Kernel density classification for spherical data

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

Classifying observations coming from two different spherical populations by using a nonparametric method appears to be an unexplored field, although clearly worth to pursue. We propose some decision rules based on spherical kernel density estimation and we provide asymptotic L_2 properties. A real-data application using global climate data is finally discussed.

Keywords: Classification, Directional data, Nonparametric methods.
2000 MSC: 62H11, 62H20

1. Introduction

Directional data arise in many scientific fields where observations are recorded as directions or angles relative to a fixed reference point. In general, the space of all directions is the unit hypersphere $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}|| = 1\}$, $d \geq 2$, that is a $(d-1)$-dimensional manifold embedded in $\mathbb{R}^d$. In particular, the special cases $d = 2$ and $d = 3$ respectively refer to the unit circle and the *ordinary* sphere. Classical examples of such data include directions of winds, marine currents, Earth's main magnetic field, rock fractures. Because of the nonlinear nature of the manifold, all the statistical methods for dealing with directional data require to be adapted. For comprehensive accounts of statistics for directional data see, for example, [15] and [12].

*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)

There are several situations in which it is necessary to classify directional data, i.e. to predict the label denoting the class of a new point on $\mathbb{S}^{d-1}$, given a training sample with known labels. Possible examples include: the identification of the presence or absence of cardiac arrhythmia using angles from the front plane of electrocardiogram waves, discrimination between closed and open fishing zones using the latitude and the longitude of the sampling locations, classification of megaspores into groups in the biological taxonomy according to the angle of their wall elements.

Concerning the classification of directional data there are several proposals in the literature, almost all having a parametric flavour. [18] and [16] introduced discriminant analysis for directional data using the von Mises-Fisher distribution. [13] extended the naive Bayes classifier to the case where the model of the conditional probability is assumed to be von Mises-Fisher, also presenting several real data examples. The Bayesian discriminant rule is discussed for different classes of spherical populations by [4]. Coming to nonparametric approaches, the only contribution seems to be provided by [14]. The authors introduced a kernel classification rule for data lying on a general compact Riemannian manifold and proved its strong consistency. The practical implementation of the rule requires the knowledge of the specific manifold geometry, and the authors deferred it to future research. Kernel density estimation is commonly used for classification tasks in the Euclidean setting, where asymptotic L_2 properties and some strategies to select the smoothing degree have been proposed by [3] and [9].

In this paper we extend nonparametric approaches based on density estimation to classify data lying on $\mathbb{S}^{d-1}$. The paper is organized as follows. Section 2 introduces some basic facts about the parametrization of functions defined on $\mathbb{S}^{d-1}$, as well as some notation. Section 3 introduces the basic tool of spherical density estimation. Section 4 presents main ideas of kernel density classification. Section 5 discusses a more general classification approach which incorporates a priori information. Finally, Section 6 presents a real-data application.

2. Preliminaries

When dealing with a non-linear manifold, such as the d-dimensional sphere, specific series expansions and integration formulas are needed. Primarily, we need to suitably represent a sphere location as follows. Given a fixed $\boldsymbol{x} \in \mathbb{S}^{d-1}$, any vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$ can be expressed according to the *tangent-normal* decomposition as $\boldsymbol{u}(\boldsymbol{\xi}, \theta) = \boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta)$, where $\theta \in (0, \pi)$ is the angle between $\boldsymbol{u}$ and $\boldsymbol{x}$, and $\boldsymbol{\xi}$ is a vector orthogonal to $\boldsymbol{x}$. Now, approximations of functions defined on a sphere will be obtained by a McLaurin expansion of $\sin(\theta)$ and $\cos(\theta)$.

Denoting as μ_d the Lebesgue measure of $\mathbb{S}^d$, with $\mu_d\left(\mathbb{S}^d\right) = \omega_d = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$ interpreted as the surface area of the d-dimensional unit sphere, for a density function f on $\mathbb{S}^{d-1}$ we have $\int_{\mathbb{S}^{d-1}} f(\boldsymbol{u}) d\mu_{d-1}(\boldsymbol{u}) = 1$. Concerning integration formulas, letting $\mathbb{T}_{\boldsymbol{x}} = \{\boldsymbol{\xi} \in \mathbb{S}^{d-1} : \boldsymbol{\xi} \perp \boldsymbol{x}\}$, for a generic function $g : \mathbb{S}^{d-1} \to \mathbb{R}$ we write

