel $K_\kappa(\boldsymbol{x}' \boldsymbol{X}_i)$ approaches the uniform density and assigns the same weight to each sample point, for any $\boldsymbol{x}$. As a consequence, for κ going to 0, the local log-likelihood optimization reduces to the standard logistic regression problem with spherical predictor.

Now, letting $\hat{\beta}_0$ be the solution for β_0 of the maximization of (6) with respect to $\{\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\}$, a p-degree local polynomial estimator for $\lambda(\boldsymbol{x})$ is

$$\hat{\lambda}_L(\boldsymbol{x}; p) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)}.$$

When $p = 0$, the resulting estimator is the *local constant* one previously discussed, while, when $p = 1$, we obtain the spherical version of the *local linear* logistic estimator studied in the Euclidean setting by Fan et al. (1995) and Signorini and Jones (2004). A closed-form expression for $\hat{\lambda}_L(\boldsymbol{x}; 1)$ does not exist, but, obviously, distinctly from $\hat{\lambda}(\boldsymbol{x}; 1)$, the estimator always takes value on $[0, 1]$.

Concerning the asymptotic properties, by reasoning as in Theorem 3 and Theorem 4 of Fan et al. (1995) with g being the logit link, and by using Result 1, we get the following

Result 2. *Given a $\mathbb{S}^{d-1} \times \{0,1\}$-valued random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, consider the estimator $\hat{\lambda}_L(\boldsymbol{x}; 1)$, $\boldsymbol{x} \in \mathbb{S}^{d-1}$. If assumption i) and assumption ii) of Result 1*

hold, then

$$\mathsf{E}[\hat{\lambda}_L(\boldsymbol{x};1)] - \lambda(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)}\mathsf{Tr}\left\{\nabla^2_{\bar{\delta}}(\boldsymbol{x})\right\}\lambda(\boldsymbol{x})(1-\lambda(\boldsymbol{x})) + o\left(b_2(\kappa)\right),$$

and

$$\mathsf{Var}[\hat{\lambda}_L(\boldsymbol{x};1)] = \frac{v_0(\kappa)}{n}\frac{\lambda(\boldsymbol{x})(1-\lambda(\boldsymbol{x}))}{f(\boldsymbol{x})} + o\left(\frac{v_0(\kappa)}{n}\right).$$

Notice that $\hat{\lambda}_L(\boldsymbol{x};1)$ shares both the asymptotic variance and the order of the asymptotic bias of $\hat{\lambda}(\boldsymbol{x};p)$, $p \in (0,1)$. Moreover, the asymptotic bias depends only on λ and the derivatives of $\bar{\lambda}$ but not on f, as it happens for $\hat{\lambda}(\boldsymbol{x};1)$.

Clearly, by virtue of Result 2, if a von Mises-Fisher kernel is employed as the weight, by recalling the approximations in (3), the estimator attains the convergence rate of order $n^{-4/(d+3)}$.

Concerning the selection of κ, a possible way is to start from a least-squares objective function, and choose the value of κ which minimizes

$$\sum_{i=1}^{n}\left(Y_i - \hat{\lambda}_{-i}(\boldsymbol{X}_i;\kappa)\right)^2,$$

where $\hat{\lambda}_{-i}(\boldsymbol{X}_i;\kappa)$ stands for the estimate of λ at $\boldsymbol{X}_i$ with the ith sample observation removed. A more natural way is to start from the leave-one-out version of the local log-likelihood, i.e. to select the value of κ maximizing

$$\sum_{i=1}^{n}\left\{Y_i \log\left(\frac{\hat{\lambda}_{-i}(\boldsymbol{X}_i;\kappa)}{1-\hat{\lambda}_{-i}(\boldsymbol{X}_i;\kappa)}\right) + \log\left(1-\hat{\lambda}_{-i}(\boldsymbol{X}_i;\kappa)\right)\right\}. \tag{7}$$

Remark 2. *A possible generalization of the discussed approach arises from considering different weights for* successes *and* failures *in the local log-likelihood expression, i.e.*

$$\sum_{i=1}^{n} Y_i q_p\left(\boldsymbol{X}_i;\beta_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p\right)K_{\kappa_1}(\boldsymbol{x}'\boldsymbol{X}_i) - \log\left(1+\exp(q_p(\boldsymbol{X}_i;\beta_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p))\right)K_{\kappa_2}(\boldsymbol{x}'\boldsymbol{X}_i),$$

with K_{κ_1} and K_{κ_2} being spherical kernels giving weight to the observations of the predictor corresponding to $Y=1$ and $Y=0$, respectively.

