Discriminant analysis with spherical data

Michail Tsagris[1] and Abdulaziz Alenazi[2]

[1] Department of Economics, University of Crete, Greece, mtsagris@uoc.gr
[2] Northern Border University, Arar, Kingdom of Saudi Arabia, ksa99ksa99@hotmail.com

Abstract

Discriminant analysis for spherical data, or directional data in general, has not been extensively studied, and most papers focus on one distribution, the von Mises-Fisher. In this work, we study more distributions, escaping the rotational symmetry bound of the aforementioned distribution and also include a non parametric classifier, the k-NN algorithm. We draw important conclusions based on extensive simulation studies and real data empirical evaluations. The conclusions are bidirectional. When the parametric assumptions are met, maximum likelihood discriminant analysis using the Kent or the ESAG distribution is advised. This was the conclusion based on the simulation studies. If these conditions are not met, as is the case usually with real data, the k-NN algorithm should be used. This was supported by the real data analysis where the k-NN algorithm superseded maximum likelihood discriminant analysis.

Keywords: spherical data, rotationally non symmetric, classification

1 Introduction

Directional data, i.e. unit vectors, are multivariate data constrained to to have unit norm and hence live on a p-dimensional sphere. Their sample space, denoted by $\mathbb{S}^{p-1}$, is given by

$$\mathbb{S}^{p-1} = \left\{ \mathbf{Y} \in \mathbb{R}^{\mathbf{p}}, \quad \mathbf{Y^T Y = 1} \right\}$$

Such data arise in many different fields, such as biology (Paterson et al., 2017), zoology (Amson et al., 2017), ecology (Vanni et al., 2017), geophysics (Rutkowska et al., 2018) and transportation (Laha and Putatunda, 2018) to name a few.

Clustering (unsupervised learning) is the task of discovering groups of observations. In the case of directional data researchers have addressed this problem by either using hierarchical clustering (Lund, 1999) the k-Means algorithm (Hornik et al., 2012) or model based clustering with the von Mises-Fisher (Banerjee et al., 2005) or the Kent distribution (Peel et al., 2001). More recently, Amayri and Bouguila (2013) included the task of feature selection into model based clustering using, mixtures of von Mises-Fisher distributions.

With discriminant analysis (supervised learning, or classification) on the other hand, the group of each observation is known and unlike clustering, the literature is far less populated. Both (Morris and Laycock, 1974) and (Figueiredo, 2009) performed supervised learning and conducted simulation studies using the von Mises-Fisher distribution. Examples of classification with directional data can be found in many scientific fields. For example, classification of the wind direction according to the year's season (Mardia and Jupp, 2000). Classifying the

constitutes measurements of magnetic remanence in rock specimens, after each specimen had been partially thermally demagnetised to the same stage (Fisher et al., 1993). Separating between the longest axis and shortest axis orientations of tabular stones measured on a slope at Windy Hills, Scotland (Fisher et al., 1993).

The drawback of the aforementioned papers is that they attack the problem with limited stepping stones. In most papers, applied or not, the von Mises-Fisher distribution is used, perhaps due to its convenient form and easiness to work with. The von Mises-Fisher tough assumes independent variables, whereas the Kent distribution has elliptical contours, allowing for correlation between the variables. Yet, there are more spherical distributions than just these two and more algorithms than simple maximum likelihood discriminant analysis. However, no one, to the best of our knowledge, has studied supervised learning for directional (or even spherical) data using more than one distribution or even more algorithms.

In this paper we focus on discriminant analysis with spherical data expanding the work of Figueiredo (2009), by including three more distributions, the Independent Angular Gaussian, or projected normal, (Mardia and Jupp, 2000), the Kent distribution (Kent, 1982), and the Elliptically Symmetric Angular Gaussian distribution (Paine et al., 2018). In addition, the, non parametric, k-NN algorithm (Cover and Hart, 1967) coupled with the cosine distance is put in the testbed for comparison. Our goal is to provide evidence as to which distribution is more suitable and whether the k-NN algorithm should be employed for supervised learning with spherical data.

The next section of the paper contains some preliminaries regarding discriminant analysis; a) the spherical distributions we will examine and their maximum likelihood estimation of their parameters and b) the standard k-NN algorithm and a variant of it are described. Section 3 contains extensive simulation studies followed by real data analysis presented in Section 4. Finally, Section 5 concludes the paper.

2 Discriminant analysis with spherical data

Discriminant analysis is the task of constructing discriminating, separating, or allocation rules or boundaries between groups of observations. The difference with clustering is that the label of each observation, or the group to which each observation belongs is known. Therefore, one should be able to predict the label of a new observation based on the available data.

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 its parameters using maximum likelihood estimation. For each group, the density of a new observation is computed and the observation is allocated at the group with the highest density value.

Since we work with spherical data, we will present, below, without loss of generality, the distributions to be used, in their spherical parametrizations and how their parameters are estimated.

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)

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

where $\kappa \geq 0$ (concentration parameter) and $\boldsymbol{\gamma} \in \mathbb{S}^2$ is the mean direction.

The corresponding log-likelihood is given by

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

and maximum likelihood estimation of the parameters does not require numerical optimization. The estimated mean direction is available in closed form given by

$$\hat{\boldsymbol{\gamma}} = \frac{\bar{\mathbf{y}}}{\parallel \bar{\mathbf{y}} \parallel},$$

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 $\parallel \cdot \parallel$ denotes the Euclidean norm on $\mathbb{R}^3$. The concentration parameter is independent of the mean direction and requires a truncated Newton-Raphson algorithm[1] (Sra, 2012).

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

where

$$A_3\left(\hat{\kappa}^{(t-1)}\right) = \frac{I_{3/2}\left(\hat{\kappa}\right)}{I_{3/2-1}\left(\hat{\kappa}\right)},$$

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

2.1.2 The Isotropic Angular Gaussian distribution

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

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

where $M_2(\alpha) = (1+\alpha^2)\Phi(\alpha)+\alpha\phi(\alpha)$ with $\Phi(.)$ and $\phi(.)$ denoting the cumulative and probability density functions, respectively, of the standard normal distribution. $\boldsymbol{\mu} \in R^3$ is the mean vector

[1] The iterative solution in (2) is the general solution. For the spherical case, the Newton-Raphson obviously has a simpler solution.

[2] 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.

and $\mathbf{V}$ is a positive definite matrix. When $\mathbf{V} = \mathbf{I_3}$, the identity matrix in the three dimensions we end up with the Isotropic Angular Gaussian (IAG), or projected normal, distribution

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

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

Estimating the vector $\boldsymbol{\mu}$ requires numerical optimization of the corresponding log-likelihood (5) and Newton-Raphson can be employed

$$\boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t - \mathbf{H}^{-1}\mathbf{J},$$

where the first derivative ($\mathbf{J}$) and the Hessian matrix ($\mathbf{H}$) can be expressed as

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

where

$$\begin{aligned}
g_i'(\boldsymbol{\mu}) &= 2\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\Phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\mathbf{y}_i^T + 2\phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right)\mathbf{y}_i^T \quad \text{and} \\
g_i''(\boldsymbol{\mu}) &= 2\mathbf{y}_i\mathbf{y}_i^T\Phi\left(\mathbf{y}_i^T\boldsymbol{\mu}\right),
\end{aligned}$$

2.1.3 The Kent distribution

The density of the Fisher-Bingham distribution (Kume and Wood, 2005) is given by

$$f(\mathbf{y};\boldsymbol{\gamma},\kappa,\mathbf{A}) = \frac{1}{C(\kappa,\mathbf{A})}\exp\left(\kappa\mathbf{y}^T\boldsymbol{\mu} - \mathbf{y}^T\mathbf{A}\mathbf{y}\right), \tag{6}$$

where $\mathbf{A} = \mathbf{A}^T$ and $C(\kappa,\mathbf{A})$ is a normalising constant which, in the general case, does not have a useful closed-form expression (Kume and Wood, 2005). The Fisher-Bingham distribution (6) has 8 free parameters, 3 more than necessary.

