ARTICLE TEMPLATE

Comparison of discriminant analysis methods on the sphere

Michail Tsagris[a] and Abdulaziz Alenazi[b]

[a]Department of Economics, University of Crete, Rethymnon, Greece; [b]Department of Mathematics, Northern Border University, Arar, Saudi Arabia

ARTICLE HISTORY
Compiled October 17, 2019

ABSTRACT
Discriminant analysis for spherical data (directional data in general) has not been studied to a great degree and most papers focus on one distribution, the rotationally symmetric (or isotropic) von Mises-Fisher. This is the first paper on maximum likelihood discriminant analysis with spherical data that considers non rotationally symmetric distributions, while the k-Nearest Neighbours algorithm is included as a model-free alternative. Extensive Monte Carlo simulations and experiments with numerous real data yield multiple conclusions regarding the algorithms' predictive performance and computational cost. Maximum likelihood discriminant analysis using rotationally non-symmetric distributions performed satisfactorily and surprisingly enough, rotationally symmetric distributions performed well in some cases. Overall, the k-NN algorithm is suggested because it is non-parametric hence flexible, computationally efficient, scalable to large sample sizes and suitable for big data, and on average is on par or outperforms the other methods.

KEYWORDS
Discriminant analysis, spherical data, non-rotational symmetry

AMS CLASSIFICATION
62H11, 62H30

1. Introduction

Directional data are multivariate data constrained to lie on a unit radius (hyper-)sphere. Such data arise in many different fields, such as biology (Paterson et al. 2017), bioinformatics (Audit and Ouzounis 2003), zoology (Amson et al. 2017), ecology (Vanni, Baldaccini, and Giunchi 2017), geophysics (Rutkowska, Kohnová, and Banasik 2018), political sciences (Gill and Hangartner 2010) and transportation (Laha and Putatunda 2018) to name a few. In mathematical terms, their sample space, denoted by $\mathbb{S}^{p-1}$, is given by

$$\mathbb{S}^{p-1} = \left\{ \mathbf{y} \in \mathbb{R}^p, \;\; \mathbf{y}^T\mathbf{y} = 1 \right\}.$$

In the special case of $p = 2$ they are termed circular or angular data and they lie on the unit circle. If $p = 3$, they lie on the unit sphere, and hence are termed spherical data; see Figure 1 for an example.

CONTACT Michail Tsagris. Email: mtsagris@uoc.gr

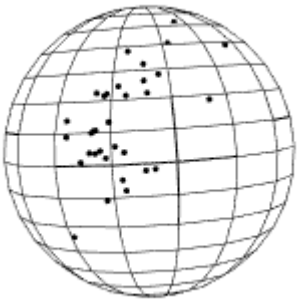


Figure 1. An example of spherical data taken from (Wood 1982). Measurements regarding previous magnetic pole position obtained using palaeomagnetic techniques.

Mathematically speaking, the sphere is an example of a manifold in 3 dimensions[1]. On the earth's surface for example, we can claim that a city (or even a county) constitutes a plane in $\mathbb{R}^2$ and data collected on that region can be considered Euclidean (locally). However, this is a Euclidean approximation (locally) to the sphere and for this reason appropriate spherical models have been developed since the middle of the 20th century. For example, Fisher (1953) proposed the von Mises-Fisher distribution and Mardia (1975) proposed the Fisher-Bingham distribution.

Clustering (unsupervised learning) with directional data has been addressed by either using hierarchical clustering (Lund 1999), the k-means algorithm (Hornik et al. 2012) or by model based clustering with the von Mises-Fisher (Banerjee et al. 2005) or the Kent distribution Peel, Whiten, and McLachlan (2001). More recently, Amayri and Bouguila (2013) included the task of feature selection into model based clustering, using mixtures of von Mises-Fisher distributions. In the context of discriminant analysis (supervised learning, or classification), both Morris and Laycock (1974) and Figueiredo (2009) conducted simulation studies considering the von Mises-Fisher distribution alone. Hamsici and Martinez (2007) explored the approximation of the multivariate Gaussian to the von Mises-Fisher distribution, while Lopez-Cruz et al. (2015) López-Cruz, Bielza, and Larrañaga (2015) proposed a naive Bayes classifier that is again based on the von Mises-Fisher distribution. Kent, Ganeiber, and Mardia (2013) on the other hand, briefly examined the use of the Kent distribution, but with no simulation studies or real data analysis whatsoever.

The drawback of the aforementioned papers is that they examine the problem of classification with distributions of limited capabilities. Broadly speaking, most papers applied or not, employ the von Mises-Fisher distribution, perhaps due to its convenient form and ease to work with, neglecting its rather unrealistic rotational symmetry assumption, the isotropic covariance matrix. Kent (1982) addressed this limitation by proposing the first rotationally non-symmetric distribution, and only recently Paine et al. (2018) proposed the second rotationally non-symmetric distribution, the Elliptically Symmetric Angular Gaussian (ESAG) distribution. Both the Kent and ESAG distributions have elliptical shape, allowing for correlation between the variables[2].

[1]In general a manifold is a topological space that locally resembles Euclidean space near each point. Each point of a p-dimensional manifold has a neighbourhood (tangent plane for example) that is homeomorphic. Homeomorphism is a mapping that preserve all the topological properties of a given space to the Euclidean space $\mathbb{R}^p$.

[2]The correspondence with the Euclidean discriminant analysis is to match the Von Mises-Fisher to Gaussian linear discriminant analysis with isotropic covariance matrix and the Kent to the quadratic discriminant

Another common drawback of these papers is that they do not compare with more alternative distributions and algorithms. The literature contains more spherical distributions and more algorithms than maximum likelihood discriminant analysis, but no one to the best of our knowledge, has studied discriminant analysis for directional (or even spherical) data by using either more distributions or more algorithms. There is no comparative evaluation of the Kent to the von Mises-Fisher distribution in the context of discriminant analysis. It is true that no "free lunch" exists and no algorithm unanimously outperforms all other algorithms, yet a concrete and solid large scale comparison of the existing methods/algorithms does not exist.

These reasons motivated us to extend the work of Figueiredo (2009), with a focus on spherical data only, by including three more distributions, the Independent Angular Gaussian (IAG), or projected normal (Mardia and Jupp 2000), the Kent distribution (Kent 1982), and the ESAG distribution (Paine et al. 2018). We also consider the k-Nearest Neighbours k-NN algorithm Cover and Hart (1967), coupled with the cosine distance, as the non-parametric competitor. The aim of this paper is to provide evidence for the suitability of each distribution and whether practitioners and researchers working with spherical data should consider maximum likelihood discriminant analysis or the k-NN algorithm. To this end we have implemented extensive Monte Carlo simulations with various scenarios assessing the predictive performance and the computational cost of each method. In addition, we have performed empirical evaluation studies using various real data from geology, medicine and sociology in order to draw safer, and more realistic, conclusions regarding the algorithms' predictive performance.

In the next section we mention some preliminaries regarding discriminant analysis; a) the spherical distributions we will examine, along with the maximum likelihood estimation of their parameters and b) the standard k-NN algorithm and a variant of it. We compare these methods via Monte Carlo simulations in Section 3 and by using real data analysis in Section 4. Finally, we conclude the paper in Section 5.

2. Discriminant analysis with spherical data

Discriminant analysis constructs discrimination rules or boundaries between groups of observations and unlike clustering, the label of each observation, or the group to which each observation belongs, is known. Examples of discriminant analysis with spherical data include the case of separating the longest axis and shortest axis orientations of tabular stones Fisher, Lewis, and Embleton (1993) and the classification of the constitutes measurements of magnetic remanence in rock specimens, after each specimen had been partially thermally demagnetised to the same stage (Fisher, Lewis, and Embleton 1993).

2.1. *Maximum likelihood discriminant analysis*

The first algorithm we will use is maximum likelihood discriminant analysis. For each group of observations, the same family of distributions is assumed and we estimate the parameters for each group using maximum likelihood estimation. In order to allocate a new observation into a group, the density of the new observation is computed for each group and the observation is allocated to the group with the highest density value. Below, we discuss the maximum likelihood discriminant analysis for spherical data.

---

analysis.

2.1.1. The von Mises-Fisher distribution

The density of the von Mises-Fisher distribution on $\mathbb{S}^2$ is given by (Mardia and Jupp 2000)

$$f\left(\mathbf{y};\boldsymbol{\gamma},\kappa\right)=\frac{\kappa}{2\pi\left(e^{\kappa}-e^{-\kappa}\right)}e^{\kappa\boldsymbol{\gamma}^T\mathbf{y}}, \tag{1}$$

where $\kappa \geq 0$ (concentration parameter, scalar), $\boldsymbol{\gamma} \in \mathbb{S}^2$ is the mean direction and $\mathbf{y} \in \mathbb{S}^2$. The corresponding log-likelihood is given by

$$\ell = n\log\frac{\kappa}{2\pi}-n\log\left(e^{\kappa}-e^{-\kappa}\right)+\kappa\sum_{i=1}^{n}\boldsymbol{\gamma}^T\mathbf{y}_i.$$

The estimated mean direction is available in closed form $\hat{\boldsymbol{\gamma}}=\frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|}$, where $\bar{\mathbf{y}}=n^{-1}\left(\sum_{i=1}^{n}\mathbf{y}_{1i},\sum_{i=1}^{n}\mathbf{y}_{2i},\sum_{i=1}^{n}\mathbf{y}_{3i}\right)^T$ and $\|\cdot\|$ denotes the Euclidean norm. The concentration parameter is independent of the mean direction and its estimation is achieved using a truncated Newton-Raphson algorithm[3] (Sra 2012)

$$\hat{\kappa}^{(t)}=\hat{\kappa}^{(t-1)}-\frac{A_3\left(\hat{\kappa}^{(t-1)}\right)-\bar{R}}{1-\left[A_3\left(\hat{\kappa}^{(t-1)}\right)\right]^2-\frac{2}{\hat{\kappa}^{(t-1)}}A_3\left(\hat{\kappa}^{(t-1)}\right)} \tag{2}$$

and similarly to Sra (2012) we will set the starting value in (2) equal to $\hat{\kappa}^{(0)}=\frac{\bar{R}\left(p-\bar{R}^2\right)}{1-\bar{R}^2}$. $A_3\left(\hat{\kappa}\right)=I_{3/2}\left(\hat{\kappa}\right)/I_{3/2-1}\left(\hat{\kappa}\right)$, $I_{\nu}\left(\hat{\kappa}\right)$ is the modified Bessel function of the first kind[4] Abramowitz and Stegun (1970) of order ν evaluated at $\hat{\kappa}$ and $\bar{R}=\frac{\|\sum_{i=1}^{n}\mathbf{y}_i\|}{n}$ is the mean resultant length.

