Athens Journal of Sciences- Volume 1, Issue 4 – Pages 225-240

Discriminant Analysis with High Dimensional von Mises - Fisher Distributions

By Mario Romanazzi[*]

This paper extends previous work in discriminant analysis with von Mises-Fisher distributions (e. g., Morris and Laycock, Biometrika, 1974) to general dimension, allowing computation of misclassification probabilities. The main result is the probability distribution of the cosine transformation of a von Mises-Fisher distribution, that is, the random variable $U_a = a^T X$, *where* $X = (X_1, \dots, X_p)^T$, *satisfying* $X^T X = 1$, *is a random direction with von Mises-Fisher distribution and* $a = (a_1, \dots, a_p)^T$, *satisfying* $a^T a = 1$, *is a fixed non-random direction. This transformation is of general interest in multivariate analysis, in particular it underlies discriminant analysis in both two-group and multiple group problem.*

Introduction

We denote with $\mathbb{R}^p$ the p-dimensional euclidean space and with $x = (x_1, \dots, x_p)^T$ a general point (vector) of the space. The scalar product of $x, y \in \mathbb{R}^p$ is the real number $x^T y = \sum_{i=1}^p x_i y_i$ and and their euclidean norms are $\|x\| = \sqrt{x^T x}$, $\|y\| = \sqrt{y^T y}$. Let $0_p = (0, \dots, 0)^T$ denote the null vector. If $x, y \neq 0_p$, the cosine of the angle between their directions is $x^T y / (\|x\| \cdot \|y\|)$. The hyper-sphere $\boldsymbol{S}_p$ centered at 0_p with unit radius is the subset of unit-length vectors of $\mathbb{R}^p$, i. e., vectors $x \in \mathbb{R}^p$ satisfying $\|x\| = 1$.

The von Mises-Fisher (vMF) distribution plays for data in $\boldsymbol{S}_p$ the same role as the normal distribution for unconstrained euclidean data. The random vectors belonging to vMF family are indexed by two parameters, the center $\mu = (\mu_1, \dots, \mu_p)^T$, $\|\mu\| = 1$, and the concentration parameter $\kappa \geq 0$. When $\kappa = 0$, the uniform distribution on $\boldsymbol{S}_p$ is obtained. Otherwise, the distribution is unimodal, with modal direction μ, and κ measures the concentration of data around μ. We write $X \sim M_p(\mu, \kappa)$, $\mu \in \boldsymbol{S}_p$ and $\kappa \geq 0$, for a p-dimensional

[*]Professor, Ca' Foscari University, Italy.

https://doi.org/10.30958/ajs.1-4-1 doi=10.30958/ajs.1-4-1

random vector with a vMF distribution. For $x \in \boldsymbol{S}_p$, the probability density function is (Mardia et al., 1979)

$$f_M(x) = c_p(\kappa) \exp(\kappa \mu^T x)\, d\boldsymbol{S}_p, \tag{1}$$

where $d\boldsymbol{S}_p$ is the probability element of $\boldsymbol{S}_p$ and

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}. \tag{2}$$

Here $I_s(\cdot)$ denotes the modified Bessel function of first type and order s.

In practice, when dealing with hyper-spherical data $x \in \boldsymbol{S}_p$, the spherical polar coordinates $\theta = (\theta_1, \dots, \theta_{p-1})^T$ of x are often considered. The corresponding transformation is

$$\begin{aligned} &x = \psi(\theta) = (\psi_1(\theta), \dots, \psi_p(\theta))^T, \\ &0 \le \theta_i \le \pi, i = 1, \dots, p-2, 0 \le \theta_{p-1} < 2\pi, \end{aligned} \tag{3}$$

where, for $i = 1, \dots, p$,

$$\begin{aligned} &\psi_i(\theta) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j, \\ &\sin\theta_0 = \cos\theta_p = 1. \end{aligned} \tag{4}$$

Writing $\alpha = (\alpha_1, \dots, \alpha_{p-1})^T$ for the spherical polar coordinates of μ, and using (3) and (4), if $X \sim M_p(\mu, \kappa)$, writing $\Theta = (\Theta_1, \dots, \Theta_{p-1})^T$ for the corresponding angular version, the density function of Θ is

$$g_M(\theta) = c_p(\kappa) \exp(\kappa \psi(\alpha)^T \psi(\theta)) a_p(\theta), \tag{5}$$

with $a_p(\theta)$, the Jacobian of the transformation, given by

$$a_p(\theta) = \prod_{j=2}^{p-1} \sin^{p-j}\theta_{j-1} \tag{6}$$

For fixed values of the parameters, vMF density is an increasing function of the cosine of the angle between $x \in \boldsymbol{S}_p$ and μ. When $\kappa > 0$, this proves the distribution to have a unique mode at $x = \mu$ and an anti-mode at the antipodal point $x = -\mu$.