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

A *spherical kernel* K_κ is a spherical probability density function with mean direction $\boldsymbol{\mu} = (0, \ldots, 0, 1)$ and concentration parameter $\kappa > 0$ such that: *i)* it is unimodal; *ii)* it is rotationally symmetric about $\boldsymbol{\mu}$. Formally, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$ and $\theta = \arccos(\boldsymbol{x}'\boldsymbol{\mu})$, we write $\boldsymbol{x} = \sin(\theta)\boldsymbol{\mu} + \cos(\theta)\boldsymbol{\xi}$, and require that the conditional distribution of $\boldsymbol{\xi} \,|\, \sin(\theta)$ is uniform on $\mathbb{S}^{d-2}$; *iii)* it is able to arbitrarily concentrate around $\boldsymbol{\mu}$, i.e., for any $W \subset \mathbb{S}^{d-1} \setminus \{\boldsymbol{\mu}\}$, $\lim_{\kappa \to \infty} \int_W K_\kappa(\boldsymbol{x}'\boldsymbol{\mu}) \omega_{d-1}(d\boldsymbol{x}) = 0$.

Observe that differently from the *linear* case where the bandwidth is a scale parameter, the concentration parameter of a spherical kernel is typically not a scale factor, and this affects the technical treatment of the directional setting.

A classical example of a spherical kernel is the von Mises-Fisher density, which is defined on $\mathbb{S}^{d-1}$ as $K_\kappa(\boldsymbol{x}'\boldsymbol{\mu}) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathscr{I}_{d/2-1}(\kappa)} \exp(\kappa \boldsymbol{x}'\boldsymbol{\mu})$, where $\mathscr{I}_u(\cdot)$ stands for the modified Bessel function of the first kind and order u. In the following we will denote by $vMF(\boldsymbol{\mu}, \kappa)$ the von Mises-Fisher distribution, with mean direction $\boldsymbol{\mu}$ and concentration parameter κ.

Remark 1. *Other than spherical densities, [10] consider kernels of the form* $L\left(\frac{1-\boldsymbol{x}'\boldsymbol{\mu}}{h^2}\right)$, *where L is typically 'half' of a bell-shaped function, such as* $\exp^{-x^2}$, *or the uniform*

density on $(0,1)$*, and* $h > 0$ *is the bandwidth parameter. Such kernels, which do not necessarily integrate to one, are reminiscent of the Euclidean ones due to the presence of a proper bandwidth h. Some results on asymptotic equivalence between the density estimators deriving from these two classes of kernels are also provided. These kernels have been used by [1] and, more recently, by [5], [6], [7] and [8].*

We introduce two quantities that will be useful to describe the asymptotic mean and variance of the kernel density estimator. For $j \in \mathbb{N}$, and a spherical kernel K_κ, set

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

$$\nu_0(\kappa) = \omega_{d-2} \int_0^\pi K_\kappa^2(\cos(\theta)) \sin^{d-2}(\theta) d\theta. \tag{3}$$

The non-negative quantity $b_j(\kappa)$ resembles the jth moment of a Euclidean kernel, with $b_0(\kappa) = 1$, while the quantity $\nu_0(\kappa)$ increases with the roughness of the shape of K_κ. Differently from the Euclidean setting, where the symmetry of the kernel implies that the moments of odd order vanish, the rotational symmetry of the spherical kernel K_κ does not make quantities $b_j(\kappa)$ null for odd j. However, these "odd moments"do not affect the asymptotic properties of the density estimators. In fact, in the expansions of convolutions involved in the derivation of the asymptotic properties, the quantity $b_j(\kappa)$ turns out to be multiplied by $\int_{\mathbb{T}_x} \boldsymbol{\xi}\boldsymbol{\xi}'^{\otimes(j-1)} d\mu_{d-2}(\boldsymbol{\xi})$, where $\boldsymbol{u}^{\otimes s}$ stands for the sth Kronekerian power of the vector $\boldsymbol{u}$, which is null for odd j. This makes the scenario comparable to the Euclidean setting.

3. Kernel density estimation

Before discussing its application to the classification task, we briefly recall some basic theory about the spherical kernel density estimation introduced by [10] and [1].

Given a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the unknown spherical density f, the kernel estimator of f at $\boldsymbol{x} \in \mathbb{S}^{d-1}$ is $\hat{f}(\boldsymbol{x}; \kappa) = n^{-1} \sum_{i=1}^n K_\kappa(\boldsymbol{x}'\boldsymbol{X}_i)$, where K_κ is a spherical kernel, with the mean value being the ith observation and the concentration parameter κ. A brief inspection of the above formula reveals that the estimate $\hat{f}(\boldsymbol{x}; \kappa)$ is built according to a sort of analogy principle: for fixed κ, it is as higher as more observations