2.1.2. The Isotropic Angular Gaussian distribution

The density of the Angular Gaussian (AG) distribution is (Mardia and Jupp 2000)

$$\begin{aligned} f\left(\mathbf{y};\boldsymbol{\mu},\mathbf{V}\right) &= \frac{1}{2\pi|\mathbf{V}|^{1/2}\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}\right)^{3/2}}\times\exp\left\{\frac{1}{2}\left[\frac{\left(\mathbf{y}^T\mathbf{V}^{-1}\boldsymbol{\mu}\right)^2}{\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}\right)}-\left(\boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu}\right)\right]\right\} \\ &\quad\times M_2\left[\frac{\left(\mathbf{y}^T\mathbf{V}^{-1}\boldsymbol{\mu}\right)^2}{\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}\right)}\right], \end{aligned} \tag{3}$$

where $M_2(\alpha)=(1+\alpha^2)\Phi(\alpha)+\alpha\phi(\alpha)$ and $\Phi(.)$, $\phi(.)$ denote the cumulative and probability density function, respectively, of the standard normal distribution. $\boldsymbol{\mu}\in\mathbb{R}^3$ is the mean vector and $\mathbf{V}$ is a positive definite matrix. When $\mathbf{V}=\mathbf{I_3}$, we end up with the IAG distribution

$$f(\mathbf{y};\boldsymbol{\mu})=\frac{1}{2\pi}\exp\left[\frac{1}{2}\left\{\left(\mathbf{y}^T\boldsymbol{\mu}\right)^2-\boldsymbol{\mu}^T\boldsymbol{\mu}\right\}\right]M_2\left(\mathbf{y}^T\boldsymbol{\mu}\right), \tag{4}$$

[3]The iterative solution in (2) is the general solution. For the spherical case a simpler form exists.

[4]The modified Bessel function in R gives us the option to scale it exponentially. This is useful because when large numbers are plugged into the Bessel function, R needs the exponential scaling to calculate the ratio of the two Bessel functions and avoid numerical overflow.

whose log-likelihood is given by

$$\ell = -n\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)^2 - \frac{n}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\mu}\right) + \sum_{i=1}^{n}\log\left\{M_2\left[\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)^2\right]\right\}. \tag{5}$$

We estimate the mean vector using the Newton-Raphson algorithm $\boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t - \mathbf{H}^{-1}\mathbf{J}$, where $\mathbf{J}$ denotes the first derivative and $\mathbf{H}$ (Hessian matrix) is the second derivative of (5), both with respect to $\boldsymbol{\mu}$

$$\begin{aligned}
\mathbf{J} &= \frac{\partial\ell}{\partial\boldsymbol{\mu}} = \sum_{i=1}^{n}\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\mathbf{y}_i^T - n\boldsymbol{\mu}^T + \sum_{i=1}^{n}\frac{g_i'(\boldsymbol{\mu})}{g_i(\boldsymbol{\mu})} \quad \text{and} \\
\mathbf{H} &= \frac{\partial^2\ell}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T} = \sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^T - n\mathbf{I}_3 + \sum_{i=1}^{n}\frac{g_i''(\boldsymbol{\mu})\,g_i(\boldsymbol{\mu}) - [g_i'(\boldsymbol{\mu})]^2}{g_i^2(\boldsymbol{\mu})},
\end{aligned}$$

where

$$\begin{aligned}
g_i(\boldsymbol{\mu}) &= \Phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\left[1 + \left(\mathbf{y}_i^T\boldsymbol{\mu}\right)^2\right] + \mathbf{y}_i^T\boldsymbol{\mu}\phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right), \\
g_i'(\boldsymbol{\mu}) &= 2\left[\mathbf{y}_i^T\boldsymbol{\mu}\Phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\mathbf{y}_i^T + \phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\right]\mathbf{y}_i^T \quad \text{and} \\
g_i''(\boldsymbol{\mu}) &= \sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^T - n\mathbf{I}_3 + 2\sum_{i=1}^{n}\frac{\Phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)}{g_i(\boldsymbol{\mu})}\mathbf{y}_i\mathbf{y}_i^T - \sum_{i=1}^{n}\frac{g_i'(\boldsymbol{\mu})\,g_i'(\boldsymbol{\mu})^T}{g_i^2(\boldsymbol{\mu})}.
\end{aligned}$$

2.1.3. The Kent distribution

Kent (1982) defined the distribution whose density is given by

$$f(\mathbf{y};\boldsymbol{\gamma},\kappa,\beta) = \frac{1}{C(\kappa,\beta)}\exp\left\{\kappa\boldsymbol{\gamma}\mathbf{y} + \beta\left[(\boldsymbol{\alpha}_1\mathbf{y})^2 - (\boldsymbol{\alpha}_2\mathbf{y})^2\right]\right\}, \tag{6}$$

where β is the ovalness parameter, k is the concentration parameter[5], $\boldsymbol{\gamma}$ is the mean direction and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ are the major and minor axis respectively with $\boldsymbol{\gamma},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2 \in \mathbb{S}^2$. The normalizing constant $C(\kappa,\beta)$ has a closed form (as a sum of infinite terms) in the spherical case Kent (1982) and the corresponding log-likelihood is

$$\ell = -nC(\boldsymbol{\gamma},\kappa,\beta) + \kappa\sum_{i=1}^{n}\boldsymbol{\gamma}^T\mathbf{y}_i + \beta\left[\sum_{i=1}^{n}\left(\boldsymbol{\alpha}_1^T\mathbf{y}_i\right)^2 - \sum_{i=1}^{n}\left(\boldsymbol{\alpha}_2^T\mathbf{y}_i\right)^2\right]. \tag{7}$$

When estimating the parameters of the Kent distribution we first estimate the matrix $\mathbf{A} = (\boldsymbol{\gamma},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)$ via moments. By choosing an orthogonal matrix $\mathbf{H}$ to rotate the mean vector $\bar{\mathbf{y}}$ to the north polar axis $(1,0,0)^T$, $\mathbf{H}$ can be written as

$$\mathbf{H} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & -\cos\phi \end{bmatrix},$$

[5]Uni-modality of the distribution requires that $|\beta| < \kappa/2$.

where θ and ϕ are the polar co-ordinates of $\bar{\mathbf{y}}$. Let $\mathbf{B} = \mathbf{H}^T\mathbf{S}\mathbf{H}$, where $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^T$. We then choose a rotation $\mathbf{K}$ about the north pole to diagonalize $\mathbf{B}_L$, where

$$\mathbf{B}_L = \begin{bmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{bmatrix}$$

is the lower 2×2 sub-matrix of $\mathbf{B}$, with eigenvalues $\lambda_1 > \lambda_2$. If we choose ψ (the angle of rotation between the mean direction and the major axis $\boldsymbol{\alpha}_2$) such that $\tan(2\psi) = 2b_{23}/(b_{22} - b_{33})$, ensuring that $\| \bar{\mathbf{y}} \| > 0$ and $\lambda_1 > \lambda_2$ then we can take

$$\mathbf{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}.$$

The moment estimate of $\mathbf{A}$ is given by $\tilde{\mathbf{A}} = \mathbf{H}\mathbf{K}$. As for the parameters κ and β we maximize (6) with respect to these two parameters using a numerical optimizer, such as the Nelder-Mead algorithm (Nelder and Mead 1965), available in R via the command *optim*.

2.1.4. The Elliptically Symmetric Angular Gaussian distribution

The ESAG distribution Paine et al. (2018) is another non rotationally symmetric distribution whose density is given by

$$f(\mathbf{y};\boldsymbol{\mu},\mathbf{V}) = \frac{1}{2\pi\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}\right)^{3/2}} \times \exp\left\{\frac{1}{2}\left[\frac{\left(\mathbf{y}^T\boldsymbol{\mu}\right)^2}{\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}} - \boldsymbol{\mu}^T\boldsymbol{\mu}\right]\right\} \times M_2\left[\frac{\left(\mathbf{y}^T\boldsymbol{\mu}\right)^2}{\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}}\right]. \quad (8)$$

The log-likelihood of (8) is given by

$$\ell = -n\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log\left(\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i\right) + \frac{1}{2}\sum_{i=1}^{n}\frac{\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)^2}{\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i} - \frac{n}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} + \sum_{i=1}^{n}M_2\left[\frac{\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)^2}{\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i}\right]. \quad (9)$$

ESAG (8) was derived from AG (3) as a result of two conditions, a) $\mathbf{V}\boldsymbol{\mu} = \boldsymbol{\mu}$ and b) $|\mathbf{V}| = 1$. The largest eigenvalue of the positive definite matrix $\mathbf{V}$ is 1 due to the first condition. The other two eigenvalues are $0 < \rho_1 \leq \rho_2$, and hence $\mathbf{V}^{-1}$ can be written as

$$\mathbf{V}^{-1} = \xi_d\xi_d^T + \sum_{j=1}^{2}\xi_j\xi_j^T/\rho_j, \quad (10)$$

where ξ_1, ξ_2 and $\xi_3 = \mu/\|\mu\|$ is a set of mutually orthogonal unit vectors. The second condition implies $\prod_{j=1}^{2}\rho_j = 1$. Once the 3 parameters in $\boldsymbol{\mu}$ are fixed, then from the two conditions there is 1 remaining degree of freedom for the eigenvalues of V, and 1 degree of freedom for its unit eigenvectors, thus similarly to the Kent distribution (6), the total number of free parameters is 5. Similarly to the Kent distribution, maximisation of (9) can be implemented in R using the command *optim*.