It is well-known that Mahalanobis' distance underlies normal-based discriminant analysis. For example, maximum likelihood discriminant rule amounts to assigning an (unclassified) data point to the nearest centroid, with respect to Mahalanobis' metric. Loosely speaking, for vMF distributions the scalar product replaces Mahalanobis' metric, that is, maximum likelihood discriminant rule amounts to assigning an (unclassified) data point to the centroid with maximum scalar product. This suggests to investigate the distribution of the random variable $\mu^T X$, or more in general, of $U_a = a^T X$,

where $\|a\| = 1$ and $X \sim M_p(\mu, \kappa)$ with $\kappa > 0$. The random variable U_a is interpreted as the cosine of the angle between the fixed direction a and the random direction X with vMF distribution. The distribution of U_a is derived in the Cosine Transformation section in a general p-dimensional setting. In the Discriminant Analysis section the implications for discrimination problems under vMF distributions are discussed. A final discussion is given in the Discussion section.

Results related to the topic of the present paper can be found in the literature but they are confined to the circular and spherical case. See, for example, Morris and Laycock (1974) and El Khattabi and Streit (1996) for treatments of discriminant analysis under vMF and several other distributions. A more recent paper devoted to clustering problems for mixtures of vMF distributions is Banerjee et al. (2005). This paper considers the general p-dimensional situation and highlights the importance of cosine similarity.

The Cosine Transformation

For a given non-random direction $a \in \boldsymbol{S}_p$, we derive the probability distribution of the random variable $U_a = a^T X$, with $X \sim M_p(\mu, \kappa)$, to be interpreted as the cosine of the angle formed by a and X. It is clear that $-1 \leq U_a \leq 1$ and the minimum and maximum values are attained when $X = -a$ and $X = a$, respectively. For a real number t, the cumulative distribution function of U_a is

$$F_U^{(a)}(t) = P(U_a \leq t) = P(X \in A_t) \quad (7)$$

for $-1 < t < 1$ and it is equal to 0 and 1 when $t \leq -1$ and $t \geq 1$, respectively. In (7) $A(t) = \{x \in \boldsymbol{S}_p : a^T X \leq t\} \subseteq \boldsymbol{S}_p$ is the inverse image on the hyper-spherical surface of the half line $(-\infty, t]$. To obtain probability (7) explicitly for general p, an invariance argument is used. For any $p \times p$ orthogonal matrix Q

$$U_a = a^T X = (Qa)^T QX = a^T Q^T QX, \quad (8)$$

i. e., the probability distribution of U_a is invariant to rotations and/or reflections simultaneously operating on a and X. This proves to be a key property to reduce the complexity of the problem. Let $\tilde{Q}$ be the orthogonal matrix satisfying

$$\tilde{Q}a = e_1 \equiv (1, 0, \ldots, 0)^T. \quad (9)$$

$\tilde{Q}$ is well-defined: its first row is vector a^T and the remaining $p - 1$ rows are any set of orthonormal vectors from the orthogonal complement of a. As vMF family is closed under orthogonal transformations

$$\tilde{X} = \tilde{Q}X \sim M_p(\tilde{Q}\mu,\ \kappa), \tag{10}$$

where $\tilde{Q}\mu = \tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_p)^T$ with $\tilde{\mu}_1 = a^T\mu$. Using this invariance property, the cumulative distribution function of U_a can be rewritten as follows

$$F_U^{(a)}(t) = P(a^T X \le t) = P\left(e_1^T \tilde{X} \le t\right) = P(\tilde{X}_1 \le t), \tag{11}$$

and therefore $F_U^{(a)}(\cdot)$ is coincident with the cumulative distribution function of $\tilde{X}_1$, the first marginal component of $\tilde{X}$, the $\tilde{Q}$-rotated vMF random vector. Here it is convenient to switch to spherical polar coordinates. Let $\tilde{\Theta}$ be the angular transformation of $\tilde{X}$. From (3) and (4),

$$\tilde{X} = \psi(\tilde{\theta}) = (\psi_1(\tilde{\theta}), \ldots, \psi_p(\tilde{\theta}))^T = (\cos\tilde{\theta}_1, \ldots, \psi_p(\tilde{\theta}))^T \tag{12}$$

implying that

$$F_U^{(a)}(t) \equiv P\left(\tilde{X}_1 \le t\right) = P\left(\cos\tilde{\Theta}_1 \le t\right). \tag{13}$$

Summing up the previous discussion, the required cumulative distribution function of U_a is the cumulative distribution function of the cosine of $\tilde{\Theta}_1$, the first marginal component of $\tilde{\Theta}$. The circular case looks particular and is dealt with in the Circular Case Section, whereas the general (hyper-)spherical case is dealt with in the Spherical and Hyper-Spherical Section.

Circular Case

In the circular case, $\tilde{\Theta}_1 \equiv \tilde{\Theta}$, with $0 \le \tilde{\Theta} < 2\pi$, and (13) is easily evaluated. The density function of $\tilde{\Theta}$ is

$$\tilde{g}_M(\theta) = (2\pi I_0(\kappa))^{-1} \exp\{\kappa\cos(\theta - \tilde{\alpha})\}, \tag{14}$$

with $\tilde{\alpha}$ satisfying

$$\tilde{\mu}_1 = a^T\mu = \psi_1(\tilde{\alpha}) = \cos\tilde{\alpha} \tag{15}$$

By (13),

$$F_U^{(a)}(t) = P\left(\cos\tilde{\Theta} \le t\right) = \int_{\arccos t}^{2\pi - \arccos t} \tilde{g}_M(\theta)d\theta, \tag{16}$$

where $0 \le \arccos t \le \pi$. Put

$$\kappa_1(a) = \kappa\cos\tilde{\alpha} = \kappa a^T\mu,\ \kappa_2(a) = \kappa\sin\tilde{\alpha} = \kappa(1 - (a^T\mu)^2)^{1/2}, \tag{17}$$