are located in a neighbourhood of the estimation point $\boldsymbol{x}$. The kernel function has the rôle of tuning the contribution of the observations. For a given sample $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, a small value of κ entails low concentration around the sample observations resulting in a undersmooth density estimate, whereas large values of κ provide a rough estimate of f. Specifically, condition i) in Section 2 makes it possible to assign a decreasing weight when the distance between $\boldsymbol{x}$ and $\boldsymbol{X}_i$ increases. Condition ii) makes the weight dependent only on the distance and not on the specific arc joining $\boldsymbol{x}$ and $\boldsymbol{X}_i$. Condition iii) makes it possible to localize the estimate by arbitrarily reducing the number of observations effectively participating to the estimation process at $\boldsymbol{x}$. An example is illustrated in Figure 1, where, for a sample drawn from a von Mises-Fisher density defined on $\mathbb{S}^2$, we obtain two density estimates by using the same kernel function with different concentration parameters. Asymptotic properties of $\hat{f}(\boldsymbol{x};\kappa)$ has been derived

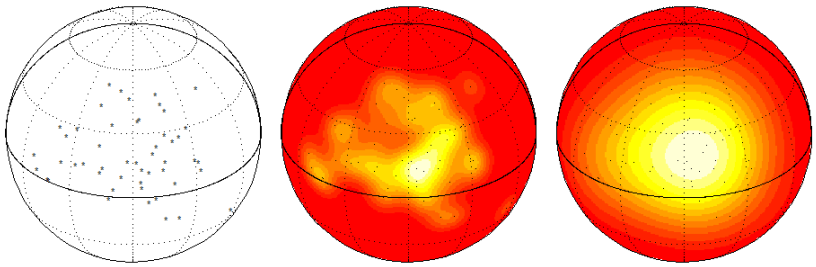

Figure 1: From left: Random sample, of size $n = 50$, drawn from a $vMF((0.40,-0.07,0.91),8)$, kernel density estimates with concentrations equal to 100 (undersmoothing), and 10 (correct smoothing).

by [10] and [11]. Since the concentration parameter of a spherical kernel is not a scale factor, the asymptotic assumptions on κ involve quantities (2) and (3). Indeed, we assume κ depending on n in such a way that, as n increases, i) $b_2(\kappa)$ and $n^{-1}v_0(\kappa)$ go to zero, and ii) $b_j(\kappa) = o(b_2(\kappa))$ for each $j > 2$. Assumption i) assures the consistency of the density estimator, while Assumption ii) enables us to retain, in the expansion of the convolution involved in the asymptotic bias calculation, terms up to the second order.

Now, for a function g defined on $\mathbb{S}^{d-1}$, let $\bar{g}(\boldsymbol{x}) = g(\boldsymbol{x}/||\boldsymbol{x}||)$ be its homogeneous extension to $\mathbb{R}^d \setminus \{\boldsymbol{0}_d\}$, with $\boldsymbol{0}_d$ being the d-dimensional zero vector, and let $\nabla^s_{\bar{g}}(\boldsymbol{x})$ be the matrix of the derivatives of total order s of $\bar{g}$ at $\boldsymbol{x}$. Then, assuming that all entries of

$\nabla^2_{\bar{f}}$ are continuous at $\boldsymbol{x}$, one has

$$\mathsf{E}[\hat{f}(\boldsymbol{x};\kappa)] - f(\boldsymbol{x}) = \frac{b_2(\kappa)}{2(d-1)} \mathsf{Tr}\left\{\nabla^2_{\bar{f}}(\boldsymbol{x})\right\} + o(b_2(\kappa)), \tag{4}$$

$$\mathsf{Var}[\hat{f}(\boldsymbol{x};\kappa)] = \frac{v_0(\kappa)}{n} f(\boldsymbol{x}) + o\left(\frac{v_0(\kappa)}{n}\right), \tag{5}$$

where $\mathsf{Tr}\{\boldsymbol{A}\}$ stands for the trace of a matrix $\boldsymbol{A}$. In the special case of the von Mises-Fisher kernel it holds that, for κ big enough, and $j \in \mathbb{Z}^+$,

$$b_j(\kappa) \sim \frac{2^{j/2}\Gamma((d+j-1)/2)}{\kappa^{j/2}\Gamma((d-1)/2)}, \quad \text{and} \quad v_0(\kappa) \sim \frac{\kappa^{(d-1)/2}}{2^{d-1}\pi^{(d-1)/2}}, \tag{6}$$

so the classical asymptotic bias-variance trade-off arises, and the asymptotic mean squared error of $\hat{f}(\boldsymbol{x};\kappa)$ is

$$\mathsf{AMSE}[\hat{f}(\boldsymbol{x};\kappa)] = \left(\frac{1}{2\kappa}\mathsf{Tr}\left\{\nabla^2_{\bar{f}}(\boldsymbol{x})\right\}\right)^2 + \frac{\kappa^{(d-1)/2} f(\boldsymbol{x})}{2^{d-1}\pi^{(d-1)/2} n}.$$

Consequently, the value of κ minimizing $\mathsf{AMSE}[\hat{f}(\boldsymbol{x};\kappa)]$ is $O(n^{2/(d+3)})$, assuring $\hat{f}(\boldsymbol{x};\kappa)$ of the convergence rate $n^{-4/(d+3)}$, which is the same as the single bandwidth kernel estimator for a density on $\mathbb{R}^{d-1}$, using a second-order kernel.