2.2. *Characteristics of the above densities*

The IAG distribution is very similar to the von Mises-Fisher distribution (Watson 1983) and they share common properties. For example the concentration parameter of the von Mises-Fisher distribution is roughly similar to the norm of the mean vector of the IAG, $\kappa \approx \| \boldsymbol{\mu} \|$. Their main difference lies in their construction. The von Mises-Fisher is a multivariate normal distribution with a covariance matrix equal to the identity matrix conditioned to lie on the unit (hyper-)sphere, $\mathbf{y} \sim \text{vMF}(\boldsymbol{\gamma}, \kappa) \equiv N_3\left(\boldsymbol{\mu}, \mathbf{I}_3 | \mathbf{y}^T\mathbf{y} = 1\right)$. IAG, on the contrary, is a multivariate normal distribution projected on the (hyper-)sphere $\mathbf{y} \sim \text{IAG}(\boldsymbol{\mu})$, where $\mathbf{y} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ and $\mathbf{x} \sim N_3(\boldsymbol{\mu}, \mathbf{I}_3)$. The von Mises-Fisher and IAG distributions are both rotationally symmetric about their mean direction $\boldsymbol{\mu}$. This is the analogue of a bivariate normal in $\mathbb{R}^2$ with isotropic or spherical covariance matrix. Their contours plot consists of many concentric circles as presented in Figure 2(a). The von Mises-Fisher (1) stems from the Kent distribution (6) with $\beta = 0$, $\text{vMF}(\boldsymbol{\gamma}, \kappa) \equiv \text{Kent}(\boldsymbol{\gamma}, \kappa, 0)$. The IAG (4) distribution corresponds to $\mathbf{V} = \mathbf{I}_3 \Leftrightarrow (\gamma_1, \gamma_2)^T = (0, 0)^T$, $\text{IAG}(\boldsymbol{\mu}) \equiv \text{ESAG}(\boldsymbol{\mu}, \mathbf{I}_3)$.

The Kent and ESAG distributions on the other hand have elliptical shape (see Figures 2(b) and 2(c)), overcoming the restraining rotational symmetry assumption. They can be seen as the analogue of a bivariate normal distribution, with some restrictions on the covariance matrix. The Kent distribution is a special case of the Fisher-Bingham distribution (Mardia 1975). Instead of 8 parameters, it has 5 (hence the name FB_5). The ESAG distribution, similarly to the Kent, comes from the more general AG distribution (3), which also has 8 parameters, with proper constraints on the covariance matrix (Paine et al. 2018). The proximity between the Kent and the ESAG can be graphically examined in the contours plot in Figure 2(c).

2.3. *The maximum likelihood discriminant boundaries*

The general rule is to allocate a new observation vector $\mathbf{x} \in \mathbb{S}^2$ in the group whose log-likelihood value has the highest value. The rule in our case with two groups is

- For the von Mises-Fisher, allocate $\mathbf{x}$ to group 1 iff

$$\log \frac{\kappa_1}{\kappa_2} - \log \frac{e^{\kappa_1} - e^{-\kappa_1}}{e^{\kappa_2} - e^{-\kappa_2}} + \left(\kappa_1 \boldsymbol{\gamma}_1^T - \kappa_2 \boldsymbol{\gamma}_2^T\right) \mathbf{x} > 0$$

  and to group 2 otherwise.
- For the IAG, allocate $\mathbf{x}$ to group 1 iff

$$\frac{1}{2}\left[\left(\mathbf{x}^T\boldsymbol{\mu}_1\right)^2 - \left(\mathbf{x}^T\boldsymbol{\mu}_2\right)^2 - \left(\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1\right) + \left(\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2\right)\right] + M_2\left[\left(\mathbf{x}^T\boldsymbol{\mu}_1\right)^2\right] - M_2\left[\left(\mathbf{x}^T\boldsymbol{\mu}_2\right)^2\right] > 0$$

  and to group 2 otherwise.
- For the Kent, allocate $\mathbf{x}$ to group 1 iff

$$\begin{aligned} &-C(\boldsymbol{\gamma}_1, \kappa_1, \beta_1) + C(\boldsymbol{\gamma}_2, \kappa_2, \beta_2) + \left(\kappa_1 \boldsymbol{\gamma}_1^T - \kappa_2 \boldsymbol{\gamma}_2^T\right) \mathbf{x} \\ &+\beta_1\left[\left(\boldsymbol{\alpha}_{2,1}^T\mathbf{x}\right)^2 - \left(\boldsymbol{\alpha}_{3,1}^T\mathbf{x}\right)^2\right] - \beta_2\left[\left(\boldsymbol{\alpha}_{2,2}^T\mathbf{x}\right)^2 - \left(\boldsymbol{\alpha}_{3,2}^T\mathbf{x}\right)^2\right] > 0 \end{aligned}$$

  and to group 2 otherwise.

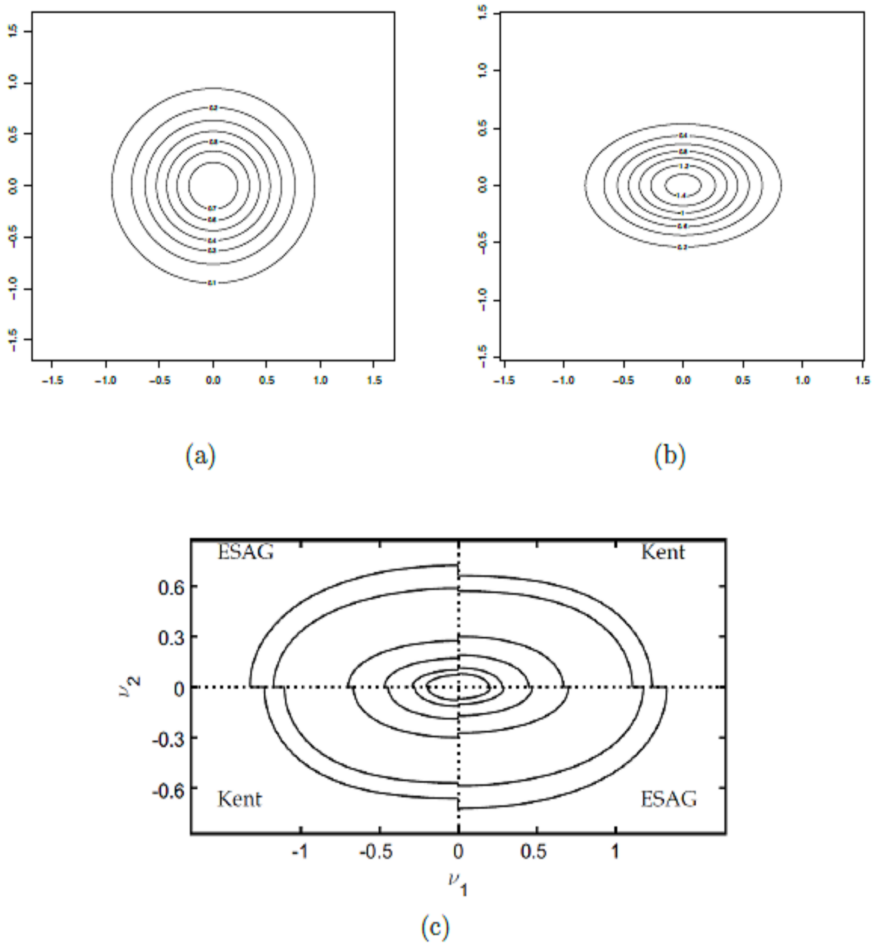


Figure 2. Contour plots of (a) the von Mises-Fisher and (b) the Kent distribution. (c) Comparison of the contour plots of the Kent and ESAG distributions.

- And for the ESAG, allocate $\mathbf{x}$ to group 1 iff

$$-\frac{3}{2}\log\frac{\left(\mathbf{x}^T\mathbf{V}_1^{-1}\mathbf{x}\right)}{\left(\mathbf{x}^T\mathbf{V}_2^{-1}\mathbf{x}\right)}+M_2\left[\frac{\left(\mathbf{x}^T\boldsymbol{\mu}_1\right)^2}{\left(\mathbf{x}^T\mathbf{V}_1^{-1}\mathbf{x}\right)}\right]-M_2\left[\frac{\left(\mathbf{x}^T\boldsymbol{\mu}_2\right)^2}{\left(\mathbf{x}^T\mathbf{V}_2^{-1}\mathbf{x}\right)}\right]$$
$$+\frac{1}{2}\left[\frac{\left(\mathbf{x}^T\boldsymbol{\mu}_1\right)^2}{\left(\mathbf{x}^T\mathbf{V}_1^{-1}\mathbf{x}\right)}-\frac{\left(\mathbf{x}^T\boldsymbol{\mu}_2\right)^2}{\left(\mathbf{x}^T\mathbf{V}_2^{-1}\mathbf{x}\right)}-\left(\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1\right)+\left(\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2\right)\right]>0$$

  and to group 2 otherwise.

We could simplify the above inequalities by assuming equal concentration parameters (see (Figueiredo 2009) for example) and/or ovalness parameters etc. In the case of the von Mises-Fisher for example, that would be the analogue of a linear Naive Bayes classifier in $\mathbb{R}^3$. We argue against this practice because parametric discriminant analysis puts constrains on the shape of the data which might be unrealistic. The von Mises-Fisher for example is a restrictive distribution; there is no reason to further restrict it.

2.4. *The k-NN algorithm*

The k-NN (k Nearest Neighbours) algorithm is an intuitive classifier that makes no parametric assumptions about the data. The standard k-NN works by determining the k observations in the sample that are closest, by some choice of distance measures, to some new observation(s). The new observation will be allocated to the sample that is most common amongst these k "nearest neighbours". In the occasion of two more samples being the most common can be broken by allocating uniformly at random amongst the tied samples (the strategy we use in our simulation studies) or else by using a secondary tie-breaking criterion. In this work we also used a variant that calculates the distances of the k "nearest neighbours" from the test observation from each group and allocates it to the group with the smallest average distance. We term this the non-standard k-NN. It is computationally more expensive, but could yield more accurate allocations.

The performance of the k-NN algorithm depends of the choice of k: small k allows for classification boundaries which are flexible but with a tendency to overfit, while the opposite is true when k is large. It also depends on the choice of distance measure, but since we are dealing with directional data we shall use the cosine distance (or cosine dissimilarity)

$$D\left(\mathbf{y}_i, \mathbf{y}_j\right) = \cos^{-1}\left(\mathbf{y}_i^T \mathbf{y}_j\right). \tag{11}$$

When the angle between the two vectors is 0, their inner product is 1 and the arc of the cosine is 0°. The maximum value of (11) is achieved when the angle between the two vectors is 180°, their inner product is -1 and the hence the arc of the cosine is $\pi/2$. However, in order to reduce the computational cost, we do not compute the arc of the cosine in (11), as the inner product is sufficient and the nearest neighbours are the observations with the highest inner product value[6].