The density function is the derivative of $F_U^{(a)}(t)$ with respect to t:

$$f_U^{(a)}(t) = \left(2\pi I_0(\kappa)\right)^{-1}(1-t^2)^{-1/2}\exp(\kappa_1(a)t) \times \left\{\exp\left(-\kappa_2(a)(1-t^2)^{1/2}\right) + \exp\left(\kappa_2(a)(1-t^2)^{1/2}\right)\right\} \tag{18}$$

The previous expression holds for all $-1 \le a^T\mu \le 1$. When $a^T\mu = \mp 1$, i. e., $a = \mp\mu$, then $\sin\tilde{\alpha} = 0$ and a simpler formula is obtained

$$f_U^{(a)}(t) = \left(\pi I_0(\kappa)\right)^{-1}(1-t^2)^{-1/2}\exp(\kappa_1(a)t). \tag{19}$$

This expression is coincident with equation (6) in Morris and Laycock (1974), giving the density function of $\mu^T X$.

Spherical and Hyper-Spherical Case

The derivation of the distribution of U_a when $p \geq 3$ relies on a preliminary lemma, possibly of independent interest.

Lemma 1. For $p \geq 3$, let $X = (X_1, \ldots, X_p)^T$ be distributed as $M_p(\mu, \kappa)$, with $\kappa > 0$ and let $\Theta = (\Theta_1, \ldots, \Theta_{p-1})^T$ and $\alpha = (\alpha_1, \ldots, \alpha_{p-1})^T$ be the spherical polar coordinates of X and μ, respectively. Partition Θ, μ and α as

$$\Theta = \left(\Theta_1, \hat{\Theta}_2^T\right)^T, \mu = (\mu_1, \hat{\mu}_2^T)^T, \alpha = (\alpha_1, \hat{\alpha}_2^T)^T \quad (20)$$

where $\hat{\theta}_2$ is the subvector of Θ obtained by dropping the first component and the same holds for $\hat{\mu}_2$ and $\hat{\alpha}_2$. Assume $0 < \alpha_1 < \pi$. Then

i. the conditional distribution $\hat{\Theta}_2 | \Theta_1 = \theta_1^{(0)}$ is the angular transformation of a vMF distribution $M_{p-1}(\mu_0, \kappa_0)$, , where

$$\mu_0 = \hat{\mu}_2 / \sin \alpha_1, \kappa_0 = \kappa \sin \alpha_1 \sin \theta_1^{(0)}, \quad (21)$$

ii. for $0 \leq \theta_1 \leq \pi$, the marginal density function $g_1(\cdot)$ of Θ_1 is

$$g_1(\theta_1) = \left(\frac{\kappa}{2\pi}\right)^{1/2} \frac{\sin^{-(p-3)/2} \alpha_1}{I_{p/2-1}(\kappa)} I_{(p-1)/2-1}(\kappa \sin \alpha_1 (1 - \cos^2 \theta_1)^{1/2}) \quad (22)$$
$$\times (1 - \cos^2 \theta_1)^{(p-1)/4} \exp(\kappa \cos \alpha_1 \cos \theta_1).$$

Proof. The proof is obtained from the factorization of the joint density $g_M(\cdot)$ as the product of the marginal density of Θ_1 and the conditional density of $\hat{\Theta}_2 | \Theta_1 = \theta_1^{(0)}$.

The following corollary deals with the special case $\alpha_1 \in \{0, \pi\}$.

Corollary 1. If $\alpha_1 \in \{0, \pi\}$, then $\psi_1(\alpha) = \pm 1$ and $\hat{\psi}_2(\alpha) = \sin \alpha_1 \psi(\hat{\alpha}_2) = 0_{p-2}$. This implies that Θ_1 and $\hat{\Theta}_2$ are independently distributed and $\hat{\Theta}_2$ has a uniform distribution on S_{p-1}. The marginal density of Θ_1 turns out to be

$$g_1(\theta_1) = \left(\frac{\kappa^{p-2}}{2^{p-2}\pi}\right)^{1/2} \frac{1}{I_{p/2-1}(\kappa)\Gamma((p-1)/2)} \sin^{p-2} \theta_1 \quad (23)$$
$$\times \exp(\kappa \cos \alpha_1 \cos \theta_1)$$

with $\cos\alpha_1 = \pm 1$ according to whether $\alpha_1 = 0$ or $\alpha_1 = \pi$, respectively (compare with Mardia et al. 1979, p. 431, equation 15.3.18).