4. Kernel density classification

Kernel density estimation is commonly used for classification purposes. Consider the case of two populations, $\mathscr{P}_1$ and $\mathscr{P}_2$, respectively described by the unknown spherical densities f_1 and f_2. Given two random samples of sizes n_1 and n_2 respectively drawn from f_1 and f_2, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$, define the classifier $\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2) = \hat{f}_1(\boldsymbol{x};\kappa_1) - \hat{f}_2(\boldsymbol{x};\kappa_2)$, then an observation $\boldsymbol{x}$ will be given label 1 (i.e. $\boldsymbol{x}$ is allocated to the population described by f_1) if $\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2) \geq 0$. The linear counterpart of this rule has been widely studied, see for example [3] and [9].

Although above rule could appear a primarily exploratory tool, the correct selection of κ_1 and κ_2 requires minimizing a loss criterion like, for example, the expectation of squared errors. Unfortunately, a tractable expression for it can be only reached by recurring to asymptotic approximations. Therefore in the rest of this section we will derive the so-called asymptotic mean squared errors (AMSE).

By virtue of results (4) and (5), we get the following

Result 1. *Given two random samples of sizes n_1 and n_2 respectively drawn from the unknown spherical densities f_1 and f_2, consider estimator $\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2)$, $\boldsymbol{x}\in\mathbb{S}^{d-1}$. Assuming that, for $j\in(1,2)$,*

i) $\lim_{n_j\to\infty} b_2(\kappa_j)=0$, *and* $b_r(\kappa_j)=o(b_2(\kappa_j))$, *for* $r\geq 2$,

ii) $\lim_{n_j\to\infty} \nu_0(\kappa_j)/n_j=0$,

iii) all entries of $\nabla^2_{\bar{f}_j}$ *are continuous at* $\boldsymbol{x}$,

it results

$$\begin{aligned}\mathsf{E}[\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2)] &= f_1(\boldsymbol{x})-f_2(\boldsymbol{x})+\frac{1}{2(d-1)}\left\{b_2(\kappa_1)\mathsf{Tr}\left\{\nabla^2_{\bar{f}_1}(\boldsymbol{x})\right\}-b_2(\kappa_2)\mathsf{Tr}\left\{\nabla^2_{\bar{f}_2}(\boldsymbol{x})\right\}\right\}\\ &+ o\left(b_2(\kappa_1)\right)+o\left(b_2(\kappa_2)\right),\end{aligned}$$

$$\mathsf{Var}[\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2)]=\frac{\nu_0(\kappa_1)}{n_1}f_1(\boldsymbol{x})+\frac{\nu_0(\kappa_2)}{n_2}f_2(\boldsymbol{x})+o\left(\frac{\nu_0(\kappa_1)}{n_1}\right)+o\left(\frac{\nu_0(\kappa_2)}{n_2}\right).$$

When K_{κ_1} and K_{κ_2} are both von Mises-Fisher kernels, by virtue of approximations (6), assumptions $i)-ii)$ of Result 1 can be respectively replaced by the assumptions that, as $n_j\to\infty$, $\kappa_j\to\infty$ and $\kappa_j/n_j\to 0$. Moreover, in this special case, we have

$$\begin{aligned}\mathsf{AMSE}[\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2)] &= \frac{1}{4}\left(\frac{\mathsf{Tr}\left\{\nabla^2_{\bar{f}_1}(\boldsymbol{x})\right\}}{\kappa_1}-\frac{\mathsf{Tr}\left\{\nabla^2_{\bar{f}_2}(\boldsymbol{x})\right\}}{\kappa_2}\right)^2\\ &+ \frac{1}{2^{d-1}\pi^{(d-1)/2}}\left(\frac{\kappa_1^{(d-1)/2}}{n_1}f_1(\boldsymbol{x})+\frac{\kappa_2^{(d-1)/2}}{n_2}f_2(\boldsymbol{x})\right),\end{aligned}$$