When $p=0$, the solution for β_0 of the maximization of the above local log-likelihood function gives the estimator of $\lambda(\boldsymbol{x})$ studied by Di Marzio et al. (2018b). This latter

is defined by using the kernel estimators, say $\hat{f}_1(\boldsymbol{x};\kappa_1)$ and $\hat{f}_2(\boldsymbol{x};\kappa_2)$, of the spherical densities f_1 and f_2 respectively characterizing the distributions of the predictor in the spaces of successes and failures, based on random samples of respective sizes n_1 and n_2, i.e.

$$\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2) = \frac{n_1\hat{f}_1(\boldsymbol{x};\kappa_1)}{n_1\hat{f}_1(\boldsymbol{x};\kappa_1)+n_2\hat{f}_2(\boldsymbol{x};\kappa_2)}. \tag{8}$$

5. Numerical examples

5.1. Simulation

In this section we use simulation experiments to test the performance of the proposed estimator for classification tasks. In particular, we consider the problem of assigning label 0 or 1 to an observation $\boldsymbol{x} \in \mathbb{S}^{d-1}$. To this end, we adopt the rule according to which $\boldsymbol{x}$ is assigned to the population with label 1 if the estimate of $\lambda(\boldsymbol{x})$ is greater or equal to 0.5.

We use $vMF(\boldsymbol{\mu},\gamma)$ to denote the von Mises-Fisher distribution on $\mathbb{S}^2$ with mean direction $\boldsymbol{\mu}$ (polar co-ordinates expressed in degrees) and concentration parameter γ. We consider different experiments using the following scenarios, where samples of sizes $n_1 = n_2 = 200$ are respectively drawn from $vMF(\boldsymbol{\mu}_1,\gamma_1)$ and $vMF(\boldsymbol{\mu}_2,\gamma_2)$:

Scenario 1: $\boldsymbol{\mu}_1 = (270,20)$, $\boldsymbol{\mu}_2 = (270,-20)$ and $\gamma_1 = \gamma_2 = 10$;

Scenario 2: $\boldsymbol{\mu}_1 = (270,20)$, $\boldsymbol{\mu}_2 = (270,-20)$ and $\gamma_1 = \gamma_2 = 20$;

Scenario 3: $\boldsymbol{\mu}_1 = (270,20)$, $\boldsymbol{\mu}_2 = (220,-20)$, $\gamma_1 = 5$ and $\gamma_2 = 10$.

In Scenario 1 the populations, which share the longitude of the mean direction and the value of the concentration parameter, generate rather overlapping groups. Scenario 2 refers to more concentrated populations generating more separated groups. Finally, in Scenario 3 two well-separated groups are generated by populations with different co-ordinates of the mean directions and different concentrations.

In the first experiment we consider the estimator (8) with K_{κ_1} and K_{κ_2} both being von Mises-Fisher kernels. The smoothing degrees are selected using the von Mises-Fisher reference rule (see, Di Marzio et al. (2018b) for details). In a second experiment