Lemma 1 and Corollary 1 provide an explicit expression of the density function of $\tilde{\Theta}_1$ and this allows (13) to be easily evaluated. For $-1 < t < 1$,

$$F_U^{(a)}(t) = P(\tilde{X}_1 \le t) = P(\cos\tilde{\Theta}_1 \le t) = 1 - P(\tilde{\Theta}_1 \le \arccos t)$$
$$= 1 - \int_0^{\arccos t} g_1(\theta_1)d\theta_1 \qquad (24)$$

The density function follows by differentiation of (24) with respect to t. Again put $\kappa_1(a) = \kappa\cos\tilde{\alpha}_1 = \kappa a^T\mu$, $\kappa_2(a) = \kappa\sin\tilde{\alpha}_1 = \kappa(1-(a^T\mu)^2)^{1/2}$. Assuming $-1 < \cos\tilde{\alpha}_1 = a^T\mu < 1$, i. e., $a \neq \mp\mu$, expression (22) is considered and the density function of U_a turns out to be

$$f_U^{(a)}(t) = \left(\frac{\kappa}{2\pi}\right)^{1/2}\frac{I_{(p-1)/2-1}(\kappa_2(a)(1-t^2)^{1/2})}{I_{p/2-1}(\kappa)}\sin^{-(p-3)/2}\tilde{\alpha}_1$$
$$\times(1-t^2)^{(p-3)/4}\exp(\kappa_1(a)t) \qquad (25)$$

If $\cos\tilde{\alpha}_1 = a^T\mu = \mp 1$, i. e., $a = \mp\mu$, expression (23) is considered and in this special case the density function of U_a turns out to be

$$f_U^{(a)}(t) = \left(\frac{\kappa^{p-2}}{2^{p-2}\pi}\right)^{1/2}\frac{1}{I_{p/2-1}(\kappa)\Gamma((p-1)/2)}(1-t^2)^{(p-3)/2}\exp(\kappa_1(a)t) \qquad (26)$$

When $p = 2$, (25) and (26) are coincident with (18) and (19), respectively. Therefore, the circular case is essentially included in the general case.

The following remark deals with the relation between normal and vMF distributions.

Remark 1. We write $X \sim N_p(\mu, \Sigma)$ for a p-dimensional random vector with a normal distribution with mean vector μ and covariance matrix Σ. Mardia (1975) proved the following characterization of vMF distribution. If $X \sim N_p(\mu, \kappa^{-1}\mathrm{I}_p)$, with $\mu^T\mu = 1$, $\kappa > 0$ and I_p denoting the identity matrix of order p, then the conditional distribution of X, given $X^TX = 1$, is $M_p(\mu, \kappa)$. It follows that, under the same hypotheses, the density of a^TX, given $X^TX = 1$, is still given by (25) and (26).

A geometrical interpretation of the set A_t in (7) is given in the remark below.

Remark 2. Suppose $p = 3$. For $x \in \mathbb{R}^3$ and $t \in \mathbb{R}$, the equation $a^T x = t$ defines the family of parallel hyperplanes orthogonal to vector a. If $-1 < t < 1$, they intersect the sphere $\boldsymbol{S}_3$ in a set of parallel circles and if $t = \pm 1$ they are tangent to $\boldsymbol{S}_3$ at $\pm a$, the intersection points of the line λa, $\lambda \in \mathbb{R}$, with $\boldsymbol{S}_3$. It follows that $A_t = \{x \in \boldsymbol{S}_3 : a^T x \leq t\}$ is the spherical cap including $-a$ whose boundary is the circle $\{x \in \boldsymbol{S}_3 : a^T x = t\}$. Rotating the sphere according to $\tilde{Q}$ (see (13)) amounts to rotate the coordinate system axes so as a is coincident with $e_1 = (1,0,0)^T$, the direction from the origin 0_3 to the North pole. With the obvious adaptations, this picture holds in the circular case and carries over the general p-dimensional case.

The Family of $\boldsymbol{U}_a$ Random Variables

The family of random variables $\{U_a, a \in \boldsymbol{S}_p\}$ has a simple structure that can be summarized by saying that they are ordered according to the scalar parameter $a^T \mu$, where μ is the center of the parent vMF distribution. The ordering is the standard stochastic ordering $\leq_{ST}$ of random variables

$$U \leq_{ST} V \text{ iff } P(U > t) \leq P(V > t), t \in \mathbb{R}. \tag{27}$$

For $X \in M_p(\mu, \kappa)$, $\mu^T \mu = 1$ and $\kappa > 0$, the expectation vector and the covariance matrix are (Mardia et al., 1979, Watamori,1995)

$$E(X) = A_p(\kappa)\mu \tag{28}$$

$$V(X) = A_p'(\kappa)\mu\mu^T + \kappa^{-1}A_p(\kappa)\left(\mathrm{I}_p - \mu\mu^T\right) = (1 - p\kappa^{-1}A_p(\kappa) - A_p^2(\kappa))\mu\mu^T + \kappa^{-1}A_p(\kappa)\mathrm{I}_p \tag{29}$$

where

$$A_p(\kappa) = I_{p/2}(\kappa)/I_{p/2-1}(\kappa) \tag{30}$$

$$A_p'(\kappa) = \frac{\partial}{\partial \kappa}A_p(\kappa) = 1 - (p-1)\kappa^{-1}A_p(\kappa) - A_p^2(\kappa). \tag{31}$$