3. Simulation studies

We have conducted extensive simulation studies to draw useful and helpful conclusions, and to get better insights into the behaviour of the two classification algorithms with spherical data. We examined the case of two groups only, compensating for the many situations and cases to be examined. Specifically, we examined many different combinations of sample sizes $n = (50, 100, 150, 200, 300, 500, 1000)$, concentration parameter values $\kappa = (5, 10, 15, 20)$, ovalness parameter values $\beta = (0, 4, 8)$ and angles between the mean directions of the two samples $\phi = (0°, 20°, 50°)$[7].

Maximum likelihood estimation of the aforementioned densities and an implementation of the k-NN algorithm with spherical (and hyper-spherical) data are publicly available in the R package *Directional* (Tsagris et al. 2018) which we will utilize in our simulation and empirical evaluation studies. All computations took place in an HP laptop with Intel Core i5-5300U CPU @ 2.3GHz and 16 GB RAM.

[6]When it comes to large scale data, with tens or even hundreds of observations, computing the arc of the cosine of thousands of numbers becomes time consuming. We further discuss the computational cost in Section 3.5

[7]We have performed simulations with more angles, but the results are similar and due to the page limit we only show these numbers

3.1. *Monte Carlo estimation of the true miss-classification error*

Evaluating the true (theoretical) miss-classification error is difficult even in the Euclidean case, where many methods exist Ganeshanandam and Krzanowski (1990). In the spherical case, its computation is more complicated. This is why we will rely on Monte Carlo and evaluate the true miss-classification error empirically. Since we have two groups only, its computation is easier as it comprises of the proportion of wrongly classified observations to one group when in fact it belongs to the other group

$$P\left(C\right) = P(\mathbf{y} \text{ is classified to } \mathrm{G}_2 \mid \mathbf{y} \in G_1) + P(\mathbf{y} \text{ is classified to } \mathrm{G}_1 \mid \mathbf{y} \in G_2)$$

To empirically estimate this quantity we will simulate $10,000,000$ values from one group and compute the proportion of values that were miss-classified in the other group

$$\widehat{P\left(C\right)} = 0.5\left[\frac{\sum_{i=1}^{n}\mathbb{I}\left[\ell_2\left(\mathbf{y}_i\right) > \ell_1\left(\mathbf{y}_i\right) \mid \ell_1\right]}{10,000,000} + \frac{\sum_{i=1}^{n}\mathbb{I}\left[\ell_1\left(\mathbf{y}_i\right) > \ell_2\left(\mathbf{y}_i\right) \mid \ell_2\right]}{10,000,000}\right], \quad (12)$$

where $\mathbb{I}(.)$ is the indicator function and $\ell_j\left(\mathbf{y}_i\right)$, $j = 1, 2$ denotes the likelihood value of group j calculated at $\mathbf{y}_i$.

3.2. *Predictive performance estimation*

We conducted a 10-fold cross-validation (CV) using stratified random sampling ensuring that the group allocation was similar for each of the 10 folds. The estimated predictive performance of each algorithm was estimated using the percentage of correct classification[8]. The estimated predictive performance of the k-NN with the chosen k is optimistically biased (the performance of the chosen model is overestimated) and the bootstrap-based bias correction method (Tsamardinos, Greasidou, and Borboudakis 2018) described below was applied.

The predicted values produced by all methods across all k folds are collected in an $n \times M$ matrix P, where n is the number of observations and M the total number of strained models. We sample with replacement a fraction of rows (or predictions) from P and denote them by in-sample values, while the non re-sampled values (rows) are termed out-of-sample values. The performance of each algorithm in the in-sample values is calculated and the model with the optimal performance is selected, followed by the computation of its performance in the out-of-sample values. This process is repeated B (for example $B = 1000$) times and the average performance is returned.

3.3. *Spherical plots of Kent simulated data*

When the ovalness parameter (β) is equal to 0, the shape of the distributions of the two groups is similar. This indicates, that a von Mises-Fisher or the IAG distribution would not differ from the Kent or the ESAG distribution. As we move to higher ovalness values ($\beta \geq 4$), the difference between the rotational symmetric and the rotationally non-symmetric distributions appears. The elliptical shape of the data is apparent; this

[8] An alternative measure would be the area under the curve (AUC) (Hand and Till 2001). The benefit of AUC over the percentage of correct classification is that AUC is independent of the group allocation and provides a better view of the discriminative power of an algorithm. A drawback of AUC is that its theoretical value even when assuming a parametric model cannot be computed.

is the point where Kent and ESAG will start outperforming the von Mises-Fisher and the IAG distributions. A final point we must put emphasis on is the proximity between the Kent and the ESAG distribution. In Figures 3-5 the lines of ESAG and Kent are indistinguishable.

3.4. *Results of the simulation studies*

3.4.1. Simulated data from the Kent$(\boldsymbol{\gamma}, \kappa, 0)$ *distribution*

Figure 3 presents the estimated predictive performance of all methods for the 4 different values of concentration parameter and the 3 different angles. For all angles greater than 0°, and regardless of the concentration values, maximum likelihood discriminant analysis has outperformed the k-NN algorithm. The difference can be deemed negligible in some cases ($\approx 0.05\%$), whereas in other cases it is substantial ($\approx 4\%$). With regards to the empirical classification rate, the k-NN algorithm usually underestimates it, even when the sample sizes are large (1,000 observations).

3.4.2. Simulated data from the Kent$(\boldsymbol{\gamma}, \kappa, 4)$ *distribution*

Figure 4 presents the estimated predictive performance for the 3 different values of concentration parameter and the 3 different angles. When the angle between the two group mean directions is 0° all methods have good performance, their predicted percentage of correct classification moves about 50%. For all other angles, and regardless of the concentration values, maximum likelihood discriminant analysis has outperformed with the ESAG and Kent distributions perform better than their corresponding rotationally symmetric distributions, the IAG and vMF respectively. The k-NN algorithm is ranked in the middle of these two families of distributions. The difference can be deemed negligible in some cases ($\approx 0.05\%$), whereas in other cases it is substantial ($\approx 4\%$). Similarly to the previous case, the k-NN underestimates the empirical rate of correct classification.

3.4.3. Simulated data from the Kent$(\boldsymbol{\gamma}, \kappa, 8)$ *distribution*

Figure 5 presents the estimated predictive performance for the 3 different angles when the ovalness parameter and the concentration parameter are equal to 8 and 20 respectively. For all angles greater than zero ($\phi > 0°$), and regardless of the concentration values, the graphs show a behaviour similar to the one observed in Figure 4. Maximum likelihood discriminant analysis with rotationally symmetric distributions were ranked first in the estimated accuracy (percentage of correct classification). In this last case, all methods either underestimate or overestimate the empirical rate of correct classification, regardless of the true angle between the mean directions of the two samples.

3.5. *Computational cost of the algorithms*

We also estimated the time required to perform a 10-fold CV with each of the 6 alternatives. Maximum likelihood discriminant analysis using the vMF, IAG, ESAG and Kent distributions and the k-NN algorithm with either the standard or the non standard version. Figure 6 presents the estimated computational cost of each method/algorithm for a range of sample sizes varying from 200 up to 10,000. Maximum likelihood dis-

criminant analysis using the vMf, Kent or the IAG distributions are by far the fastest procedures.

The reason for this is that when obtaining the maximum likelihood estimates of the vMF and the Kent distribution, only one matrix multiplication occurs. The concentration and the ovalness parameters are then estimated using a numerical optimiser in R. On the contrary, MLE of the IAG distribution requires several matrix-vector multiplications explaining why this is slower. Nevertheless, the Newton-Raphson is given with appropriate starting values and the maximization is fast. Similarly, MLE in the ESAG distribution relies on the a numerical optimiser in R which internally performs matrix-vector multiplications.

The k-NN algorithm, on the other hand, computes distances between vectors, and hence it depends upon the sample size, but our C++ implementation is memory efficient; we can efficiently treat large sample sizes. For example, when computing the distances of each candidate vector from the available sample, we store the indices of the vectors with the k smallest distances and based upon them we allocate the candidate vector to a group. We give the option to predict the group membership for a range of values of k and parallel computations[9] can offer a significant boost in the speed.

Finally, when it comes to large scale (or massive) data, that cannot be loaded into R, the R package *bigmemory* (Kane et al. 2018) is necessary. It helps load the data onto R and allows for maximum likelihood discriminant analysis using the vMF and Kent distributions, and of the k-NN algorithm.

4. Real data analysis

We also compared the above methods using real spherical data, forming 15 pairs of samples, in order to obtain a better and more realistic image of the discriminant analysis.

4.1. *Description of the data*

Below we list the data and provide some information about them.

- **Paleomagnetic** (Wood 1982): A set of 33 estimates of a previous magnetic pole position (Table 2 in Schmidt (1976)) were obtained using palaeomagnetic techniques. Each estimate is associated with a different site coming from a large area of Tasmania. Following Figueiredo (2009) we will use the same labels for these data. The first group contains the observations $(9, 10, 11, 12, 14, 16, 23, 24, 30)$.
- **Ordovician** (Fisher, Lewis, and Embleton 1993): There are two groups of 50 measurements each from L_0^1 axes (intersections between cleavage and bedding planes of F, folds) in Ordovician turbidites, collected in the same sub-domain.
- **Stones** (Fisher, Lewis, and Embleton 1993): There are 202 measurements of the longest axis (101) and shortest axis (101) orientations of tabular stones on a slope at Windy Hills, Scotland.
- **Magnetic** (Fisher, Lewis, and Embleton 1993): These are measurements of magnetic remanence in specimens of Mesozoic Dolerite from Prospect, New South Wales, after successive partial demagnetisation stages (200° and 350°) for each of 62 specimens as part of an experiment to determine the blocking temperature spectrum of components of magnetisation.

[9] In our case we used a laptop with 4 cores, from which 2 are physical cores.

- **Vectorcardiogram** (Downs, Liebman, and Mackay 1971): The data are derived from vectorcardiogram measurements of the electrical activity of the heart of children of different ages (2-10 and 11-19 years old) and genders. Sometimes, used as a summary for clinical diagnosis, is a unit vector defined as the directional component of the vector at a particular extremum. The data come from for two different lead placement systems, the Frank system and for the McFee system for each of 98 children of different ages and gender.
- **Judgements** (Fisher, Lewis, and Embleton 1993): In a sociological study of the attitudes of 48 individuals to 16 different occupations, judgements were made according to 4 different criteria (Earnings, Social Status, Reward, Social Usefulness), giving rise to 4 samples (each of 48 multivariate measurements). From so-called external analysis of the occupational judgements, each multivariate measurement was reduced to a (spherical) unit vector, yielding the 4 samples of unit vectors. Eeach response was transformed to a unit vector according to 4 different criteria (Earnings, Social Status, Reward, Social Usefulness).
- **Midatlantic ridge**: This is data set which compares hand selected landmarks of contents and the midatlantic ridge. 70 landmarks of the Somalian and 70 landmarks of the Arabian plate are available.