and the optimal values of κ_1 and κ_2 could be found as the minimizers of $\mathsf{AMSE}[\hat{h}(\boldsymbol{x};\kappa_1,\kappa_2)]$. Closed form expressions, for κ_1 and κ_2 are available only for the circle case:

$$\hat{\kappa}_1=\left\{\left[2\pi^{1/2}n_1(f_1''(\boldsymbol{x}))^2-(2\sqrt{\pi}n_1f_1''(\boldsymbol{x}))^{5/3}(2\pi^{1/2}n_2)^{-2/3}f_2''(\boldsymbol{x})^{1/3}\right]/f(\boldsymbol{x})\right\}^{2/5},$$

$$\hat{\kappa}_2=\left\{\left[2\pi^{1/2}n_2(f_2''(\boldsymbol{x}))^2-(2\pi^{1/2}n_2f_2''(\boldsymbol{x}))^{5/3}(2\pi^{1/2}n_1)^{-2/3}f_1''(\boldsymbol{x})^{1/3}\right]/f(\boldsymbol{x})\right\}^{2/5}.$$

These formulas instructively show that the classification problem cannot be faced by naively selecting the two smoothing degrees, each optimally chosen for its density estimation on the basis of a single sample. In fact, we can note that each smoothing degree depends on *both* f_1 and f_2, and on *both* n_1 and n_2. When n_1 (n_2, resp.) is fixed and n_2 (n_1, resp.) goes to infinity, the optimal value of κ_1 (κ_2, resp.) increases to the optimal value of κ obtained for the case of the circular, standard density estimation.

5. A more general decision rule

In this section we will extend the approach discussed in Section 4 by considering some a priori information. Let $(\boldsymbol{X}, Y)$ be random variables valued in $\mathbb{S}^{d-1} \times \{0,1\}$, and set $\lambda(\boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. Also, denote the density functions in the spherical covariate space for the successes ($Y = 1$) and for the failures ($Y = 0$) by f_1 and f_2, respectively, and let p_1 be the proportion of successes in the population, and $p_2 = 1 - p_1$. Then, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$, we have $\lambda(\boldsymbol{x}) = \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})}$.

We discuss a nonparametric approach to estimate $\lambda(\boldsymbol{x})$ based on the kernel estimation of the densities appearing in its formulation. The Bayes-type classification rule assigns label 1 to a new unlabelled point $\boldsymbol{x}$ if $p_1 f_1(\boldsymbol{x}) \geq p_2 f_2(\boldsymbol{x})$. The corresponding practical rule is to assign label 1 to $\boldsymbol{x}$ if the estimate of $\lambda(\boldsymbol{x})$ is greater than or equal to 0.5. In particular, given a $\mathbb{S}^{d-1} \times \{0,1\}$-valued random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, assume that it has been arranged in such a way that the first n_1 pairs are successes and the last $n_2 = n - n_1$ ones are failures. Then, replacing p_j appearing in $\lambda(\boldsymbol{x})$ with n_j/n, $j \in (1,2)$, a kernel estimator of $\lambda(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{S}^{d-1}$, can be defined as

$$\hat{\lambda}(\boldsymbol{x}; \kappa_1, \kappa_2) = \frac{n_1 \hat{f}_1(\boldsymbol{x}; \kappa_1)}{n_1 \hat{f}_1(\boldsymbol{x}; \kappa_1) + n_2 \hat{f}_2(\boldsymbol{x}; \kappa_2)}, \tag{7}$$

where the $\hat{f}_j$s, $j \in (1,2)$, are kernel estimators of the f_js. Estimators like the above, which satisfy $0 \leq \hat{\lambda}(\boldsymbol{x}; \kappa_1, \kappa_2) \leq 1$, have been studied in the Euclidean setting by [17].

When we use the same kernel function for both $\hat{f}_1$ and $\hat{f}_2$ in (7), and, in addition, a single smoothing parameter, i.e. $\kappa_1 = \kappa_2 = \kappa$, it reduces to the *local constant* estimator for spherical-linear regression studied by [2]. In the general case where $\kappa_1 \neq \kappa_2$, reasoning as in the Euclidean setting (see [17]), we consider in estimator (7) the *true* p_j, $j \in (1,2)$, instead of n_j. Therefore we focus on the estimation of the two densities, by ignoring the error of magnitude $O(n^{-1/2})$ implied by the replacement of p_j by n_j/n.