It is known that $A_p(\kappa)$ is non-decreasing and $0 < A_p(\kappa) < 1$ because $I_s(t)$ is a decreasing function of the order s for a given value of the argument t. From (28) and (29), for $a, b \in \boldsymbol{S}_p$, expectation and variance of U_a and covariance of U_a, U_b turn out to be

$$E(U_a) = A_p(\kappa)a^T\mu, \tag{32}$$

$$Var(U_a) = \left(1 - p\kappa^{-1}A_p(\kappa) - A_p^2(\kappa)\right)(a^T\mu)^2 + \kappa^{-1}A_p(\kappa), \tag{33}$$

$$Cov(U_a, U_b) = \left(1 - p\kappa^{-1}A_p(\kappa) - A_p^2(\kappa)\right)a^T\mu\mu^T b + \kappa^{-1}A_p(\kappa)a^T b. \tag{34}$$

For fixed values of μ and κ, $E(U_a)$ and $Var(U_a)$ only depend on $a \in S_p$. It is easily checked that $-A_p(\kappa) \leq E(U_a) \leq A_p(\kappa)$ and the minimum and

maximum values are attained when $a = -\mu$ and $a = \mu$, respectively. Moreover, $A'_p(\kappa) \leq Var(U_a) \leq \kappa^{-1} A_p(\kappa)$ and the minimum and maximum values are reached when $a = \pm\mu$ and when $a^T\mu = 0$, i. e., a is orthogonal to μ, respectively.
We now prove the ordering property stated above.

Proposition 1. If $X \sim M_p(\mu, \kappa)$, $\kappa > 0$, and $a, b, \mu \in \boldsymbol{S}_p$, then $a^T\mu < b^T\mu$ implies $U_a \leq_{ST} U_b$.
Proof. A geometrical argument will be used. From (7)

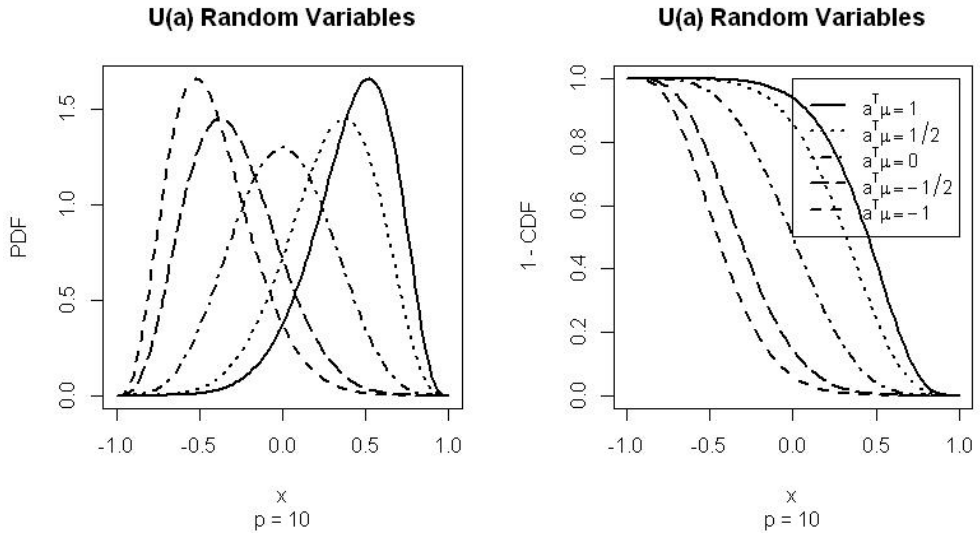
$$P(U_a > t) = P(X \in A_t^C) \tag{35}$$

where, according to Remark 2, A_t^C is the hyper-spherical cap of $\boldsymbol{S}_p$ including a whose boundary is the $(p-1)$-dimensional sphere $\{x \in \boldsymbol{S}_p : a^T x = t\}$. As the random vector X is unimodal and symmetric about μ, the probability of A_t^C reaches the maximum value when $a = \mu$ because in this case A_t^C is a symmetric neighbourhood of the mode μ. Conversely, the probability reaches the minimum value when $a = -\mu$ because then A_t^C is a symmetric neighbourhood of the anti-mode $-\mu$. In general, the probability of A_t^C is an increasing function of the cosine $a^T\mu$.

The following example provides an illustration of the properties of U_a random variables.

Example 1. We assume $X \sim M_p(\mu, \kappa)$, with $p = 10$, $\mu = e_1$ and $\kappa = 5$. Moreover, $a = (c, a_2, \ldots, a_{10})^T$ so as $a^T\mu = c$ and $c \in \{-1, -1/2, 0, 1/2, 1\}$. Figure 1 shows the density and survival functions, respectively, of U_a. The behaviour of survival functions confirms the stochastic ordering of Proposition 1. It is also clear that the value of $a^T\mu$ determines not only location, but also spread and shape. The distribution is generally asymmetric, but when $a^T\mu$ it is symmetric.