The datasets **Vectorcardiogram** and **Judgements** had more than two groups. Since we wanted to draw conclusions for the two sample case, we considered all possible combinations of 2 samples for these two datasets, making a total of 15 pairs for all datasets. All pairs are graphically presented in Figures 7-8 and Table 1 presents the p-values of the rotational symmetry hypothesis test for each group of observations. The p-values were produced by using the log-likelihood ratio test for the IAG versus the ESAG.

4.2. *Predictive performance estimation*

To estimate the predictive performance of each method, we repeated the 10-fold CV protocol 50 times, on each real dataset for reducing the variability of the estimation. The exception is with the paleomagnetic dataset, for which the leave-one-out CV (LOOCV) was implemented due to its small sample size.

4.3. *Results of the empirical evaluation studies*

Table 1 presents the p-values of the rotational symmetry assumption for each sample, whereas Table 2 summarizes the predictive performance of each algorithm applied to each dataset. IAG, ESAG and the k-NN algorithm were the only ones that managed to outperform all the others in some datasets. Maximum likelihood discriminant analysis with the von Mises-Fisher or the Kent distribution were never selected as "winners" in any dataset. The k-NN algorithm produced the highest predictive performance in 9 out of 16 pairs, whereas IAG and ESAG MLE discriminant analysis were chosen in 6 out of 16 pairs. We cannot make a decision whether to use maximum likelihood discriminant analysis or the k-NN based on the whether the assumption of rotational symmetry is rejected or not. Hence the p-values in Table 1 do not seem to affect the performances of the algorithms in Table 2 and are rather unrelated.

Overall, the maximum likelihood discriminant analysis is on par with the k-NN algorithm. However, when the data are only separable in higher dimensions, as is the case with the midatlantic data (see Figure 12), the k-NN algorithm seems to be the

only option.

Table 1. P-values of the rotational symmetry tests of the IAG versus ESAG distributions for each group of the datasets. The test of von Mises versus Kent produced similar results and hence omitted.

Dataset	p-value	Dataset	p-value
Paleomagnetic group 1	0.632	Frank system boys	0.410×10^{-6}
Paleomagnetic group 2	0.096	Frank system girls	0.971
Ordovician group 1	0.139	McFee system boys	0.360×10^{-6}
Ordovician group 2	0.466	McFee system girls	0.684
Stones group 1	0.634	Judgements earnings	0.899×10^{-5}
Stones group 2	0.610×10^{-28}	Judgements social status	0.083
Magnetic group 1	0.511×10^{-27}	Judgements reward	0.569×10^{-6}
Magnetic group 2	0.880×10^{-27}	Judgements social usefulness	0.196×10^{-6}
Frank system 2-10 years	0.003	Midatlantic Somalian	0.019
Frank system 11-19 years	0.147	Midatlantic Arabic	6.33×10^{-18}
McFee system 2-10 years	0.227		
McFee system 11-19 years	0.019		

Figures 9-11 show the results of the repeated CV. For each pair of samples and for each method a boxplot of their estimated predictive performance allows for a comparison among them. In general, all methods exhibit the same variability more or less. A remarkable difference in the performances can be seen in Figures 10(a)-(c), where the k-NN algorithm has outperformed the maximum likelihood method. If we see Figures 8(a)-8(c) we can justify this behaviour. For example it is obvious that rotationally symmetric distributions cannot separate such data. By examining the other Figures, we can draw similar conclusions.

In the simulation studies we showed that the predictive performance of the ESAG and the Kent distributions is nearly equal. In conjunction with the computational efficiency of the Kent compared to the ESAG distribution (centiseconds to seconds) the conclusion is that Kent is to be preferred. The empirical evaluation study though supports the ESAG distribution over the Kent distribution. The ESAG distribution performed 9 times better than the Kent distribution, Kent performed better in 3 times and in the other 3 times they performed equally well.

Figure 12 shows the Midatlantic ridge data and the box plot of the cross-validate estimated of the performance of all methods. It can be seen that the two groups cannot be separated adequately by the parametric models considered. Surprisingly enough, the estimated accuracy of the ESAG and Kent distribution is as high as 80%. The k-NN algorithm has outperformed (by far) though the parametric models achieving almost perfect classification.

5. Conclusions

We compared maximum likelihood and the k-NN algorithm in the context of discriminant analysis with spherical data. The first method employed 4 distributions, 2 with rotational symmetry and 2 without this restrictive assumption.

In the simulation studies, maximum likelihood discriminant analysis with non-rotational symmetric distributions outperformed k-NN, but in many cases the difference between them was 1% or less. The extensive simulation studies showed that when the rotational symmetry holds, the choice of the distribution does not matter.

Table 2. Average estimated predictive performance of all methods based on repeated 10-fold CV. The highest performances are highlighted with bold.

Dataset	vMF	IAG	ESAG	Kent	S k-NN	NS k-NN
Paleomagnetic	0.970	0.970	0.939	0.939	**1.000**	**1.000**
Ordovician	0.570	**0.579**	0.525	0.509	0.432	0.440
Stones	0.867	0.875	0.891	0.888	**0.905**	0.903
Magnetic	0.528	0.518	0.492	0.520	**0.548**	0.538
Frank system age	0.608	**0.627**	0.566	0.584	0.566	0.559
McFee system age	0.603	**0.604**	0.582	0.595	0.555	0.528
Frank system gender	0.496	0.540	0.555	0.510	0.557	**0.562**
McFee system gender	0.529	**0.544**	0.510	0.502	0.486	0.455
Judgements earnings-social status	0.496	0.487	0.581	0.519	0.744	**0.749**
Judgements earnings-reward	0.416	0.384	0.757	0.686	0.790	**0.802**
Judgements earnings-social usefulness	0.593	0.581	0.752	0.735	0.855	**0.868**
Judgements social status-reward	0.545	0.549	**0.589**	0.582	0.526	0.546
Judgements social status-social usefulness	0.631	0.644	**0.664**	0.656	0.650	0.649
Judgements reward-social usefulness	0.551	0.554	0.550	0.550	0.583	**0.632**
Midatlantic ridge	0.590	0.582	0.780	0.794	0.993	**0.999**

When this assumption does not hold the more general distributions were clearly superior to the restrictive distributions. The Kent and ESAG distributions performed better than the von Mises-Fisher and IAG, as expected. This pattern was observed regardless of the concentration and the ovalness values. The k-NN algorithm on the other hand was always in between these two families of distributions. It was also shown to underestimate the true percentage of correct classification, even when the sample sizes were $1,000$ for each group.

In real data, the k-NN algorithm outperformed maximum likelihood discriminant analysis in most cases, providing strong evidence to support its use. The k-NN algorithm outperformed the maximum likelihood discriminant analysis in most cases. Among those cases, the non standard version of the k-NN was chosen 6 times, the standard version was chosen 2 times and they tied in one time. Further, when comparing between the two families of distributions only, the result was "tie", with and IAG and vMF "winning" exactly half of the times. These two rotationally symmetric distributions half of the times performed better than their rotationally non-symmetric distributions. Examining the boxplots of the real data more carefully one can see the superiority of the k-NN algorithm. When maximum likelihood discriminant analysis performed better, the difference with the k-NN algorithm was at most 10%. When k-NN outperformed the parametric models the difference could be more than 25% (see for example Figures 10(b) and 10(c)).

Relying heavily on the empirical evaluation studies we favour the k-NN algorithm because it is computationally efficient and scalable to large sample sizes and since we are living in the era of large scale or massive data, these two features are highly appreciable. Further, a prioritization scheme would be to use the k-NN algorithm first, followed by the Kent and ESAG distributions. To our surprise, the von Mises-Fisher and IAG distribution performed well and should be also utilised, for the task of discrimination/classification. When prioritizing the algorithms, rotational symmetric distributions should be used last. The ESAG and the Kent distributions are available in the R package *Directional* (Tsagris et al. 2018) but only for the spherical case. The k-NN algorithm on the other hand and the von Mises-Fisher distribution, also

available in the R package *Directional*, are applicable to higher dimensions. Even though our conclusions are limited to spherical data only, we managed to draw some useful conclusions. A natural question arises as to what should be the general strategy? Which results should one trust? We will weigh heavier the real data analysis results. Simulation studies can help validate a model when the assumptions hold true. They can help validate a model when the assumptions hold true. This is exactly our key argument and key point of discussion. In real life, the assumptions do not hold true and the need for a model or algorithm robust to model miss-specification, k-NN for example, is apparent. Ideally, we would like to suggest an algorithm or method that works well not only under ideal conditions, but works well in realistic scenarios. Real data will not obey any parametric assumptions and the noise to signal ratio can be really high.

Our final conclusion, based our evidence in the classification setting, is that one should employ numerous algorithms and methods; there is no panacea. Real data can be highly complex and there is need for development of more advanced machine learning discriminant analysis algorithms. In our case, the k-NN algorithm clearly outperformed maximum likelihood discriminant analysis.

Our future plans include a) develop more flexible algorithms for spherical data and b) adopt the current and future algorithms to large scale data, for example data with millions of observations that cannot be loaded onto R.

References

Abramowitz, Milton, and Irene Stegun. 1970. *Handbook of mathematical functions*. New York: Dover Publishing Inc.

Amayri, Ola, and Nizar Bouguila. 2013. "On online high-dimensional spherical data clustering and feature selection." *Engineering Applications of Artificial Intelligence* 26 (4): 1386–1398.

Amson, Eli, Patrick Arnold, Anneke H van Heteren, Aurore Canoville, and John A Nyakatura. 2017. "Trabecular architecture in the forelimb epiphyses of extant xenarthrans (Mammalia)." *Frontiers in Zoology* 14 (1): 52.

Audit, Benjamin, and Christos A Ouzounis. 2003. "From genes to genomes: universal scale-invariant properties of microbial chromosome organisation." *Journal of Molecular Biology* 332 (3): 617–633.

Banerjee, Arindam, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. "Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research* 6 (Sep): 1345–1382.

Cover, Thomas, and Peter Hart. 1967. "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory* 13 (1): 21–27.