Concerning the asymptotic L_2 properties of estimator (7), we start by defining the quantities $m(\boldsymbol{x}) = p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})$ and $\hat{m}(\boldsymbol{x}; \kappa_1, \kappa_2) = p_1 \hat{f}_1(\boldsymbol{x}; \kappa_1) + p_2 \hat{f}_2(\boldsymbol{x}; \kappa_2)$. Then, we write $\hat{\lambda}(\boldsymbol{x}; \kappa_1, \kappa_2) = \frac{p_1 \hat{f}_1(\boldsymbol{x}; \kappa_1)}{\hat{m}(\boldsymbol{x}; \kappa_1, \kappa_2)}$, which can be re-written as

$$\hat{\lambda}(\boldsymbol{x}; \kappa_1, \kappa_2) = \frac{p_1 f_1(\boldsymbol{x}) \left(\frac{\hat{f}_1(\boldsymbol{x}; \kappa_1) - f_1(\boldsymbol{x})}{f_1(\boldsymbol{x})} + 1 \right)}{m(\boldsymbol{x}) \left(\frac{\hat{m}(\boldsymbol{x}; \kappa_1, \kappa_2) - m(\boldsymbol{x})}{m(\boldsymbol{x})} + 1 \right)}.$$

Now, since $\hat{f}_1(\boldsymbol{x};\kappa_1) - f_1(\boldsymbol{x})$ and $\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2) - m(\boldsymbol{x})$ are asymptotically *small*, expanding $\hat{f}_1(\boldsymbol{x},\kappa_1)$ and $\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2)$ respectively around $f_1(\boldsymbol{x})$ and $m(\boldsymbol{x})$, it results

$$\begin{aligned}\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2) &= \lambda(\boldsymbol{x})\left(1+\frac{\hat{f}_1(\boldsymbol{x};\kappa_1)-f_1(\boldsymbol{x})}{m(\boldsymbol{x})}-\frac{\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2)-m(\boldsymbol{x})}{m(\boldsymbol{x})}\right)\\ &\quad + O_p(\{\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2)-m(\boldsymbol{x})\}^2)+O_p(\{\hat{f}_1(\boldsymbol{x};\kappa_1)-f_1(\boldsymbol{x})\}^2),\end{aligned}$$

where O_p indicates the order in probability. Now, recalling the definition of $\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2)$, we have

$$\begin{aligned}\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2) &= \lambda(\boldsymbol{x})+\frac{p_1(1-\lambda(\boldsymbol{x}))(\hat{f}_1(\boldsymbol{x};\kappa_1)-f_1(\boldsymbol{x}))-p_2\lambda(\boldsymbol{x})(\hat{f}_2(\boldsymbol{x};\kappa_2)-f_2(\boldsymbol{x}))}{m(\boldsymbol{x})}\\ &\quad + O_p(\{\hat{m}(\boldsymbol{x};\kappa_1,\kappa_2)-m(\boldsymbol{x})\}^2)+O_p(\{\hat{f}_1(\boldsymbol{x};\kappa_1)-f_1(\boldsymbol{x})\}^2).\end{aligned}$$

Starting from above equation, by using quantities (4) and (5), we obtain the following

Result 2. *Given two random samples of sizes n_1 and n_2 respectively drawn from the unknown spherical densities f_1 and f_2, consider estimator $\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2)$, $\boldsymbol{x}\in\mathbb{S}^{d-1}$. If assumptions $i)-iii)$ of Result 1 hold, and n_1 and n_2 go to infinity in such a way that $n_1/n_2\to p_1/p_2$, $b_2(\kappa_1)\approx b_2(\kappa_2)$, and $\nu_0(\kappa_1)\approx\nu_0(\kappa_2)$, it results*

$$\begin{aligned}\mathsf{E}[\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2)]-\lambda(\boldsymbol{x}) &= \frac{p_1p_2\left(b_2(\kappa_1)f_2(\boldsymbol{x})\mathsf{Tr}\{\nabla^2_{\tilde{f}_1}(\boldsymbol{x})\}-b_2(\kappa_2)f_1(\boldsymbol{x})\mathsf{Tr}\{\nabla^2_{\tilde{f}_2}(\boldsymbol{x})\}\right)}{2(d-1)m^2(\boldsymbol{x})}\\ &\quad + o(b_2(\kappa_1)),\end{aligned}$$

$$\mathsf{Var}[\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2)]=\frac{\lambda(\boldsymbol{x})(1-\lambda(\boldsymbol{x}))}{n\,m(\boldsymbol{x})}\left[(1-\lambda(\boldsymbol{x}))\nu_0(\kappa_1)+\lambda(\boldsymbol{x})\nu_0(\kappa_2)\right]+o\left(\frac{\nu_0(\kappa_1)}{n}\right).$$