Figure 1. *Probability Density (Left) and Survival Functions (Right) of* U_a *Random Variables for* $p = 10$



Discriminant Analysis

In this section the usual model of discriminant analysis is considered. The distribution of the random vector $X = (X_1, \ldots, X_p)^T$ depends on a partition of the sample space in $G \geq 2$ classes C_g, $1 \leq g \leq G$. The prior probabilities of the classes are $P(C_g) \equiv \omega_g \geq 0$, $1 \leq g \leq G$, $\sum_{g=1}^{G} \omega_g = 1$. These classes often correspond to the distinct modalities of an observable stratification variable, e. g., gender or education level of the units of a human population. The conditional random vector X, given C_g, is denoted with $X|C_g \equiv X^{(g)}$ and is assumed to belong to vMF family for all $1 \leq g \leq G$, with class-dependent parameters, that is

$$X^{(g)} \sim M_p(\mu^{(g)}, \kappa_g) \tag{36}$$

Let $x \in \boldsymbol{S}_p$ be the data vector observed on a unit to be allocated to one of the classes. Bayes' allocation rule (BR) states that it must be assigned to the class with maximum posterior probability, according to Bayes' theorem. It amounts to compute

$$P(C_g|x) = \frac{\omega_g f_M^{(g)}(x)}{\sum_{l=1}^{G} \omega_l f_M^{(l)}(x)} \tag{37}$$

and to determine the optimum class C_{g_0} satisfying

$$g_0 = \text{argmax}_{1\le g\le G} P(C_g|x) = \text{argmax}_{1\le g\le G} \ln\left(\omega_g f_M^{(g)}(x)\right). \quad (38)$$

Here, $f_M^{(g)}(x)$ is the vMF density function value of x, given C_g class, $1 \le g \le G$. The natural logarithm in (38) turns out to be

$$\lambda_{BR}(g;x) = \ln \omega_g + \ln c_p(\kappa_g) + \kappa_g \mu^{(g)T} x, \quad (39)$$

i. e., an affine transformation of a U_a-type random variable, with $a \equiv \mu^{(g)}$, the class center. This makes clear the connection between the results in the Cosine Transformation Section and discriminant analysis under vMF distributions.

A remarkable property of Bayes' rule is that it maximizes the posterior probability of correct allocation, with respect to the prior $\omega_1, \dots, \omega_G$. Of course it assumes that the prior distribution exists and is known. An alternative is maximum likelihood allocation rule (LR) that assigns a unit with data vector x to the class providing maximum likelihood to x. In the present context, it amounts to maximize the function

$$\lambda_{LR}(g;x) = \ln c_p(\kappa_g) + \kappa_g \mu^{(g)T} x, \quad (40)$$

with respect to $1 \le g \le G$. Bayes' and maximum likelihood rule agree for a uniform prior distribution. In the rest of the paper, the latter rule is always considered.

Two-Class Discrimination

When $G = 2$, the problem considerably simplifies. In terms of maximum likelihood rule, a unit with data point x is assigned to class C_1, say, iff $\lambda_{LR}(1;x) \ge \lambda_{LR}(2;x)$, that is,

$$\bar{\mu}^T x \ge \gamma, \quad (41)$$

where

$$\bar{\mu} = \frac{\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}}{\|\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}\|}, \quad (42)$$

and

$$\gamma = \frac{\ln(c_p(\kappa_2)/c_p(\kappa_1))}{\|\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}\|}. \quad (43)$$

Once again, the discriminant variable $D_{12} = \bar{\mu}^T X$ is a U_a-type random variable, whose class-conditional distributions are known. This allows an easy evaluation of several summaries, including misclassification probabilities.

The probability of erroneously allocating to C_1 a unit belonging to C_2 is

$$P_{1|2} = P(\bar{\mu}^T X^{(2)} \ge \gamma) \quad (44)$$

and the probability of erroneously allocating to C_2 a unit belonging to C_1 is

$$P_{2|1} = P(\bar{\mu}^T X^{(1)} < \gamma) \quad (45)$$

Moreover, the expectation of the difference of the two discriminant variables $\bar{\mu}^T X^{(1)}$, $\bar{\mu}^T X^{(2)}$ is

$$\Delta_{12} = E\left(\bar{\mu}^T X^{(1)} - \bar{\mu}^T X^{(2)}\right) = \bar{\mu}^T\left(A_p(\kappa_1)\mu^{(1)} - A_p(\kappa_2)\mu^{(2)}\right). \quad (46)$$

Remark 3. Once again, there is a simple geometrical interpretation of (41). Suppose, for simplicity, $p = 3$. The plane $\bar{\mu}^T x = \gamma$, $-1 \leq \gamma \leq 1$, intersects the sphere $\boldsymbol{S}_3$ in a circle. Therefore, the domain of allocation to C_1, i. e., the set $\{x \in \boldsymbol{S}_3 : \bar{\mu}^T x \geq \gamma\}$ is the spherical cap cut by such a plane which includes $\bar{\mu}$ and the domain of allocation to C_2 is the complementary spherical cap, which includes $-\bar{\mu}$. The boundary circle is the set of equal density of the two classes. The normal direction of the cutting plane, the $\bar{\mu}$-vector, depends on the location and concentration parameters of the two classes. Bayes' rule is only a minor modification because the prior probabilities affect the constant term γ, only, and make the cutting plane shift nearer to $\mu^{(1)}$ or $\mu^{(2)}$ according to the relative size of ω_1 and ω_2. This holds for general dimension p.