Downs, Thomas, Jerome Liebman, and Wilma Mackay. 1971. "Statistical methods for vectorcardiogram orientations." *Vectorcardiography* 2: 216–222.

Figueiredo, Adelaide. 2009. "Discriminant analysis for the von Mises-Fisher distribution." *Communications in Statistics-Simulation and Computation* 38 (9): 1991–2003.

Fisher, Nicholas I, Toby Lewis, and Brian JJ Embleton. 1993. *Statistical analysis of spherical data*. Cambridge university press.

Fisher, Ronald Aylmer. 1953. "Dispersion on a sphere." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 217 (1130): 295–305.

Ganeshanandam, S, and WJ Krzanowski. 1990. "Error-rate estimation in two-group discriminant analysis using the linear discriminant function." *Journal of Statistical Computation and Simulation* 36 (2-3): 157–175.

Gill, Jeff, and Dominik Hangartner. 2010. "Circular data in political science and how to handle it." *Political Analysis* 18 (3): 316–336.

Hamsici, Onur C, and Aleix M Martinez. 2007. "Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification." *Journal of Machine Learning Research* 8 (Jul): 1583–1623.

Hand, David J, and Robert J Till. 2001. "A simple generalisation of the area under the ROC curve for multiple class classification problems." *Machine learning* 45 (2): 171–186.

Hornik, Kurt, Ingo Feinerer, Martin Kober, and Christian Buchta. 2012. "Spherical k-means clustering." *Journal of Statistical Software* 50 (10): 1–22.

Kane, M.J., J.W. Emerson, P. Haverty, and C. Jr. Determan. 2018. "bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped Files." R package version 4.5.33.

Kent, J.T. 1982. "The Fisher-Bingham Distribution on the Sphere." *Journal of the Royal Statistical Society, Series B* 44: 71–80.

Kent, J.T., A.M. Ganeiber, and K.V. Mardia. 2013. "A new method to simulate the Bingham and related distributions in directional data analysis with applications." *arXiv preprint arXiv:1310.8110* .

Laha, Arnab Kumar, and Sayan Putatunda. 2018. "Real time location prediction with taxi-GPS data streams." *Transportation Research Part C: Emerging Technologies* 92: 298–322.

López-Cruz, Pedro L, Concha Bielza, and Pedro Larrañaga. 2015. "Directional naive Bayes classifiers." *Pattern Analysis and Applications* 18 (2): 225–246.

Lund, Ulric. 1999. "Cluster analysis for directional data." *Communications in Statistics-Simulation and Computation* 28 (4): 1001–1009.

Mardia, K.V. 1975. "Statistics of Directional Data (with discussion)." *Journal of the Royal Statistical Society, Series B* 37: 349–393.

Mardia, K.V., and P.E. Jupp. 2000. *Directional Statistics*. John Wiley & Sons.

Morris, James Edward, and PJ Laycock. 1974. "Discriminant analysis of directional data." *Biometrika* 61 (2): 335–341.

Nelder, J.A., and R. Mead. 1965. "A simplex algorithm for function minimization." *Computer Journal* 7 (4): 308–313.

Paine, PJ, Simon P. Preston, Michail Tsagris, and Andrew T.A. Wood. 2018. "An elliptically symmetric angular Gaussian distribution." *Statistics and Computing* 28 (3): 689–697.

Paterson, John R, James G Gehling, Mary L Droser, and Russell DC Bicknell. 2017. "Rheotaxis in the Ediacaran epibenthic organism Parvancorina from South Australia." *Scientific Reports* 7: 45539.

Peel, David, William J Whiten, and Geoffrey J McLachlan. 2001. "Fitting mixtures of Kent distributions to aid in joint set identification." *Journal of the American Statistical Association* 96 (453): 56–63.

Rutkowska, Agnieszka, Silvia Kohnová, and Kazimierz Banasik. 2018. "Probabilistic properties of the date of maximum river flow, an approach based on circular statistics in lowland, highland and mountainous catchment." *Acta Geophysica* 1–14.

Schmidt, PW. 1976. "The non-uniqueness of the Australian Mesozoic palaeomagnetic pole position." *Geophysical Journal of the Royal Astronomical Society* 47 (2): 285–300.

Sra, S. 2012. "A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$." *Computational Statistics* 27 (1): 177–190.

Tsagris, M., G. Athineou, A. Sajib, E. Amson, and M.J. Waldstein. 2018. "Directional: directional statistics." R package version 3.3.

Tsamardinos, Ioannis, Elissavet Greasidou, and Giorgos Borboudakis. 2018. "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation." *Machine Learning* 107 (12): 1895–1922.

Vanni, Lorenzo, N Emilio Baldaccini, and Dimitri Giunchi. 2017. "Cue-conflict experiments between magnetic and visual cues in dunlin Calidris alpina and curlew sandpiper Calidris ferruginea." *Behavioral Ecology and Sociobiology* 71 (4): 61.

Watson, G.S. 1983. *Statistics on Spheres*. New York: Wiley.

Wood, Andrew. 1982. "A bimodal distribution on the sphere." *Applied Statistics* 31 (1): 52–58.

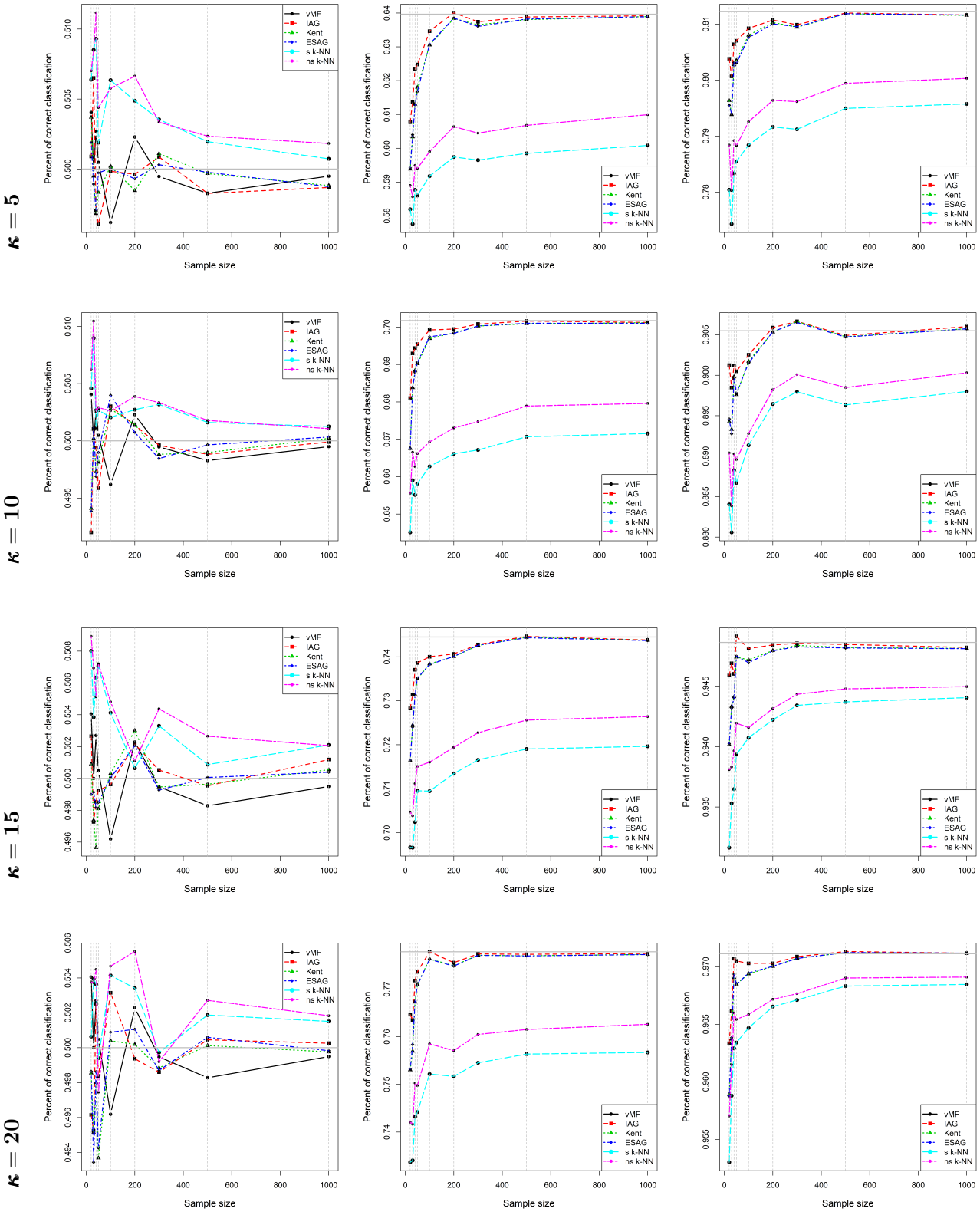


Figure 3. Estimated performance of each method for a variety of sample sizes, when the data have been generated from a Kent $(\boldsymbol{\gamma}, \kappa, 0)$ distribution, for different values of κ and ϕ. The grey line is the estimated percentage of correct classification (12).