The FB_5 or Kent distribution (Kent, 1982)

$$f(\mathbf{y};\boldsymbol{\gamma},\kappa,\beta) = \frac{1}{C(\kappa,\beta)}\exp\left\{\kappa\boldsymbol{\alpha}_1\mathbf{y} + \beta\left[\left(\boldsymbol{\alpha}_2\mathbf{y}\right)^2 - \left(\boldsymbol{\alpha}_3\mathbf{y}\right)^2\right]\right\}, \tag{7}$$

where β is the ovalness parameter, k is the concentration parameter, and $\boldsymbol{\alpha}_1$ (or $\boldsymbol{\gamma}$), $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are the mean direction, major and minor axis respectively. The angle of rotation ψ between the mean direction and the major axis is the fifth parameter of the Kent distribution.

The normalizing constant $C(\kappa,\beta)$ has a closed form in the spherical case, given by Kent (1982). In higher dimensions though, approximations have been suggested (Kume and Wood, 2005; Kume et al., 2013) and only recently exact calculation of this constant was achieved (Kume and Sei, 2018). To ensure ensure correct behaviour (uni-modality) of the density, the necessary condition is $|\beta| < \kappa/2$.

Its corresponding log-likelihood is

$$\ell = -nC(\boldsymbol{\gamma}, \kappa, \beta) + \kappa \sum_{i=1}^{n} \boldsymbol{\gamma}^T \mathbf{y}_i + \beta \left[\sum_{i=1}^{n} \left(\boldsymbol{\alpha}_2^T \mathbf{y}_i \right)^2 - \sum_{i=1}^{n} \left(\boldsymbol{\alpha}_3^T \mathbf{y}_i \right)^2 \right]. \quad (8)$$

To estimate the orthogonal matrix $\mathbf{A}$ we will use the moment estimation (Kent, 1982). 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

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

where θ and ϕ are the polar co-ordinates of $\bar{\mathbf{y}}$. Let $\mathbf{B} = \mathbf{H}^T\mathbf{S}\mathbf{H}$, where $\mathbf{S} = n^{-1}\sum \mathbf{y}_i\mathbf{y}_i^T$. 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 $l_1 > l_2$. If we choose ψ such that $\tan(2\psi) = 2b_{23}/(b_{22} - b_{33})$, ensuring that $\| \bar{\mathbf{y}} \| > 0$ and $l_1 > l_2$ 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 (7) with respect to these two parameters using the command *optim* in R.

2.1.4 The ESAG distribution

The Elliptically Symmetric Angular Gaussian (ESAG) distribution was recently defined (Paine et al., 2018) is a non rotationally symmetric distribution

$$f(\mathbf{y}; \boldsymbol{\mu}, \mathbf{V}) = \frac{1}{2\pi \left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \right)^{3/2}} \times \exp\left\{ \frac{1}{2} \left[\frac{\left(\mathbf{y}^T\boldsymbol{\mu} \right)^2}{\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \right)} - \left(\boldsymbol{\mu}^T\boldsymbol{\mu} \right) \right] \right\} \times M_2 \left[\frac{\left(\mathbf{y}^T\boldsymbol{\mu} \right)^2}{\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \right)} \right]. \quad (9)$$

The log-likelihood of (9) is given by

$$\begin{aligned} \ell &= -n\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n} \log\left(\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i \right) + \frac{1}{2}\sum_{i=1}^{n} \frac{\left(\mathbf{y}_i^T\boldsymbol{\mu} \right)^2}{\left(\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i \right)} - \frac{n}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\mu} \right) \\ &\quad + \sum_{i=1}^{n} M_2 \left[\frac{\left(\mathbf{y}_i^T\boldsymbol{\mu} \right)^2}{\left(\mathbf{y}_i^T\mathbf{V}^{-1}\mathbf{y}_i \right)} \right]. \end{aligned} \quad (10)$$

ESAG (9) 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 eigenvalues are $0 < \rho_1 \leq \rho_2$, and hence $\mathbf{V}^{-1}$ can be written as

$$\mathbf{V}^{-1} = \xi_d\xi_d^T + \sum_{j=1}^{d-1} \xi_j\xi_j^T/\rho_j, \quad (11)$$

where $\xi_1, \ldots, \xi_{d-1}$ and $\xi_d = \mu/\|\mu\|$ is a set of mutually orthogonal unit vectors. The second condition implies $\prod_{j=1}^{d-1} \rho_j = 1$. 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. The total number of free parameters is thus 5, just like in Kent distribution (7), the same as for the trivariate normal in a tangent space $\mathbb{R}^2$ to the sphere.

A more convenient parameterisation for the covariance matrix $\mathbf{V}$ exists, such that $\mathbf{V}$ has 2 free parameters. Let us define the two unit vectors $\tilde{\xi}_1$ and $\tilde{\xi}_2$ which are orthogonal to each other and to the mean direction $\xi_3 = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$:

$$\begin{aligned} \tilde{\xi}_1 &= \left(-\mu_0^2, \mu_1\mu_2, \mu_1\mu_3\right)^T /(\mu_0\|\boldsymbol{\mu}\|) \quad \text{and} \\ \tilde{\xi}_2 &= (0, -\mu_3, \mu_2)^T /\mu_0, \end{aligned} \tag{12}$$

where $\mu_0 = (\mu_2^2+\mu_3^2)^{1/2}$; then $\tilde{\xi}_1$ and $\tilde{\xi}_2$ in (12) are smooth functions of $\boldsymbol{\mu}$ except at $\mu_2 = \mu_3 = 0$, where there is indeterminacy. To enable the axes of symmetry, ξ_1 and ξ_2, to be an arbitrary rotation of $\tilde{\xi}_1$ and $\tilde{\xi}_2$, we can define

$$\begin{aligned} \xi_1 = &\quad \cos\psi\,\tilde{\xi}_1 &+ &\quad \sin\psi\,\tilde{\xi}_2 \\ \xi_2 = &-\sin\psi\,\tilde{\xi}_1 &+ &\quad \cos\psi\,\tilde{\xi}_2, \end{aligned} \tag{13}$$

where $\psi \in (0, \pi]$ is the angle of rotation. Substituting ξ_1 and ξ_2 from (13) into (11), and putting $\rho_1 = \rho$ and $\rho_2 = 1/\rho$ where $\rho \in (0, 1]$, gives the parameterisation

$$\begin{aligned} \mathbf{V}^{-1} &= \left(\rho^{-1}\cos^2\psi + \rho\sin^2\psi\right)\tilde{\xi}_1\tilde{\xi}_1 + \left(\rho^{-1}\sin^2\psi + \rho\cos^2\psi\right)\tilde{\xi}_2\tilde{\xi}_2^T \\ &\quad + \frac{1}{2}(\rho^{-1} - \rho)\sin 2\psi\left(\tilde{\xi}_1\tilde{\xi}_2^\top + \tilde{\xi}_2\tilde{\xi}_1\right) + \xi_3\xi_3^\top. \end{aligned} \tag{14}$$

To overcome the disadvantage that ρ and ψ are restricted, Paine et al. (2018) used the unrestricted parameters γ_1 and γ_2

$$\gamma_1 = 2^{-1}(\rho^{-1} - \rho)\cos 2\psi \quad \text{and} \quad \gamma_2 = 2^{-1}(\rho^{-1} - \rho)\sin 2\psi.$$

Then $\mathbf{V}^{-1}$ in (14) becomes

$$\mathbf{V}^{-1} = I_3 + \gamma_1\left(\tilde{\xi}_1\tilde{\xi}_1^T - \tilde{\xi}_2\tilde{\xi}_2^T\right) + \gamma_2\left(\tilde{\xi}_1\tilde{\xi}_2^T + \tilde{\xi}_2\tilde{\xi}_1^T\right) + \left\{(\gamma_1^2 + \gamma_2^2 + 1)^{1/2} - 1\right\}\left(\tilde{\xi}_1\tilde{\xi}_1^T + \tilde{\xi}_2\tilde{\xi}_2^T\right).$$

Unfortunately, the derivatives of (10) are not available and MLE of the ESAG distribution is implemented using a numerical optimizer, such as the Nelder-Mead algorithm (Nelder and Mead, 1965), available in R via the command *optim*.