Two particular cases are of interest. First, suppose $\kappa_1 = \kappa_2$ and $\mu^{(1)} \neq \mu^{(2)}$. Then

$$\bar{\mu} = \frac{\mu^{(1)} - \mu^{(2)}}{\|\mu^{(1)} - \mu^{(2)}\|}, \gamma = 0. \quad (47)$$

It can be shown that in such a case $\bar{\mu}$ is orthogonal to the bisector direction $(\mu^{(1)} + \mu^{(2)})/\|\mu^{(1)} + \mu^{(2)}\|$ of the angle formed by $\mu^{(1)}$ and $\mu^{(2)}$. Moreover, $\gamma = 0$ implies that the spherical caps corresponding to the allocation domains of the two classes are hemispheres. Therefore, the misclassification probabilities of maximum likelihood rule are equal. This could no longer hold with Bayes' rule because then $\gamma = \ln(\omega_2/\omega_1)/\|\mu^{(1)} - \mu^{(2)}\|$ which is not equal to zero if $\omega_1 \neq \omega_2$.

Second, suppose $\mu^{(1)} = \mu^{(2)} = \mu$ with $\mu^T\mu$=1 and, without loss of generality, $\kappa_1 > \kappa_2 \geq 0$. Then

$$\bar{\mu} = \mu, \gamma = \frac{\ln(c_p(\kappa_2)/c_p(\kappa_1))}{|\kappa_1 - \kappa_2|} \quad (48)$$

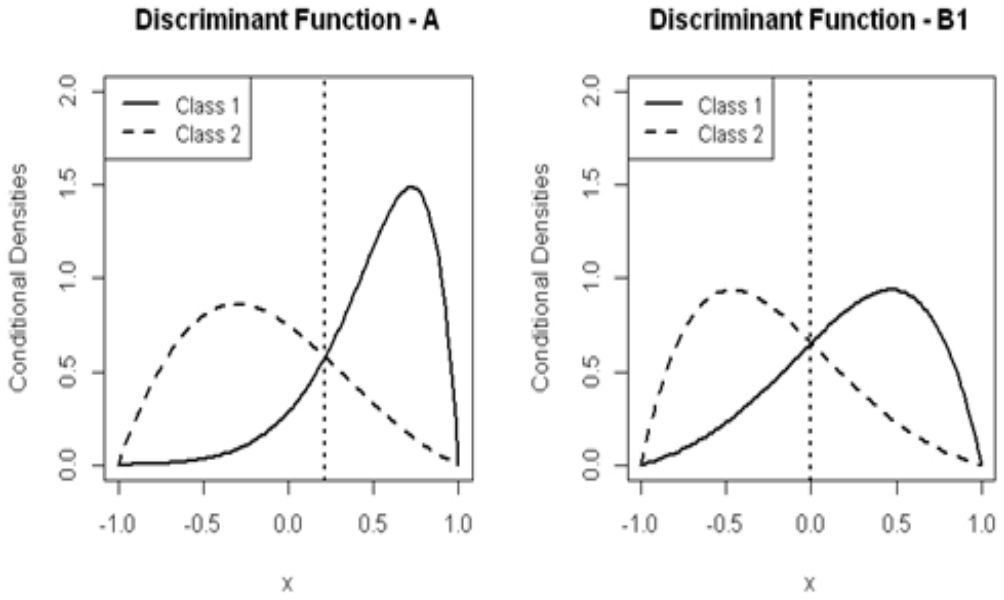
Table 1. *Conditional Allocation Probabilities* $P_{i|j}$, $i,j \in \{1,2\}$, *under vMF Distributions of Example 2*

	Case A – True Class		Case B1 – True Class	
Predicted Class	C_1	C_2	C_1	C_2
C_1	0.843	0.201	0.738	0.262
C_2	0.157	0.799	0.262	0.738
	Case B2 – True Class		Case C – True Class	
Predicted Class	C_1	C_2	C_1	C_2
C_1	0.876	0.124	0.702	0.442
C_2	0.124	0.876	0.298	0.558

Therefore, the allocation domain of C_1 is the spherical cap cut by the plane with normal direction μ which includes μ and the allocation domain of C_2 is the complementary spherical cap, not including μ. This appears coherent with model features because, in this case, the conditional random vectors $X^{(1)}$, $X^{(2)}$ have the same center μ and $X^{(1)}$ is more concentrated around it. As in the general case, the boundary of the allocation domains is the equal density set of the two classes.

Example 2. We consider two vMF random vectors, $X^{(1)} \sim M_5(\mu^{(1)}, \kappa_1)$, $X^{(2)} \sim M_5(\mu^{(2)}, \kappa_2)$ and four different cases, labelled A, B1, B2, and C. In case A, $\mu^{(1)} = (1,1,0,0,0)^T/\sqrt{2}$, $\mu^{(2)} = (1,-1,0,0,0)^T/\sqrt{2}$, $\kappa_1 = 4$, $\kappa_2 = 2$. In case B1, centers are as in case A and $\kappa_1 = \kappa_2 = 2$. Case B2 is the same as B1 except that $\kappa_1 = \kappa_2 = 4$. Finally, in case C $\mu^{(1)} = \mu^{(2)} = (1,1,0,0,0)^T/\sqrt{2}$ and concentration parameters are as in case A. Figures 2 and 3 show the density functions of the discriminant functions $\vec{\mu}^T X^{(1)}$ and $\vec{\mu}^T X^{(2)}$ together with the γ thresholds for allocation to C_1, according to maximum likelihood rule. In case C, (26) is used because $\vec{\mu}^T \mu^{(1)} = \vec{\mu}^T \mu^{(2)} = 1$. The pictures confirm the thresholds to coincide with the crossing points of the conditional densities. The allocation probabilities are reported in Table 1. As suggested by the density plots, case C gives the worst results. According to the underlying geometry, when the concentration parameters are equal, as in cases B1 and B2, the allocation probabilities for B1 and B2 are exactly the same.