If K_{κ_1} and K_{κ_2} are both von Mises-Fisher kernels, with $\kappa_1\approx\kappa_2$, by virtue of Result 2 and by using approximations (6), we easily obtain the asymptotic bias and variance

$$\mathsf{E}[\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2)]-\lambda(\boldsymbol{x})=\frac{p_1p_2}{2m^2(\boldsymbol{x})}\left(\frac{f_2(\boldsymbol{x})\mathsf{Tr}\{\nabla^2_{\tilde{f}_1}(\boldsymbol{x})\}}{\kappa_1}-\frac{f_1(\boldsymbol{x})\mathsf{Tr}\{\nabla^2_{\tilde{f}_2}(\boldsymbol{x})\}}{\kappa_2}\right)+o\left(\frac{1}{\kappa_1}\right),$$

$$\begin{aligned}\mathsf{Var}[\hat{\lambda}(\boldsymbol{x};\kappa_1,\kappa_2)] &= \frac{\lambda(\boldsymbol{x})(1-\lambda(\boldsymbol{x}))}{2^{d-1}\pi^{(d-1)/2}nm(\boldsymbol{x})}\left[(1-\lambda(\boldsymbol{x}))\kappa_1^{(d-1)/2}+\lambda(\boldsymbol{x})\kappa_2^{(d-1)/2}\right]\\ &\quad + o\left(n^{-1}\kappa_1^{(d-1)/2}\right).\end{aligned}$$

As for the optimal smoothing, in the Euclidean setting the standard approach is to consider the values which minimize a weighted version of the mean squared error. For practical implementation the smoothing parameters can be selected by minimizing an empirical version of the weighted mean squared error (for details see [17]).

6. Global temperatures

We consider values taken from Berkeley Earth[1] who have extracted data from a variety of sources, combined these with some bias-correction, and produced "anomaly" values. These show, for 15984 equal-area grid points regularly spaced on the earth's surface, and for each month in the period 1850–2015, the difference between the monthly average at that location and the 1951–1980 mean. Note that an anomaly could be small, and the fact that there is always a recorded anomaly means that it does not indicate anything abnormal. Although these anomalies (given in $^{\circ}C$) are real-valued, we consider, for our classification example, only the *sign* of the anomaly. Examples of this classification are shown for two successive months in Figure 2. In general, the number of missing values decreased over time, and these were more prevalent closer to to the South pole. In the figure, we can note some similarities between the two months, with an agreement of just over 68% (of the non-missing values), which suggests that a nearest-neighbour classifier would achieve an error rate of about 32% on this naive version of the data. In our classification task we will consider data with "overlapping"

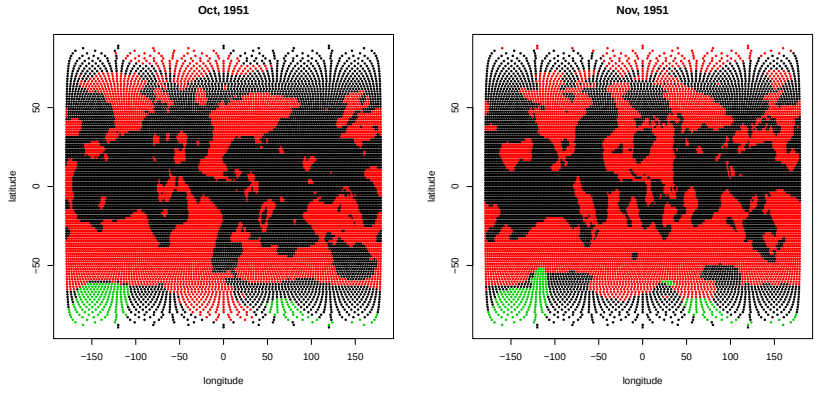

Figure 2: Sign of temperature anomalies at each grid point location in two consecutive months. Red: negative; Black: positive; Green: missing data.

class distributions. Given some training data, which will be a random sample (of size

[1] http://berkeleyearth.org/data/

2000) of monthly anomalies and locations from a 3-month period, we will predict the classes of monthly data occurring in the next quarter (three month period) which is also of a random sample of 2000 locations. We use Equation (7) with a common smoothing parameter ($\kappa = \kappa_1 = \kappa_2$) for both classes, but do not use the information about the month (within the quarter) of either the training or test data. We choose the smoothing parameter in two ways: (i) using leave-one-out cross-validation for the data from the current month, say κ_{cvc}; (ii) choosing κ to minimize the error rate, when data from the previous 3 months are used to classify data in the current quarter. Given that the error rate is a non-continuous function, we use, instead,

$$\kappa_{\text{cvf}} = \text{argmax}_\kappa \sum_{i:Y_i=1} \log \hat{\lambda}\left(\boldsymbol{x}_i; \kappa\right) + \sum_{i:Y_i=2} \log\left(1 - \hat{\lambda}\left(\boldsymbol{x}_i; \kappa\right)\right).$$

In general, we anticipate $\kappa_{cvf} < \kappa_{cvc}$, since there will be less class overlap in this case.