2.2 The maximum likelihood discriminant boundaries

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

$$\text{Allocate } \mathbf{x} \text{ to group 1 iff } \ell_1(\mathbf{x}) > \ell_2(\mathbf{x}). \tag{15}$$

The rule (15) translates into

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

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

- For the IAG allocate $\mathbf{x}$ to group 1 iff

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

- For the Kent allocate $\mathbf{x}$ to group 1 iff

$$\begin{aligned} &-C\left(\boldsymbol{\gamma}_1, \kappa_1, \beta_1\right) + C\left(\boldsymbol{\gamma}_2, \kappa_2, \beta_2\right) + \left(\kappa_1 \boldsymbol{\gamma}_1^T - \kappa_2 \boldsymbol{\gamma}_2^T\right) \mathbf{x} \\ &+\beta_1\left[\left(\boldsymbol{\alpha}_2^{1T} \mathbf{x}\right)^2 - \left(\boldsymbol{\alpha}_3^{1T} \mathbf{x}\right)^2\right] - \beta_2\left[\left(\boldsymbol{\alpha}_2^{2T} \mathbf{x}\right)^2 - \left(\boldsymbol{\alpha}_3^{2T} \mathbf{x}\right)^2\right] > 0. \end{aligned}$$

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

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

Looking at the above inequalities it is straightforward to simplify them 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 discriminant analysis model in $\mathbb{R}^3$. But, the von Mises-Fisher is already restrictive; there is no reason to further restrict the boundary. More generally, parametric discriminant analysis puts constrains on the shape of the data which might be unrealistic as well, but this is a discussion which we will engage later.

2.3 Characteristics of the above densities

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 R^2 with isotropic or spherical covariance matrix. Their contour plot would consists of many concentric circles as presented in Figure 1(a). The Kent and ESAG distributions on the other hand allows for elliptical contours (see Figure 1(b) and 1(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 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, $\kappa \approx \| \boldsymbol{\mu} \|$, with their

main difference lying in their construction. The von Mises-Fisher is a multivariate normal with a covariance matrix equal to the identity matrix conditioned to lie on the unit (hyper-)sphere, $\mathbf{Y} \sim \text{vMF}(\boldsymbol{\gamma}, \kappa) \equiv N_3\left(\boldsymbol{\mu}, \mathbf{I}_3 | \mathbf{Y}^T\mathbf{Y} = 1\right)$ whereas the IAG is a multivariate normal 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 (1) comes the Kent distribution (7) 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 distribution is a special case of the Fisher-Bingham distribution (Bingham, 1974). Instead of 8 parameters, it has 5 (hence the name FB_5). The 5th parameter is the angle of rotation ψ between the mean direction and the major axis. The ESAG distribution, similarly to the Kent, comes from the more general AG distribution (3) with proper constraints on the covariance matrix (Paine et al., 2018). The proximity between the Kent and the ESAG can be graphically seen in the contour plots in Figure 1(c).

Maximum likelihood estimation of the above densities is publicly available in the R package *Directional* (**?**) which we will utilize in our simulation studies.

2.4 The k-NN algorithm

The k-NN (k Nearest Neighbours) algorithm is an intuitive classifier that assumes no parametric model. It involves determining the k observations in the training sample that are closest, by some choice of distance measures, to the new test observation, then allocating the test observation to the group most common amongst these k "nearest neighbours". Ties caused by two or more groups jointly being most common can be broken by allocating uniformly at random amongst the tied groups (the strategy we use in our simulation studies or else by using a secondary tie-breaking criterion. This is the standard k-NN. n this work we also used a variant of this which calculates the distances of the k "nearest neighbours" of the test observation from each group and allocates it to the group with the smallest average distance. This is the non-standard k-NN and is computationally more expensive, but might lead to better performance.

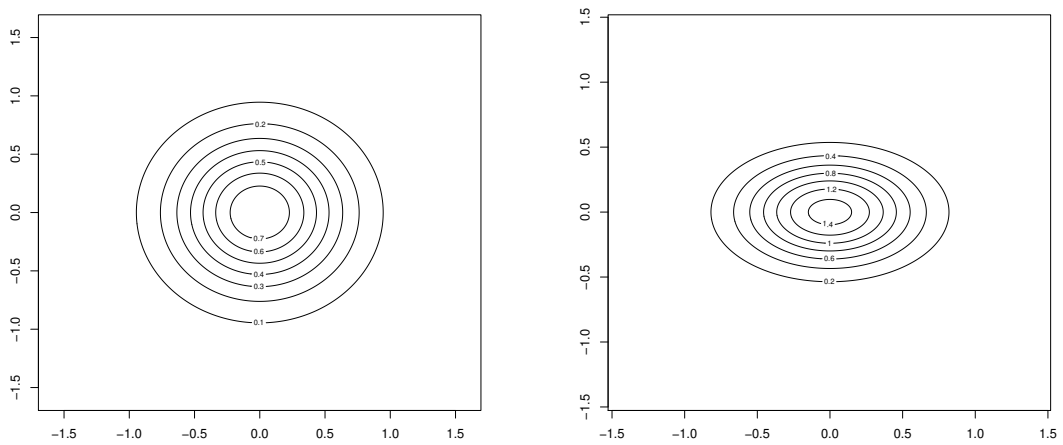
Performance of k-NN depends of the choice of k: small k allows for classification boundaries which are flexible but which have a tendency to overfit, with the opposites true when k is large. It also depends on the choice of distance measure. Since we are dealing with directional data we shall use the cosine distance (or cosine similarity)

$$D(\mathbf{X}_i, \mathbf{X}_j) = \cos^{-1}\left(\mathbf{X}_i^T\mathbf{X}_j\right). \tag{16}$$

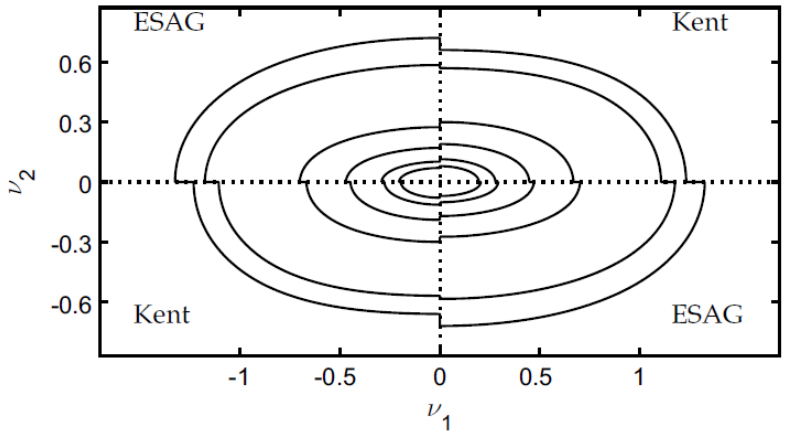
When the angle between the two vectors is 0, the (16), their inner product is 1 and the arc of the cosine is 0°. The maximum value of (16) is achieved when the the angle between the two vectors is 180°, their inner product is -1 and the hence the arc of the cosine is $\pi/2$.

3 Simulation studies

We have conducted extensive simulation studies so as to produce useful and helpful conclusions, but also 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



(a) (b)



(c)

Figure 1: Contour plots of the von Mises-Fisher, the Kent and a comparison of the Kent with the ESAG distribution.

and cases to be examined. In specific, we examined many different combinations of sample sizes $n = (50, 100, 150, 200, 300, 500, 1000)$, concentration parameter values $\kappa = (0, 5, 10, 15, 20)$, ovalness parameter values $\beta = (0, 2, 4, 6, 8)$ and angles between the mean directions of the two samples $\phi = (0^\circ, 10^\circ, 20^\circ, 30^\circ, 40^\circ, 50^\circ, 60^\circ)$. Our goal is to cover many situations which can arise in practice.

3.1 Monte Carlo estimation of the true miss-classification error

Evaluating the true (theoretical) miss-classification error is difficult even in the multivariate 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 to empirically evaluate the true error. Since we have two groups only, its computation is easier. The error consists of two elements, wrongly classifying an observation to one group when in fact it belongs

to the other group

$$\begin{aligned} P(E) &= P(\mathbf{y} \text{ is classified to } \mathrm{G}_2 \mid \mathbf{y} \text{ belongs to } \mathrm{G}_1) + P(\mathbf{y} \text{ is classified to } \mathrm{G}_1 \mid \mathbf{y} \text{ belongs to } \mathrm{G}_2) \\ &= P(G_2|G_1) + P(G_1|G_2) \end{aligned}$$

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

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

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

3.2 Predictive performance estimation

We conducted a 10-fold cross-validation 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, in order to compare with the empirical miss-classification error[3].