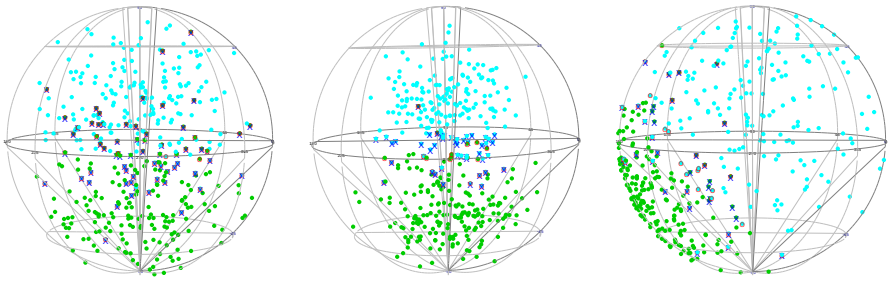

Figure 1: From left: Misclassified observations using KDE classification (marked by 'X') and using local likelihood with $p = 0$ (marked by 'o') for one dataset drawn from $vMF(\boldsymbol{\mu_1}, \gamma_1)$ (cyan points) and $vMF(\boldsymbol{\mu_2}, \gamma_2)$ (green points) in scenarios 1, 2 and 3.

we consider estimator (4). Also in this case we use the von Mises-Fisher kernel as the weight, by selecting the concentration parameter by least squares cross-validation. Figure 1 illustrates the misclassified observations obtained according to the rule for estimators (8) and (4) by using one dataset for each of the described scenarios. In Table 1, for each experiment, we report as the accuracy measure the average misclassification rate over 200 simulated datasets. The results show that the binary regression estimator slightly outperforms the kernel density classifier (KDE), especially when the groups are well-separated. Moreover, the results for $n_1 = n_2$ show that, in the considered scenarios, estimator (4) performs slightly better than the same estimator using two concentration parameters (which leads to the same classification rule as the kernel density one).

Table 1: Estimate of the misclassification rates for kernel density classification and local binary regression with $p = 0$, using 200 samples of sizes $n_1 = n_2 = 200$ respectively drawn from $f_j = vMF(\boldsymbol{\mu}_j, \gamma_j), j \in (1,2)$, given in scenarios 1–3. For both classification rules we use a von Mises-Fisher kernel: for KDE, κ_1 and κ_2 are selected according to the von Mises-Fisher reference rule, and for the local binary regression estimator κ is selected by least squares cross validation.

Classification rule	Misclassification rate		
	Scenario 1	Scenario 2	Scenario 3
KDE estimator	0.178	0.090	0.112
$p = 0$ estimator	0.147	0.065	0.086

5.2. Handwritten digit recognition

We apply our methods to the digits dataset used in the StatLog project (Michie et al., 1994). The dataset consists of 18,000 examples of the digits 0 to 9 (i.e. $q = 10$ classes) extracted from hand-written postcodes in Germany. These numbers were initially digitised onto 16×16 images with 256 grey levels; examples are shown in Figure 2. To enable meaningful comparisons with previously obtained results, we have used the same train-test split of the data which has 900 examples of each number (0–9) in the training set and the test set, and an averaging over 4×4 pixels resulting in 16 real-valued variables. These data were then transformed to the unit sphere by simply normalizing each observation replacing $\boldsymbol{X}_i$ by $\boldsymbol{X}_i / ||\boldsymbol{X}_i||$.

Figure 2: Examples of 10 handwritten, digitised digits with resolution 16×16 and 256 grey scales, extracted from postcodes in Germany (Michie et al., 1994).

Our implementation, which corresponds to a 1-degree local polyomial estimator, used logistic regression with weights obtained from a spherical kernel. The smoothing parameter was selected — for each pair of classes $(j,k) \in \{0,1,\ldots,9\} \times \{0,1,\ldots,9\}$ —using cross-validation (i.e. Equation (7)), which yielded solutions for the smoothing parameter ranging from 0.9 to 38.8. Then, for each element of the test set, we compute the probability of membership of class j, given an alternative of class k, say P_{jk} with $P_{jk} = 1 - P_{kj}$ (also setting $P_{jj} = 1$), using the corresponding $\kappa_{jk} (= \kappa_{kj})$ found by cross-validation. Finally, we allocate this observation to the class $\text{argmax}_k \min_j P_{jk}$. The error rate for 9000 observations in the test set, was 0.043, which is much better than the unweighted multinomial logistic regression (error 0.086), a simple linear discriminant (0.114) and just better than the top rank classifier (k-nearest neighbour, with an error rate of 0.047) of those given in (Michie et al., 1994, p. 136). The confusion matrix for this classifier is shown in Figure 3, in which it can be seen that the most common

classification mistakes were to recognize an ”8” as a “0”, a “2” as a “3”, and a ”7” as a ”9”.