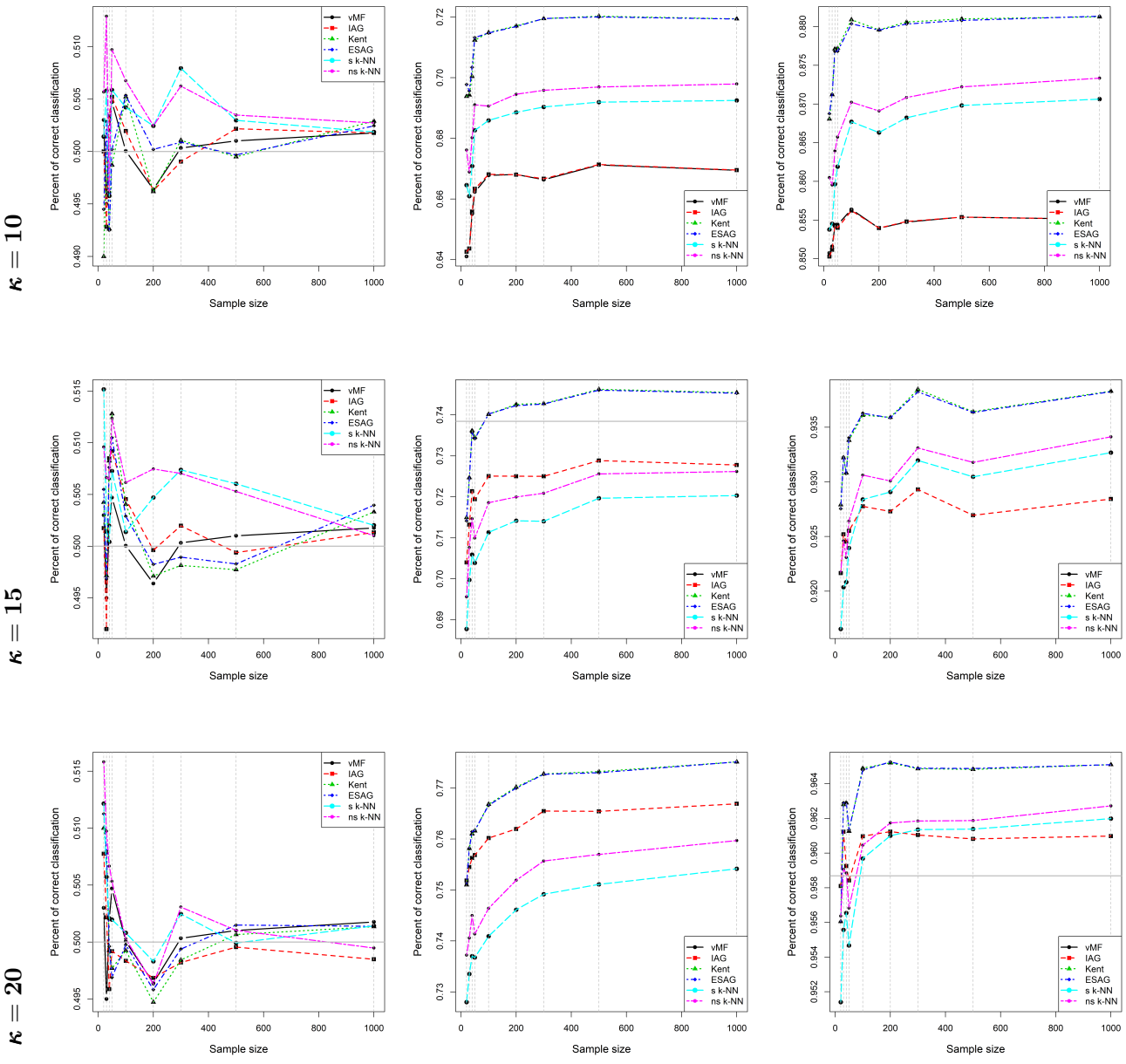
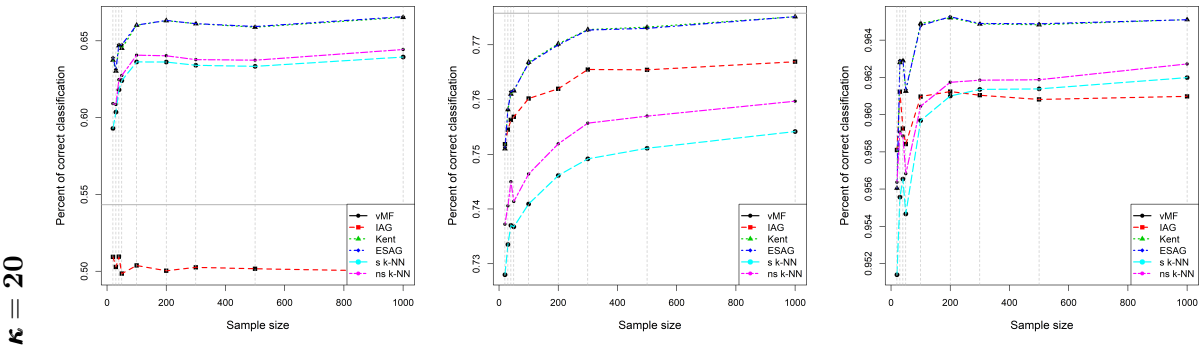


Figure 4. Estimated performance of each method for a variety of sample sizes, when the data have been generated from a Kent $(\boldsymbol{\gamma}, \kappa, 4)$ distribution, for different values of κ and ϕ. The grey line is the estimated percentage of correct classification (12).



$\phi = 0°$ $\phi = 20°$ $\phi = 50°$

Figure 5. Estimated performance of each method for a variety of sample sizes, when the data have been generated from a Kent $(\boldsymbol{\gamma}, 20, 8)$ distribution, for different values ϕ. The grey line is the estimated percentage of correct classification (12).

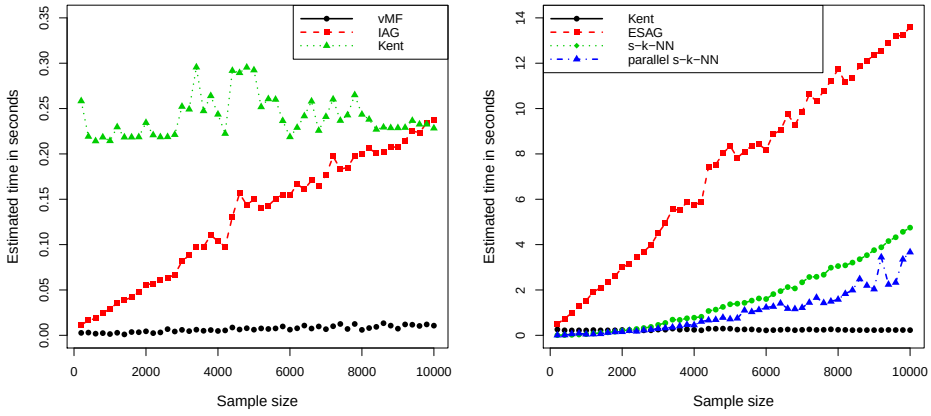


Figure 6. (a) Estimated time versus sample sizes for the vMF, IAG and Kent distributions. (b) Estimated time versus sample size for the ESAG and the k-NN algorithm, using the standard version (with and without parallel computations). The non standard version of the k-NN is not presented as it was 3 times slower than ESAG.

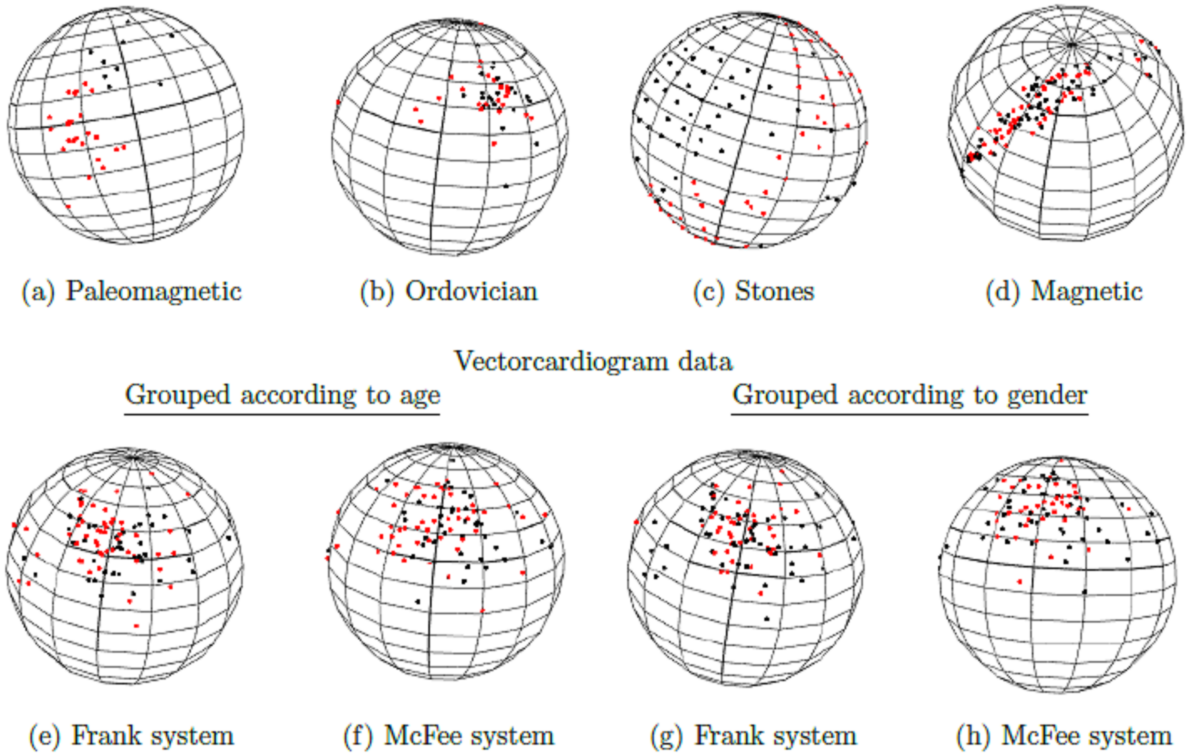


Figure 7. Spherical plots of the Paleomagnetic, Ordovician, stones, Magnetic and Vectorcardiogram data with different colours indicating the two groups.