In each fold, both versions of the k-NN algorithm (the standard and the non standard) were tested using a range of k values. The estimated predictive performance of the best k is optimistically biased (the performance of the chosen model is overestimated) and a bootstrap-based bias correction method was applied Tsamardinos et al. (2018).

The predicted values produced by all methods across all k folds are collected in a matrix P of dimensions $n \times M$, where n is the number of samples and M the number of trained models. We sample with replacement a fraction of rows (or predictions) from P and denote them by in-sample values. On average, the newly created set will be comprised by 63.2% of the original individuals Efron and Tibshirani (1994). The non re-sampled values (rows) can be termed out-of-sample values. The performance of each algorithm in the in-sample values is calculated and the model (or configuration) with the optimal performance is selected, followed by the calculation of performance in the out-of-sample values. This process is repeated $B = 1000$ times and the average performance is returned.

The only computational overhead is with the repetitive re-sampling and calculation of the predictive performance, i.e. no model is fitted nor trained. A possible drawback is that the final estimated performance usually underestimates the true performance, but this negative bias is smaller than the optimistic uncorrected performance.

3.3 Spherical plots of the Kent simulated data

Figure 2 presents simulated data of some cases examined in the simulation studies. Simulated data from the Kent distribution are plotted on the sphere. These can be used as guides as to what to expect from the simulation studies.

[3] 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 when assuming a parametric model cannot be computed.

When the ovalness parameter (β) is equal to 0 or even equal to 2, the shape of the data 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 rotational non symmetric distributions appears. The elliptical shape of the data is apparent; this is the point where Kent and ESAG will start outperforming the von Mises-Fisher and the IAG distributions.

3.4 Results of the simulation studies

3.4.1 Simulated data from the Kent $(\boldsymbol{\gamma}, \kappa, 0)$ ($\equiv$ von Mises-Fisher) distribution

Figures 3 and 4 present the results of the estimated predictive performance for all methods for the 4 different values of concentration parameter and the 6 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 the k-NN algorithm. The difference can be deemed negligible in some cases ($\approx 0.05\%$), whereas in other cases it is substantial ($\approx 4\%$).

3.4.2 Simulated data from the Kent $(\boldsymbol{\gamma}, \kappa, 2)$ distribution

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

3.4.4 Simulated data from the Kent $(\boldsymbol{\gamma}, \kappa, 6)$ distribution

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

4 Real data analysis

In order to have a better and more realistic image of the discriminant analysis we also compared the above methods using real spherical data[4]. We managed to form 20 pairs of groups, hoping to obtain evidence and insight on how discriminant analysis performs with real 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)) obtained using palaeomagnetic techniques. Each estimate is associated with a different site, the 33 sites being spread over 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 et al., 1993): Two groups of 50 measurements each, of L_0^1 axes (intersections between cleavage and bedding planes of F, folds) in Ordovician turbidites, collected in the same sub-domain.

- **Stones** (Fisher et al., 1993): 202 measurements of the longest axis (101) and shortest axis (101) orientations of tabular stones on a slope at Windy Hills, Scotland.

[4]All datasets are available upon request from the corresponding author.

- **Magnetic** (Fisher et al., 1993): Measurements of magnetic remanence in specimens of Mesozoic Dolerite from Prospect, New South Wales, after successive partial demagnetisation stages (200° and $35O^\circ$) for each of 62 specimens as part of an experiment to determine the blocking temperature spectrum of components of magnetisation.

- **Vectorcardiogram** (Downs et al., 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. The vectorcardiogram involves three leads being connected to the torso produce a time-dependent vector that traces approximately closed curves, each representing a heartbeat cycle, in $\mathbb{R}^3$. Sometimes used as a summary for clinical diagnosis is a unit vector defined as the directional component of the vector at a particular extremum across the cycles. The data comprise such unit-vectors derived from data for two different lead placement systems, the Frank system and for the McFee system for each of 98 children of different ages and gender. We will examine both systems independently using the age groups and the gender groups. These combinations produce 4 pairs of datasets in total.

- **Judgments** (Fisher et al., 1993): In a sociological study of the attitudes of 48 individuals to 16 different occupations, judgments 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 judgments, 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).

Table 1 presents the p-values of the rotational symmetry hypothesis test for each group of observations. The p-values were produced by using Rivest's (Rivest, 1986) test for the von Mises-Fisher versus the Kent and the log-likelihood ratio test for the IAG versus the ESAG.

To estimate the predictive performance of each method, we repeated the 10-fold CV protocol 50 times, on each real dataset [5] for reducing the variability of the estimation.

5 Conclusions

We compared maximum likelihood and the k-NN algorithm in the context of discriminant analysis. The first method employed 4 distributions, 2 with rotational symmetry and 2 without this assumption. Our extensive simulation studies and the empirical evaluation studies allowed us to draw multiple conclusions.

In the simulation studies, when comparing parametric discriminant analysis to the non parametric k-NN algorithm, the former always produced better results than the latter. This pattern was observed regardless of the concentration and the ovalness values. With regards to the distributions, when the rotational symmetry holds true, all distributions produced similar results. When the assumption did not hold, the Kent and ESAG distributions performed better

[5] The exception is with the paleomagnetic dataset, for which the leave-one-out CV (LOOCV) was implemented due to its small sample size.

Dataset	von Mises-Fisher Vs Kent	IAG Vs ESAG
Paleomagnetic group 1	0.656	0.632
Paleomagnetic group 2	0.109	0.096
Ordovician group 1	0.106	0.139
Ordovician group 2	0.379	0.466
Stones group 1	0.267	0.634
Stones group 2	0.20×10^{-25}	0.610×10^{-28}
Magnetic group 1	0.188×10^{-12}	0.511×10^{-27}
Magnetic group 2	0.128×10^{-12}	0.880×10^{-27}
Frank system 2-10 years	0.002	0.003
Frank system 11-19 years	0.094	0.147
McFee system 2-10 years	0.045	0.227
McFee system 11-19 years	0.005	0.019
Frank system boys	0.450×10^{-5}	0.410×10^{-6}
Frank system girls	0.410	0.971
McFee system boys	0.807×10^{-5}	0.360×10^{-6}
McFee system girls	0.218	0.684
Judgments earnings	0.295×10^{-5}	0.899×10^{-5}
Judgments social status	0.217	0.083
Judgments reward	0.315×10^{-5}	0.569×10^{-6}
Judgments social usefulness	0.943×10^{-5}	0.196×10^{-6}

Table 1: P-values of the rotational symmetry tests, von Mises-Fisher versus Kent and IAG versus ESAG distributions for each group of the datasets.

than the von Mises-Fisher and IAG, as expected. The k-NN was shown to underestimate the true percentage of correct classification, even when the sample sizes were $1,000$ for each group.

In the empirical evaluation studies, the conclusions were the opposite. The k-NN algorithm outperformed the maximum likelihood discriminant analysis. In real life it is rather unusual to find datasets following a parametric model and perhaps this is why the distribution free k-NN algorithm performed so well.

A natural question arises as to what should be the general strategy? Which results should one trust? We will put more weight on the real data analysis results. With simulation studies, it is hard to evaluate the true performance of a classifier, even if the data are generated from distributions with strange shapes.[6] This is because real data will not obey any parametric assumptions and the noise to signal ratio can be really high. In the classification setting, we believe one should gather and use as many real data as possible to compare their methods. Simulation studies can help validate a model when the assumptions hold true. And this is exactly our point of discussion. In real life, the assumptions do not hold true. Hence, there is need for a model or algorithm robust to model miss-specification and k-NN is such an example.

Based on our findings, 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 dis-

[6]Our goal is not to suggest an algorithm or method that works well under ideal conditions, but works well in realistic scenarios.

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
Judgments earnings-social status	0.496	0.487	0.581	0.519	0.744	**0.749**
Judgments earnings-reward	0.416	0.384	0.757	0.686	0.790	**0.802**
Judgments earnings-social usefulness	0.593	0.581	0.752	0.735	0.855	**0.868**
Judgments social status-reward	0.545	0.549	**0.589**	0.582	0.526	0.546
Judgments social status-social usefulness	0.631	**0.644**	**0.664**	0.656	0.650	0.649
Judgments reward-social usefulness	0.551	0.554	0.550	0.550	0.583	**0.632**

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

tribution performed well and should be also utilised, for the task of discrimination/classification. Unfortunately our conclusions are limited to spherical data only. In addition, the ESAG distribution though was only recently suggested (Paine et al., 2018) and has not been extended to higher dimensions. The Kent distribution on the other hand has been extended (Scealy and Welsh, 2011), yet it is not available in any R package. The k-NN algorithm on the other hand and the von Mises-Fisher distribution, available in the R package *Directional* (**?**), are applicable to higher dimensions.