	true digit									
	0	1	2	3	4	5	6	7	8	9
0	853	4	2	0	5	2	2	2	8	2
1	4	886	3	3	1	1	0	6	0	8
2	0	0	858	22	1	3	3	1	16	1
3	0	0	11	839	0	5	0	2	7	14
4	4	1	2	0	876	1	4	4	0	4
5	0	0	7	18	1	868	5	0	11	4
6	8	0	0	0	3	8	880	0	5	0
7	2	5	0	2	3	0	0	870	1	22
8	21	2	12	9	1	5	6	4	845	11
9	8	2	5	7	9	7	0	11	7	834

Figure 3: Confusion matrix for local multinomial logistic classifier applied to German handwritten postcode digits. Columns represent true label, and rows the predicted label.

Although the error rate is very good, we note that this approach was computationally intensive, with the multinomial logistic model entailing the estimation of $q(q-1)/2 = 45$ smoothing parameters in the training phase, and a further fitting of $nq(q-1)/2 = 405{,}000$ models in the testing phase. Whilst it would be straightforward to consider $p = 2$ (including interaction terms, if desired) this would take an excessive amount of time without a common choice of κ across all class pairs.

Using a classification rule based on a kernel density estimator, a single smoothing parameter (for all classes) was selected by leave-one-out cross-validation on the training data. This value of κ ($= 140.6$) was then used to classify the test data. For this classifer, the error rate of 0.039 was unexpectedly somewhat better than the result given in Michie et al. (1994) (0.068) for data which have not been transformed to the sphere.

References

Batschelet, E., 1981. Circular Statistics in Biology. Academic Press, New York, NY.

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C., 2017. Circular local likelihood. TEST 21, 1863–8260.

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C., 2018. Kernel density classification for spherical data, submitted.

Di Marzio, M., Panzera, A., Taylor, C.C., 2009. Local polynomial regression for circular predictors. Statistics & Probability Letters 79, 2066–2075.

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

Di Marzio, M., Panzera, A., Taylor, C.C., 2018. Nonparametric rotations for sphere-sphere regression. Journal of the American Statistical Association, forthcoming.

Fan, J., Heckman, N., Wand, M.P., 1995. Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. Journal of the American Statistical Association 90, 141–150.

Fernandes, K., Cardoso, J.S., 2016. Discriminative directional classifiers. Neurocomputing 207, 141–149.

Fisher, N., Lewis, T., Embleton, B., 1987. Statistical Analysis of Spherical Data. Cambridge University Press.

García-Portugués, E., Van Keilegom, I., Crujeiras, R.M., González-Manteiga, W., 2016. Testing parametric models in linear-directional regression. Scandinavian Journal of Statistics 43, 1178–1191.

Hall, P., Watson, G., Cabrera, J., 1987. Kernel Density Estimation with Spherical Data. Biometrika 74, 751–762.

Hamsici, O.C., 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.

Klemela, J., 2000. Estimation of densities and derivatives of densities with directional data. Journal of Multivariate Analysis 73, 18–40.

Ley, C., Verdebout, T., 2017. Modern Directional Statistics. Chapman & Hall/CRC Press, Boca Raton, Florida.

Ley, C., Verdebout, T., 2018. Applied Directional Statistics: Modern Methods and Case Studies. Chapman & Hall/CRC Press, Boca Raton, Florida.

Mardia, K.V., Jupp, P.E., 2008. Directional Statistics. Chichester: J. Wiley.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (eds) 1994. Machine Learning, Neural and Statistical Classification. Chichester: Ellis Horwood.

Signorini, D.F., Jones, M.C., 2004. Kernel Estimators for Univariate Binary Regression. Journal of the American Statistical Association 99, 119–126.