Figure 2. *Density Curves of Random Variables* $\bar{\mu}^T X^{(1)}$ *and* $\bar{\mu}^T X^{(2)}$ *and* γ *Threshold (Vertical Line) in Cases A (Left) and B1 (Right) of Example 2*



Discussion

Bayes and maximum likelihood discriminant rules for vMF distributions depend on the scalar products $a^T X^{(g)}$, $g = 1, \ldots, G$, where the coefficient vector a is a function of distribution parameters. This implies that the allocation regions are spherical caps when $G = 2$, or intersections of them, when $G > 2$. However, if $\boldsymbol{S}_p$ is seen as a subset of $\mathbb{R}^p$, the allocation regions are halfspaces when $G = 2$, or intersections of them, when $G > 2$. The boundary hyperplanes depend on distribution parameters, centers and concentration parameters, according to a well-understood mechanism. That the discrimination rules, in $\mathbb{R}^p$, produce a linear separation of centers is not surprising (recall Remark 1) because vMF distributions are restrictions on $\boldsymbol{S}_p$ of p-variate normal distributions with a scalar covariance matrix.

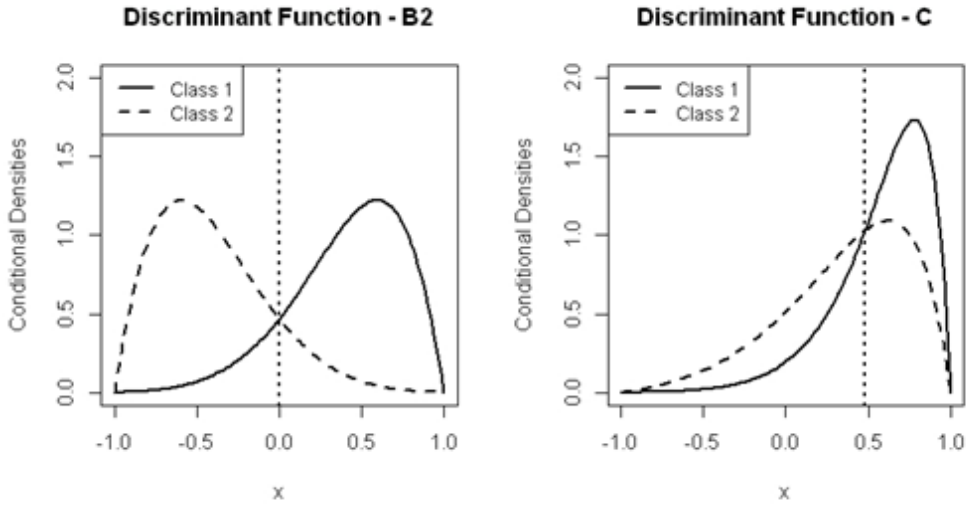
The central role of the scalar product in parametric discriminant analysis can be of help in devising sensible distribution-free rules. An example is k-th nearest neighbour classification rule with (suitable rescaled) scalar product replacing euclidean distance as the measure of neighbourhood width.

With sample data, unknown parameters are replaced by suitable estimates, e. g., maximum likelihood (ML) estimates. Recently, there were several contributions aimed to improve the performance of the ML estimator of the concentration parameter, e. g., Song et al. (2012), Sra (2012) and references

therein. An R package implementing the improved ML estimation is described by Hornik and Grun (2013).

The main application of the theory developed in this work is discriminant analysis of data belonging to S_p under vMF distribution, for all p. As shown by Banerjee et al. (2005), two important fields of application are text categorization and supervised classification of gene expression data. In both cases dimension p is typically greater than 3, a situation that requires the general results derived in the Cosine Transformation section.

Figure 3. *Density Curves of Random Variables* $\bar{\mu}^T X^{(1)}$, $\bar{\mu}^T X^{(2)}$ *and* γ *Threshold (Vertical Line) in Cases B2 (Left) and C (Right) of Example 2*



References

Banerjee, A., Dhillon, I. S. Gosh, J., and Sra, S. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. Journal of Machine Learning Research 6 1345-1382.

El Khattabi, S., and Streit, F. 1996. Identification analysis in directional statistics. Computational Statistics & Data Analysis 23 45-63.

Hornik, K., Grun, B. 2013. movMF: An R package for fitting mixtures of von Mises-Fisher distributions. R package version 0.1-2. Morris, J. E., and Laycock, P. J. 1974. Discriminant analysis of directional data. Biometrika 61 335-341.

Mardia, K. V. 1975. Characterizations of directional distributions. In Statistical Distributions in Scientific Work, Patil, G. P., Kotz, S., Ord, J. K. (eds), Reidel, Dordrecht, vol. 3, 365-385.

Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. Multivariate Analysis. Academic Press. London.

Song, H, Liu, J., Wang, G. 2012. High-order parameter approximation for von Mises-Fisher distributions. Applied Mathematics and Computation 218 11880-11890.

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

Watamori, Y. 1995. Statistical inference of Langevin distribution for directional data. Hiroshima Mathematical Journal 26 25-74