Closing this paper we will mention that real data can be highly complex, hence more advanced discriminant analysis algorithms should be used. The machine learning field is rich in such algorithms and statisticians coping with spherical (or hyper-spherical) should borrow, or at least consider them.

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of mathematical functions*. New York: Dover Publishing Inc.

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

Amson, E., Arnold, P., van Heteren, A. H., Canoville, A., and Nyakatura, J. A. (2017). Trabecular architecture in the forelimb epiphyses of extant xenarthrans (mammalia). *Frontiers in zoology*, 14(1):52.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere

using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Downs, T., Liebman, J., and Mackay, W. (1971). Statistical methods for vectorcardiogram orientations. *Vectorcardiography*, 2:216–222.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

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

Fisher, N. I., Lewis, T., and Embleton, B. J. (1993). *Statistical analysis of spherical data*. Cambridge university press.

Ganeshanandam, S. and Krzanowski, W. (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.

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

Hornik, K., Feinerer, I., Kober, M., and Buchta, C. (2012). Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

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

Kume, A., Preston, S., and Wood, A. (2013). Saddlepoint Approximations for the Normalizing Constant of Fisher–Bingham Distributions on Products of Spheres and Stiefel Manifolds. *Biometrika*, 100(4):971–984.

Kume, A. and Sei, T. (2018). On the exact maximum likelihood inference of Fisher–Bingham distributions using an adjusted holonomic gradient method. *Statistics and Computing*, 28(4):835–847.

Kume, A. and Wood, A. T. (2005). Saddlepoint approximations for the Bingham and Fisher–Bingham normalising constants. *Biometrika*, 92(2):465–476.

Laha, A. K. and Putatunda, S. (2018). Real time location prediction with taxi-gps data streams. *Transportation Research Part C: Emerging Technologies*, 92:298–322.

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

Mardia, K. and Jupp, P. (2000). *Directional Statistics*. John Wiley & Sons.

Morris, J. E. and Laycock, P. (1974). Discriminant analysis of directional data. *Biometrika*, 61(2):335–341.

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

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

Paterson, J. R., Gehling, J. G., Droser, M. L., and Bicknell, R. D. (2017). Rheotaxis in the ediacaran epibenthic organism parvancorina from south australia. *Scientific Reports*, 7:45539.

Peel, D., Whiten, W. J., and McLachlan, G. J. (2001). Fitting mixtures of kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96(453):56–63.

Rivest, L.-P. (1986). Modified Kent's statistics for testing goodness of fit for the Fisher distribution in small concentrated samples. *Statistics & probability letters*, 4(1):1–4.

Rutkowska, A., Kohnová, S., and Banasik, K. (2018). Probabilistic properties of the date of maximum river flow, an approach based on circular statistics in lowland, highland and mountainous catchment. *Acta Geophysica*, pages 1–14.

Scealy, J. and Welsh, A. (2011). Regression for Compositional Data by using Distributions defined on the Hyper-sphere. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73:351–375.

Schmidt, P. (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., Athineou, G., Sajib, A., Amson, E., and Waldstein, M. J. (2018). *Directional: directional statistics*. R package version 3.3.

Tsamardinos, I., Greasidou, E., and Borboudakis, G. (2018). Bootstrapping the Out-of-sample Predictions for Efficient and Accurate Cross-Validation. *Machine Learning*, To appear.

Vanni, L., Baldaccini, N. E., and Giunchi, D. (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. (1983). *Statistics on Spheres*. New York: Wiley.

Wood, A. (1982). A bimodal distribution on the sphere. *Applied Statistics*, pages 52–58.

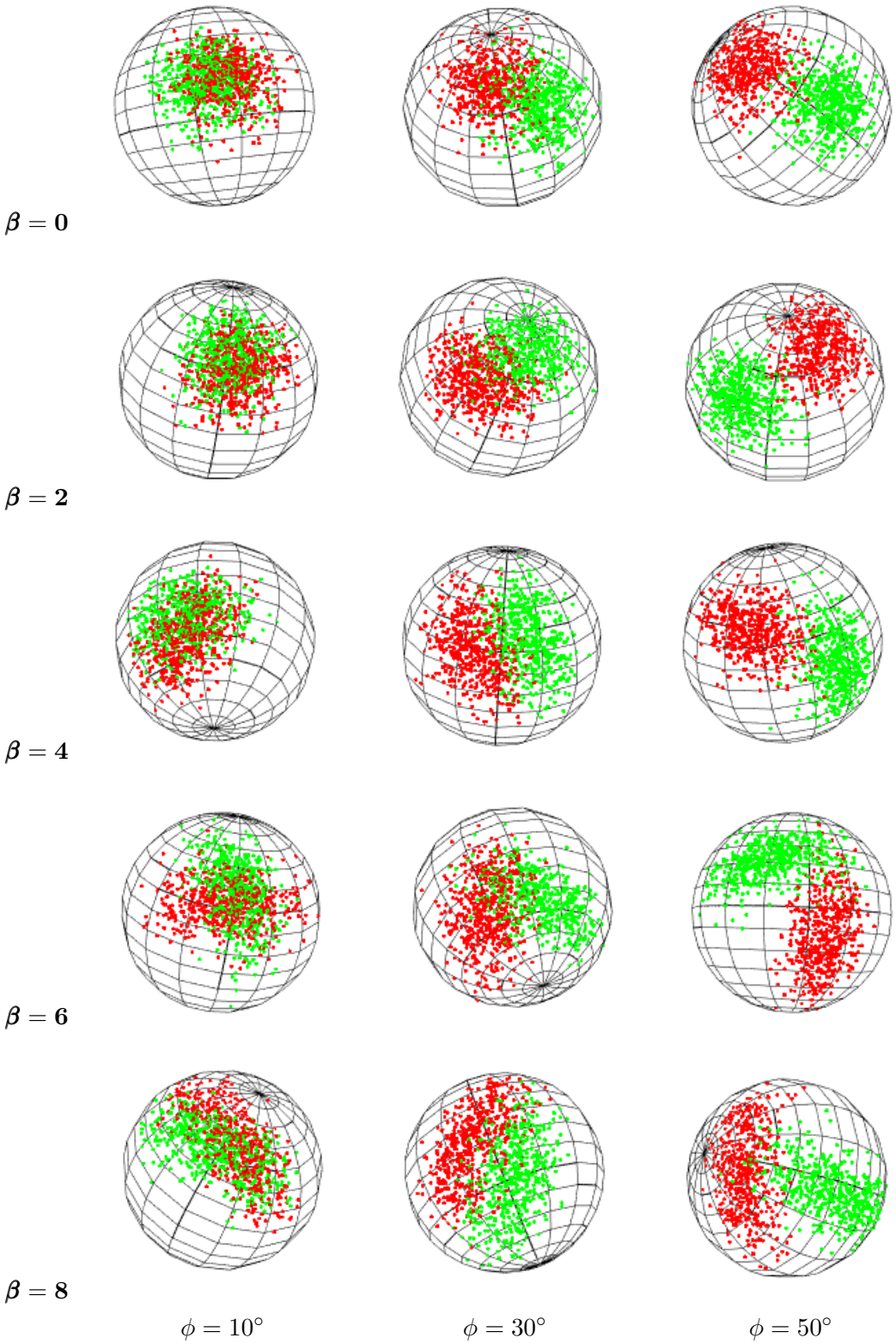


Figure 2: Spherical plots of 500 points from each population generated from Kent $(\boldsymbol{\gamma}, 20, \boldsymbol{\beta})$ distribution. The ϕ numbers denote the angle between the mean vector of each population.