We predict one-quarter ahead monthly temperature anomalies from January-March 1950 to April-June, 1980, with independent samples drawn (for previous, current and future quarters) for each of the predicted months from the next quarter. The error rates (for 120 quarters) are shown in Figure 3, for the two ways of selecting κ, and for the "default" rule, in which all observations are allocated to the most common class (in the current month). We also show the selected smoothing parameters ($\kappa_{cvc}, \kappa_{cvf}$) for each quarter. The average error rates are 0.449, 0.381, 0.407 respectively, for the default classifier, and Equation (7) using κ_{cvc} and κ_{cvf} respectively. The smoothing parameter selections had averages of $\overline{\kappa}_{cvc} = 167.1$ and $\overline{\kappa}_{cvf} = 16.0$ which was consistent with our expectations. However, it is somewhat surprising that such different values of smoothing parameter have given quite similar error rates (on average).

The error rates for this problem are quite high, although still better than the default classifier. The reasons for this are that the month-by-month changes are non-negligible, the population boundaries are rather inhomogeneous, and that the classification task has been derived from thresholding a real-valued response, for which the inter-quartile range is less than one degree centigrade. Given the agreement between consecutive months noted in the above example (Figure 2) we might consider a nearest neighbour classifier as more suited to this problem; the average error rate for these data is 0.399 which is slightly more than that when using κ_{cvc} as above.

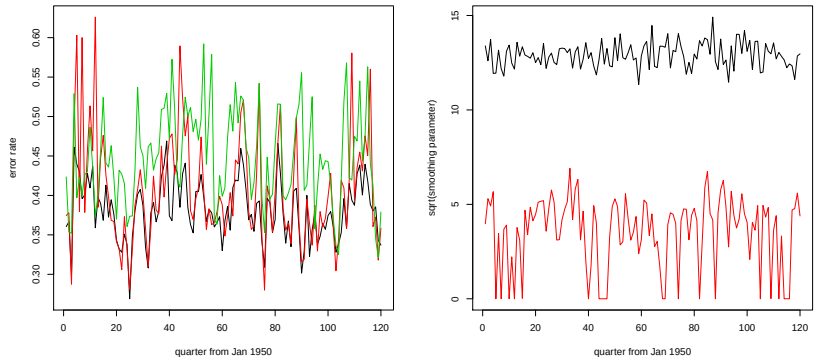

Figure 3: Left: Error rates for 120 months using default classification rule (green), κ_{cvc} estimated from current month (black) and κ_{cvf} estimated from previous month to current month forecast (red). Right: square root of smoothing parameter values selected according to: leave-one-out cross-validation (κ_{cvc}) (black) and κ_{cvf} (red).

[1] Bai, Z.D., Rao, C.R., Zhao, L.C., 1988. Kernel estimators of density function of directional data. Journal of Multivariate Analysis 27, 24-39.

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

[3] Di Marzio, M., Taylor, C.C., 2005. Kernel density classification and boosting: an L_2 analysis. Statistics and Computing 15, 113–123.

[4] El Khattabi, S., Streit, F., 1996. Identification analysis in directional statistics. Computational Statistics & Data Analysis 23, 45–53.

[5] García-Portugués, E., 2013. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. Electronic Journal of Statistics 7, 1655–1685.

[6] García-Portugués, E., Crujeiras, R.M., González-Manteiga, W., 2013. Kernel density estimation for directional–linear data. Journal of Multivariate Analysis 121, 152–175.

[7] García-Portugués, E., Crujeiras, R.M., González-Manteiga, W., 2015. Central limit theorems for directional and linear random variables with applications. Statistica Sinica 25, 1207–1229.

[8] 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.

[9] Hall, P., Kang, K.H., 2005. Bandwidth choice for nonparametric classification. The Annals of Statistics 33, 284–306.

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

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

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

[13] Lopez-Cruz, P.L., Bielza, C., Larranaga, P., 2015. Directional naive Bayes classifiers. Pattern Analysis and Applications 18, 225–246.

[14] Loubes, J.M., Pelletier, B., 2008. A kernel-based classifier on a Riemannian manifold. Statistics & Decisions 26, 35–51.

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

[16] Morris, J.E., Laycock, P.J., 1974. Discriminant Analysis for Directional Data. Biometrika 61, 335–341.

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

[18] Watson, G.S., 1956. Analysis of dispersion on a sphere. Royal Astronomical Society Geophysical Supplement 7, 153–159.