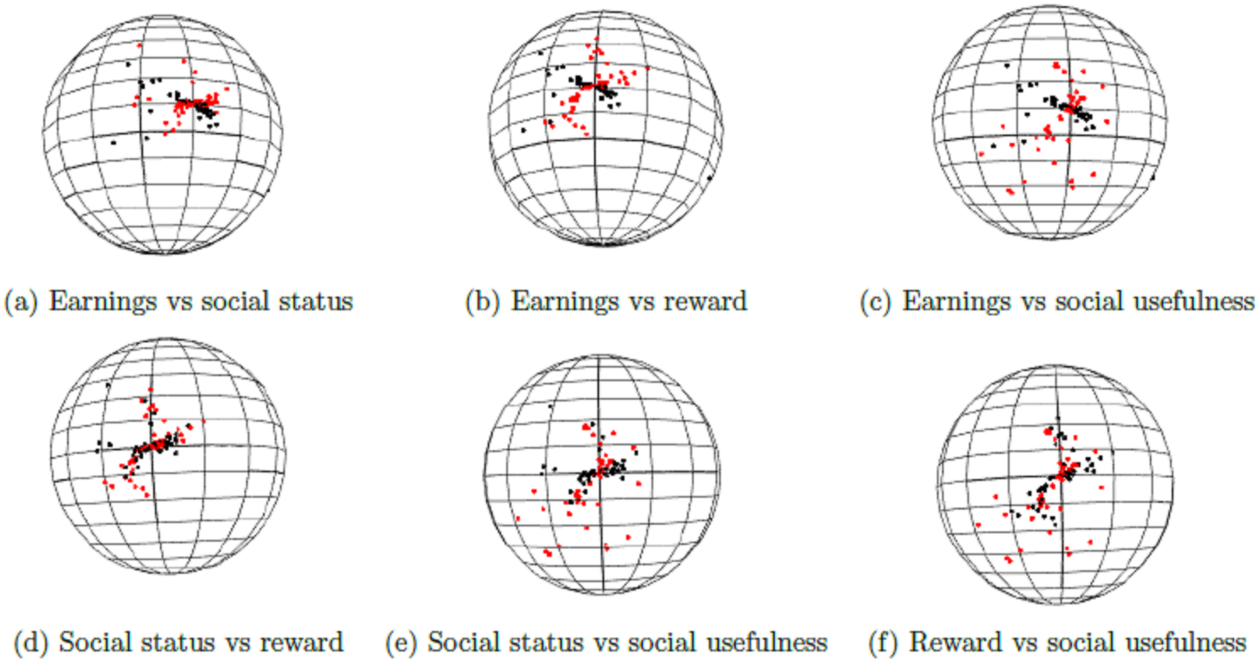
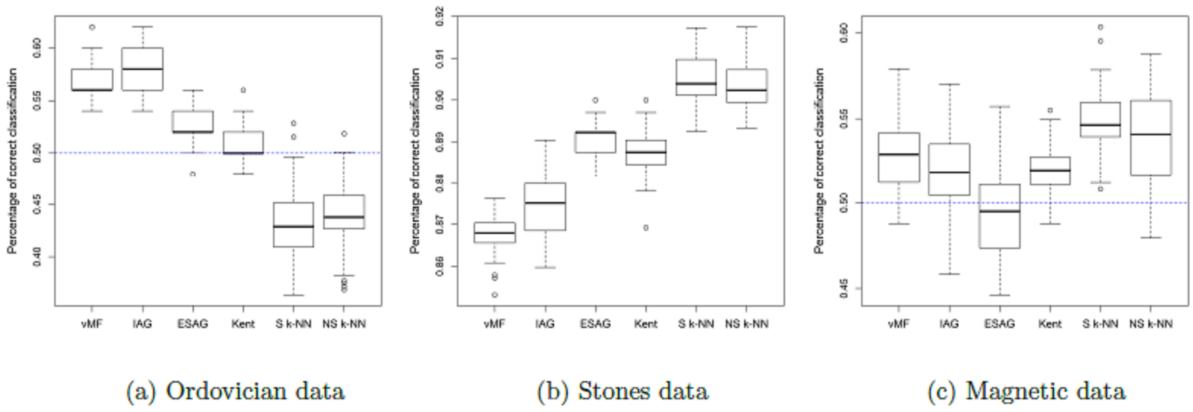
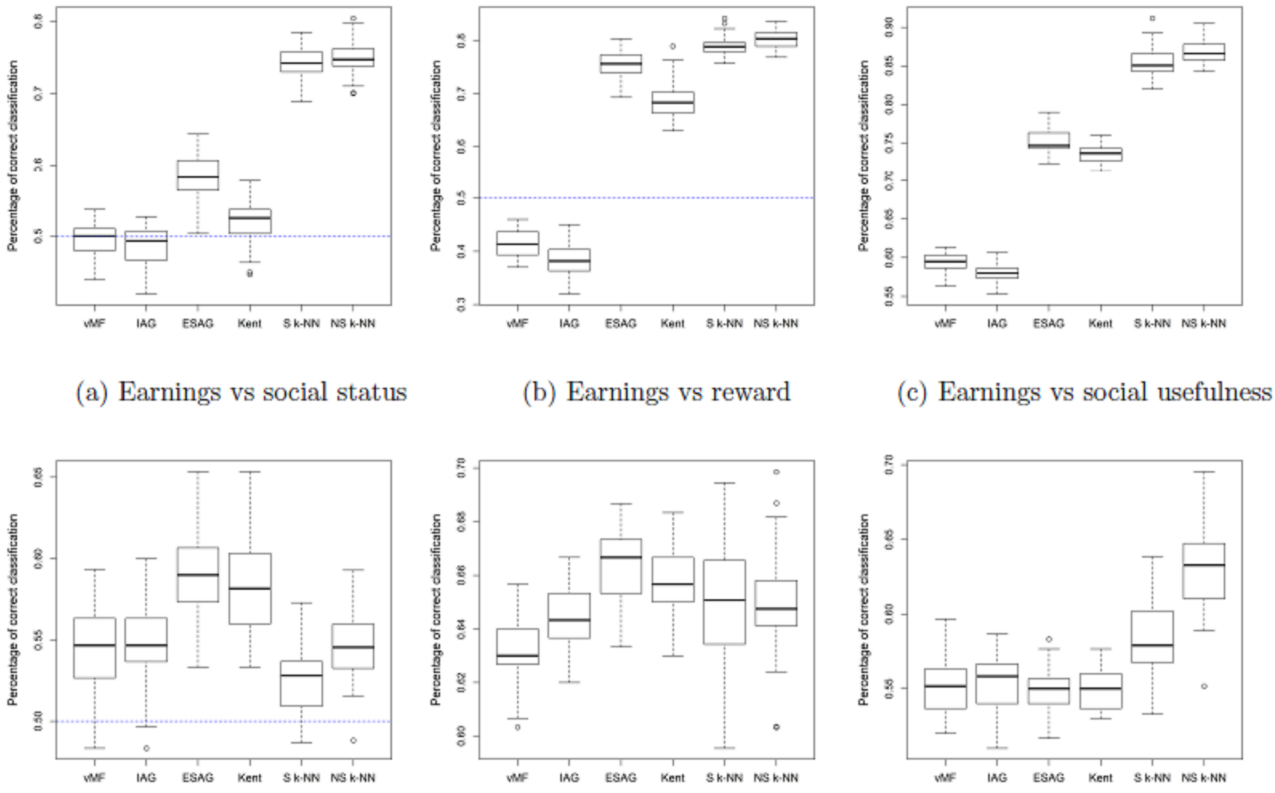


Figure 8. Spherical plots of the Judgements data with different colours indicating the two groups.



(a) Ordovician data (b) Stones data (c) Magnetic data

Figure 9. Box plots of the estimated predictive performance of all methods based on repeated 10-fold CV applied to the Paleomagnetic, Ordovician, stones, and Magnetic data.



(a) Earnings vs social status (b) Earnings vs reward (c) Earnings vs social usefulness

(d) Social status vs reward (e) Social status vs social usefulness (f) Reward vs social usefulness

Figure 10. Box plots of the estimated predictive performance of all methods based on repeated 10-fold CV applied to the Judgements data.

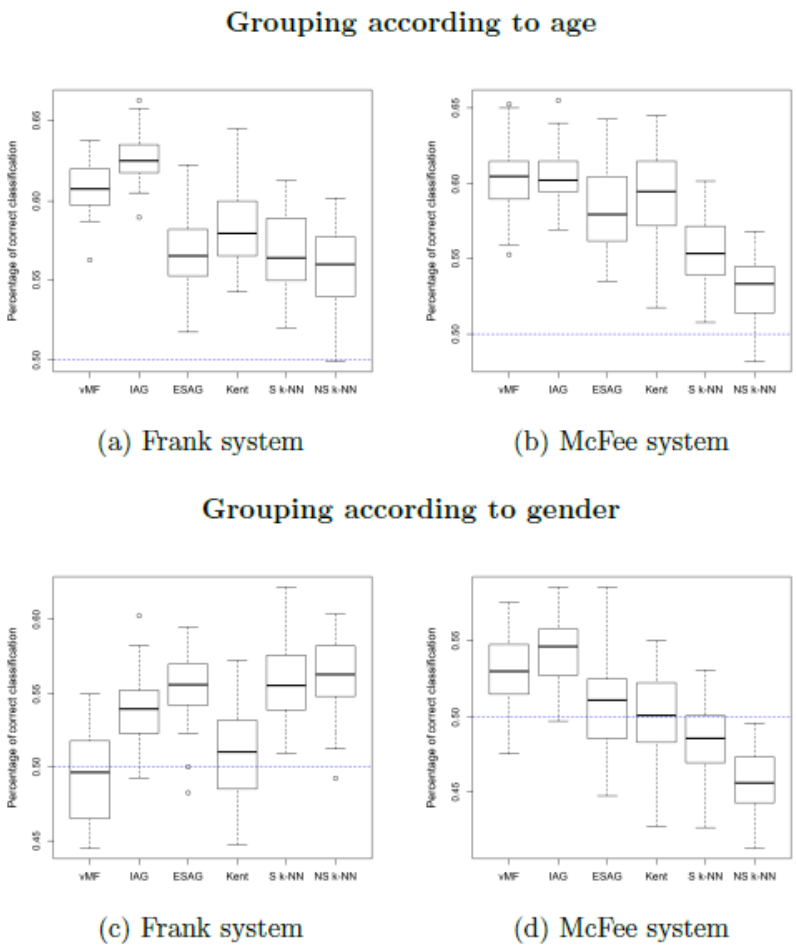


Figure 11. Box plots of the estimated predictive performance of all methods based on repeated 10-fold CV applied to the Vectorcardiogram data.

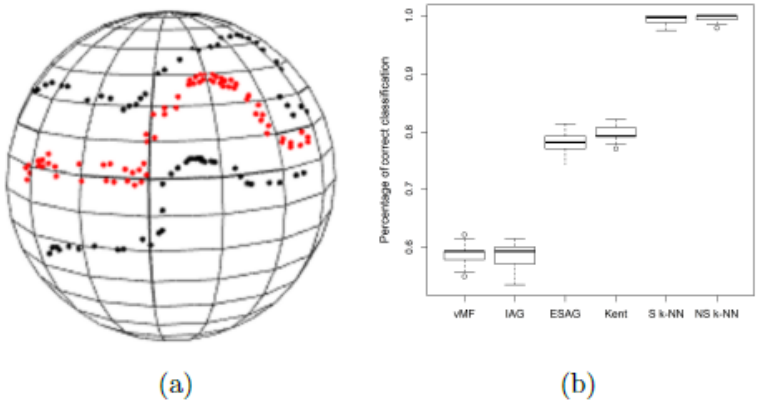


Figure 12. (a) Spherical plot of the Midatlantic data with different colours indicating the two groups. (b) Box plot of the estimated predictive performance of all methods based on repeated 10-fold CV applied to the Midatlantic data.