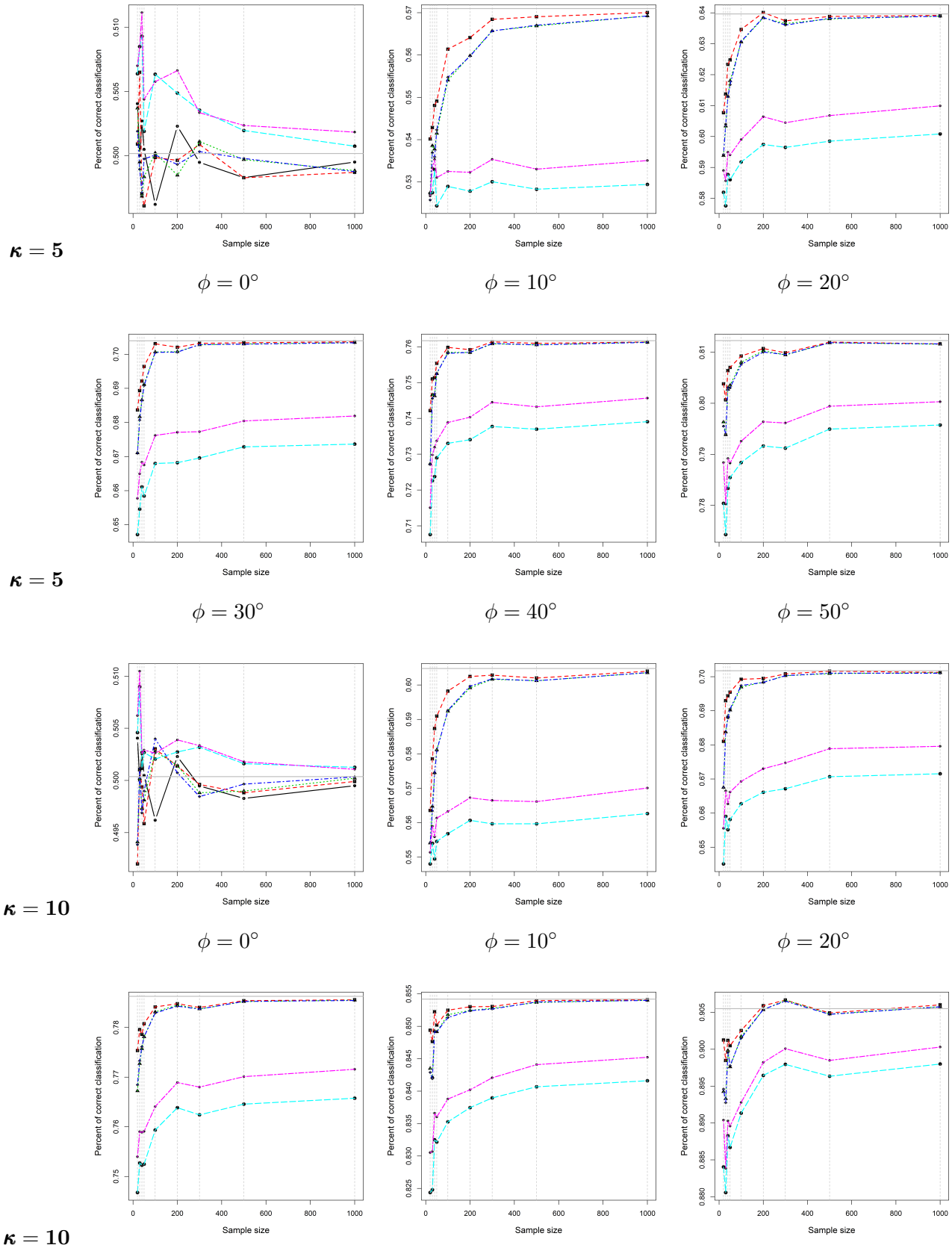
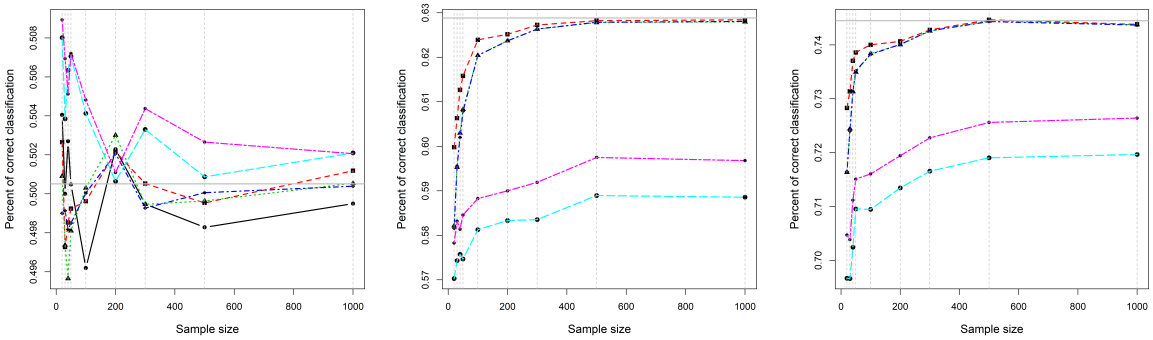
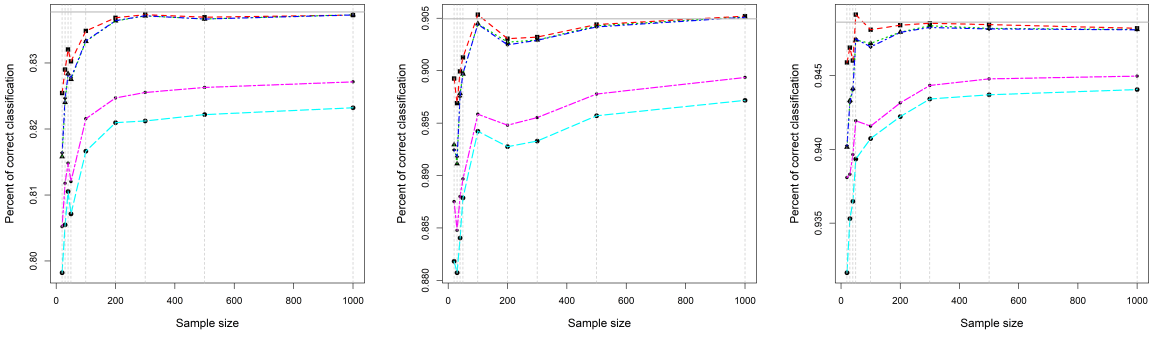


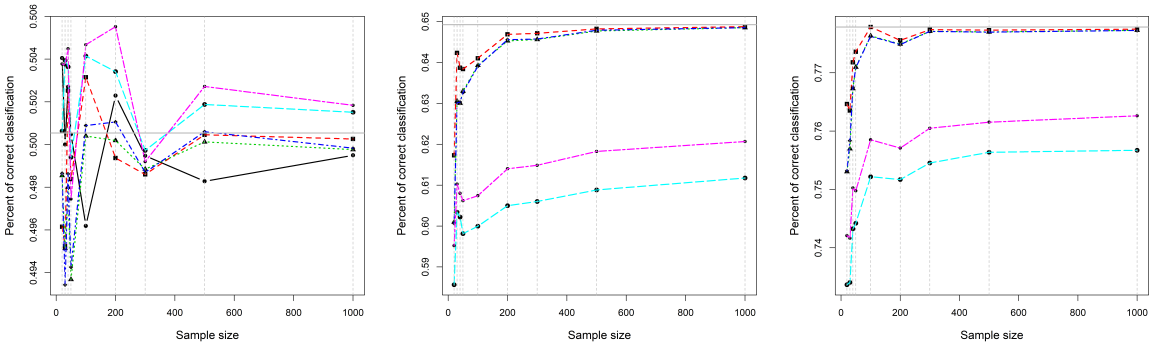
Figure 3: Data generated from a Kent $(\boldsymbol{\gamma}, \kappa, 0)$ distribution.



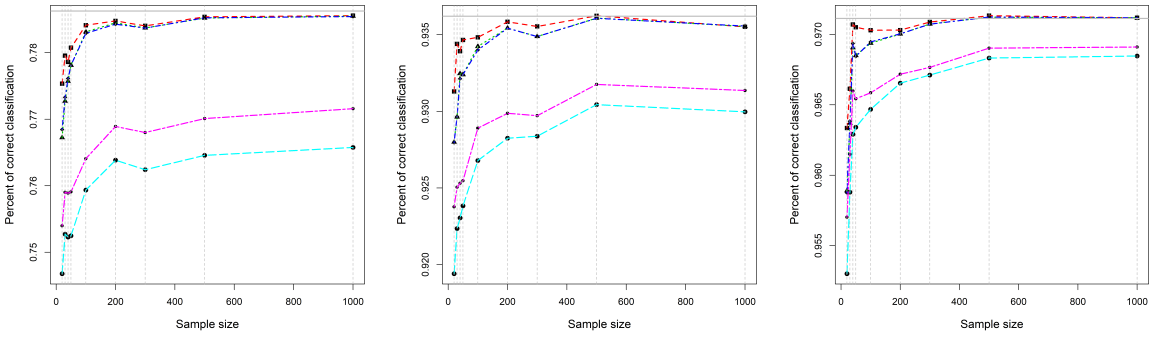
$\boldsymbol{\kappa = 15}$ $\phi = 0°$ $\phi = 10°$ $\phi = 20°$



$\boldsymbol{\kappa = 15}$ $\phi = 30°$ $\phi = 40°$ $\phi = 50°$

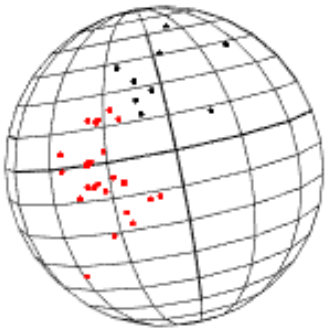


$\boldsymbol{\kappa = 20}$ $\phi = 0°$ $\phi = 10°$ $\phi = 20°$

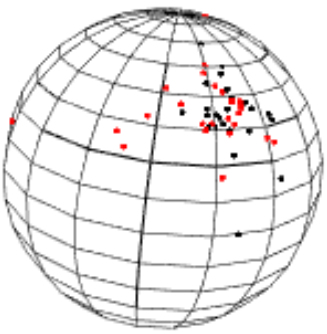


$\boldsymbol{\kappa = 20}$ $\phi = 30°$ $\phi = 40°$ $\phi = 50°$

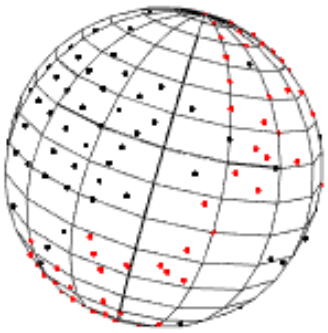
Figure 4: Data generated from a Kent $(\boldsymbol{\gamma}, \kappa, 0)$ distribution.



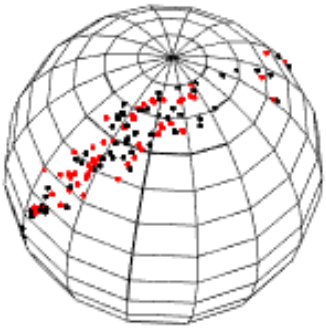
(a) Paleomagnetic data



(b) Ordovician data



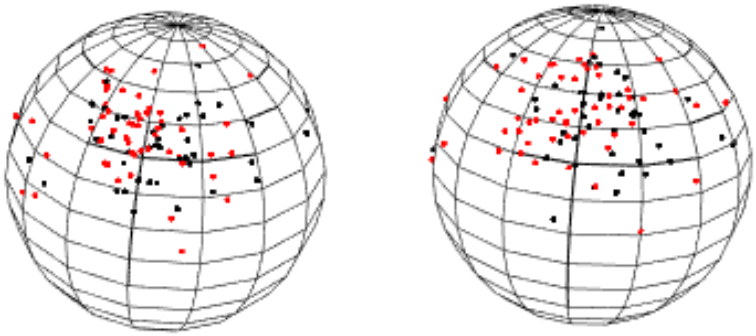
(c) Stones data



(d) Magnetic data

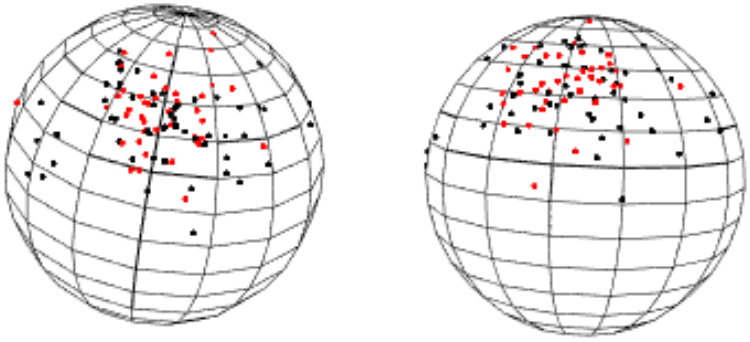
Figure 5: Spherical plots of the real datasets with different colours indicating the two groups.

Grouping according to age



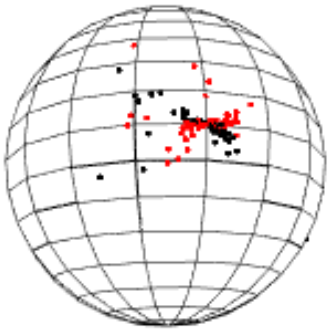
(a) Frank system (b) McFee system

Grouping according to gender

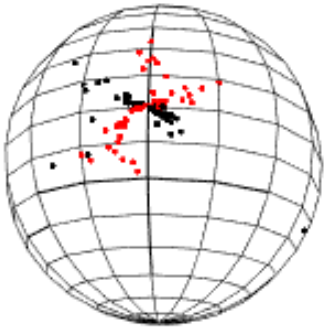


(c) Frank system (d) McFee system

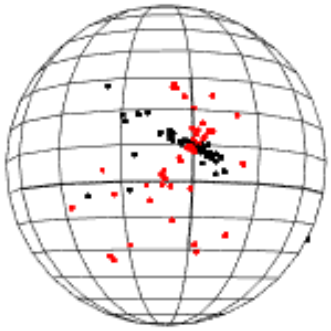
Figure 6: Spherical plots of the vectorcardiogram data with different colours indicating the two groups (age group and gender group).



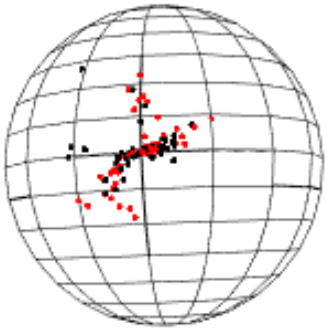
(a) Earnings vs social status



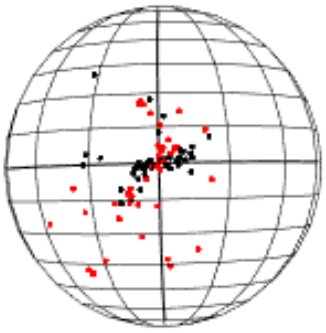
(b) Earnings vs reward



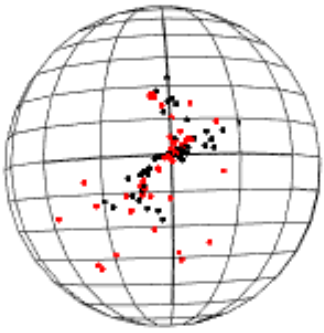
(c) Earnings vs social usefulness



(d) Social status vs reward

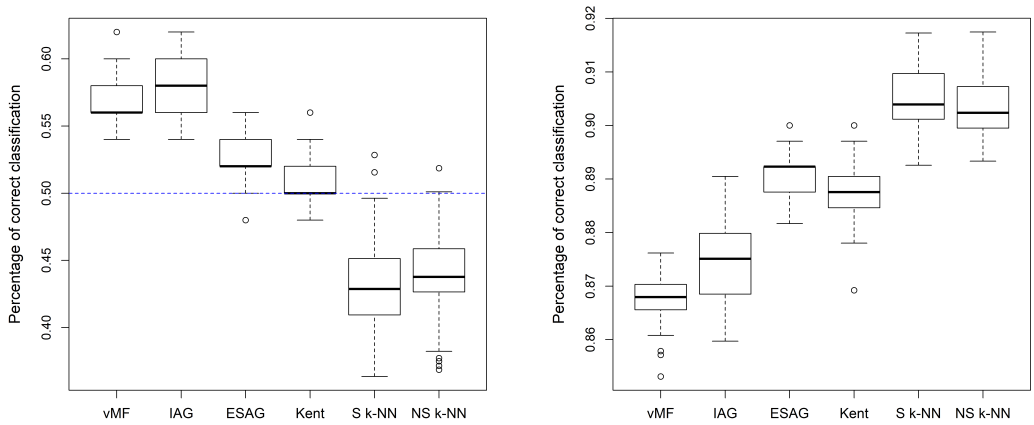


(e)Social status vs social usefulness



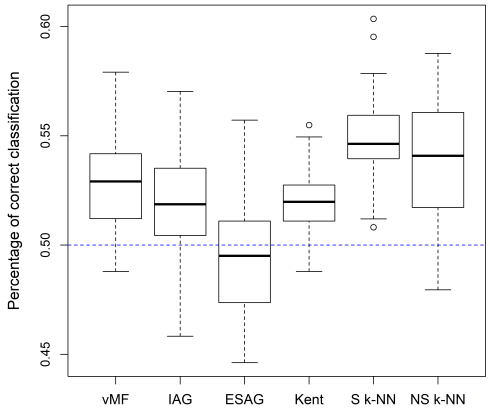
(f) Reward vs social usefulness

Figure 7: Spherical plots of the Judgments data with different colours indicating the two groups.



(a) Ordovician data

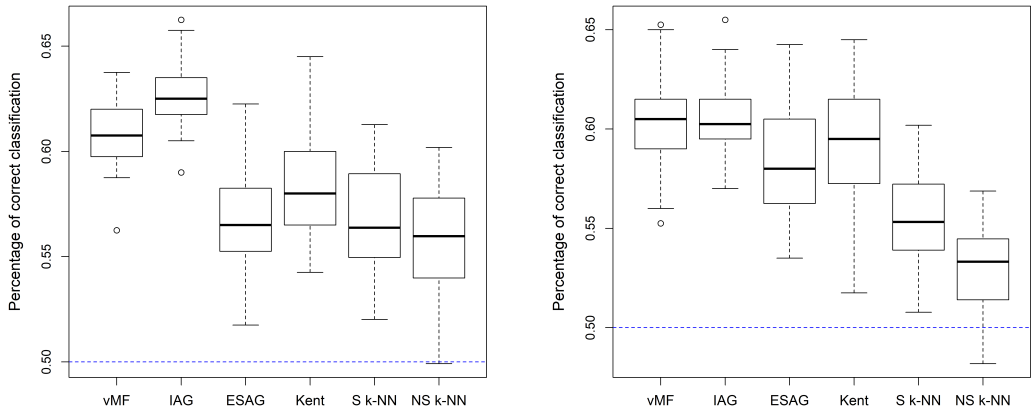
(b) Stones data



(c) Magnetic data

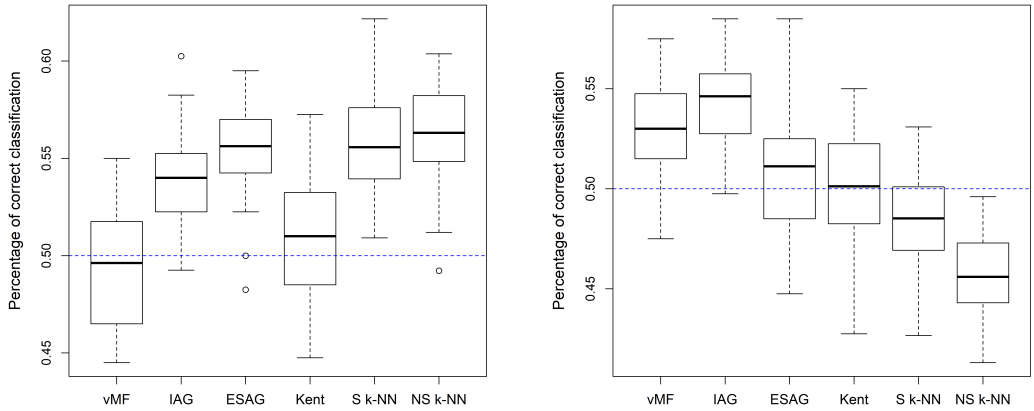
Figure 8: Box plots of the estimated predictive performance of all methods based on repeated 10-fold CV.

Grouping according to age



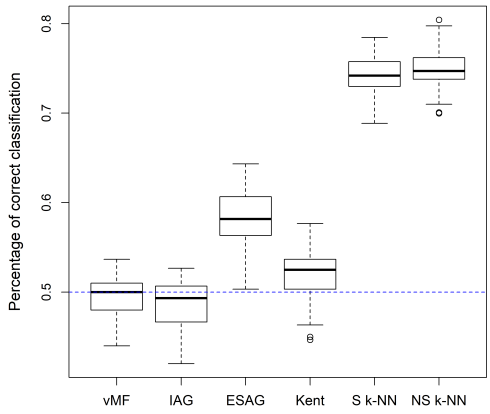
(a) Frank system (b) McFee system

Grouping according to gender

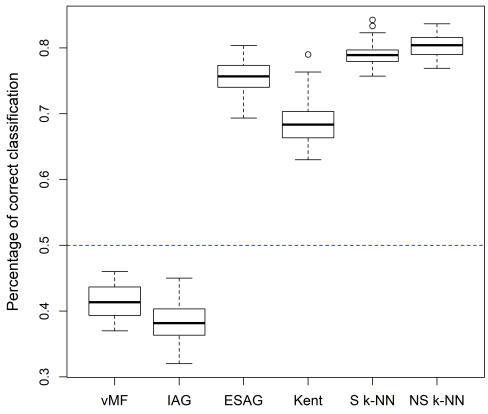


(c) Frank system (d) McFee system

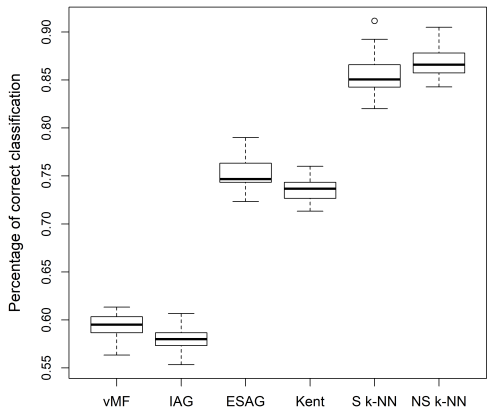
Figure 9: Box plots of the estimated predictive performance of all methods based on repeated 10-fold CV.



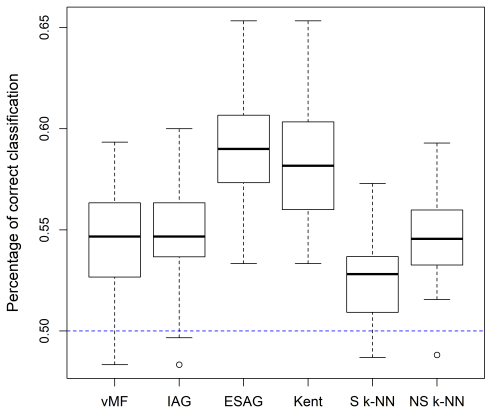
(a) Earnings vs social status



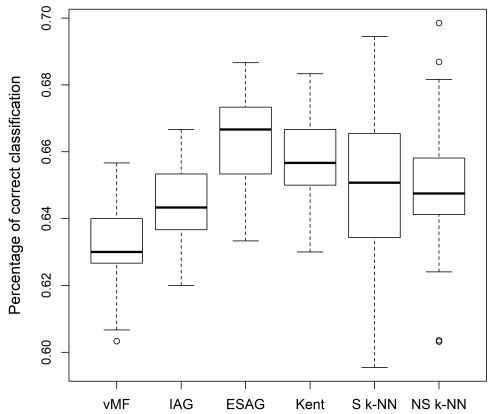
(b) Earnings vs reward



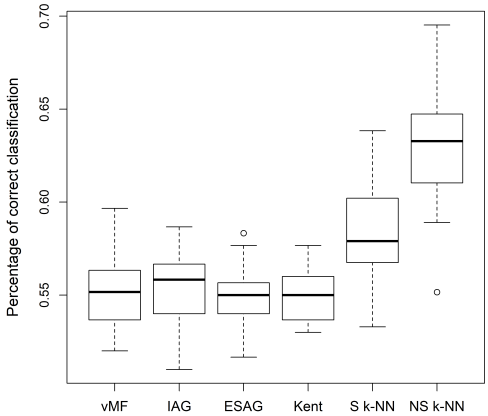
(c) Earnings vs social usefulness



(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.