

Discriminant analysis of von Mises-Fisher distributions

Analisi discriminante delle distribuzioni di von Mises-Fisher

Mario Romanazzi

Abstract The geometry of von Mises-Fisher distributions offers useful clues to discriminant analysis. The discriminant functions turn out to depend on the cosine of a non random vector with a von Mises-Fisher random vector and their probability density functions are explicitly derived. This considerably improves our understanding of discriminant analysis with hyper - spherical data in arbitrary dimension, especially in the two-group case. Specific applications include misclassification probabilities and ROC curve.

Abstract *La geometria delle distribuzioni di von Mises-Fisher è di grande aiuto nell'analisi discriminante. Le funzioni discriminanti sono il coseno di un vettore non stocastico con un vettore aleatorio della famiglia von Mises-Fisher e le loro funzioni di densità sono ricavate in forma esplicita. Il risultato consente un più agevole trattamento dei problemi di classificazione con dati sferici in spazi di dimensione arbitraria, specialmente nel caso di due gruppi. Particolare interesse rivestono nelle applicazioni le probabilità di errata classificazione e la curva ROC.*

Key words: Directional data, Bayes' discriminant rule, Maximum likelihood discriminant rule, Misclassification probability, ROC curve

1 Introduction

There is a renewed interest in classification problems with data distributed on the surface of hyper-spheres (Banerjee et al, 2005). Fields of applications include compositional data (Stephens, 1982), text categorization and gene expression analysis, where data are frequently normalized to unit euclidean norm. Dimension is typically higher than just two or three so as computational efficiency is an important requirement.

Mario Romanazzi
Ca' Foscari University, Venezia, e-mail: romanaz@unive.it

In this work the classical discrimination problem with von Mises-Fisher distributions is addressed. Related contributions are Morris and Laycock (1974) and El Khattabi and Streit (1996), both confined to circular and spherical data, hence of little help in the applications mentioned at the outset. Here the problem is attacked from a different point of view. The von Mises-Fisher density function depends on the data through the cosine with a fixed vector. For all dimensions $p \geq 2$, the cosine transformation originates a random variable defined in $[-1, 1]$ with an explicitly available probability density function. This random variable is tightly related to the discriminant functions thus allowing an easier derivation of several parameters, like misclassification probabilities and ROC curve. The results are particularly encouraging in the two-group situation.

The structure of the paper is as follows. In Section 2 some basic notions related to von Mises-Fisher distributions are reviewed and the probability distribution of the cosine transformation is derived. Section 3 is devoted to the discriminant problem and Section 4 gives an application.

2 von Mises-Fisher distribution and cosine transformation

Let $\mathbb{R}^p$, $p \geq 2$ denote the p -dimensional euclidean space with general point $x = (x_1, \dots, x_p)^T$ and euclidean norm $\|x\| = (x^T x)^{1/2}$. Let 0_p denote the zero vector of $\mathbb{R}^p$. The cosine of $x, y \in \mathbb{R}^p$, $x, y \neq 0_p$, is $x^T y / (\|x\| \|y\|)$. The unit radius hypersphere of $\mathbb{R}^p$, centered at 0_p , is $\mathcal{S}_p = \{x \in \mathbb{R}^p : \|x\| = 1\}$.

The von Mises-Fisher distribution (vMF) is a classical symmetric model for data distributed on $\mathcal{S}_p$, indexed by 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 $\mathcal{S}_p$ is obtained and, when $\kappa > 0$, the distribution is unimodal, with modal direction μ , and κ measures the concentration of data around μ . We write $X \sim \mathcal{M}_p(\mu, \kappa)$ for a p -dimensional random vector of this family. Let $I_s(\cdot)$ denote the modified Bessel function of first kind and order s . For $x \in \mathcal{S}_p$, the probability density function is

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

where $d\mathcal{S}_p$ is the probability element of $\mathcal{S}_p$ and $c_p(\kappa) = \kappa^{p/2-1} / ((2\pi)^{p/2} I_{p/2-1}(\kappa))$.

The density contours are indexed by the rescaled cosine values $\kappa \mu^T x$, $x \in \mathcal{S}_p$, with a maximum at $x = \mu$, the modal direction, and a minimum at the antipodal point $x = -\mu$. It is well-known that vMF family is closed under rotations and this property will be exploited in the following developments.

We will also need the spherical polar coordinates of $X \sim \mathcal{M}_p(\mu, \kappa)$, forming the $p-1$ -dimensional random vector $\Theta = (\Theta_1, \dots, \Theta_{p-1})^T$. The corresponding transformation is $X = \psi(\Theta) = (\psi_1(\Theta), \dots, \psi_p(\Theta))^T$, $0 \leq \Theta_i \leq \pi$, $i = 1, \dots, p-2$, $0 \leq \Theta_{p-1} < 2\pi$, where, for $i = 1, \dots, p$,

$$\psi_i(\Theta) = \cos \Theta_i \prod_{j=0}^{i-1} \sin \Theta_j, \quad \sin \Theta_0 = \cos \Theta_p = 1. \quad (2)$$

For a fixed vector $a \in \mathcal{S}_p$ and $X \sim \mathcal{M}_p(\mu, \kappa)$, consider the random variable $U_a = a^T X$, to be interpreted as the cosine of the angle formed by a and X . Of course $-1 \leq U_a \leq 1$, the minimum and maximum being reached when $X = \mp a$, respectively. The cumulative distribution function (cdf) $F_U^{(a)}$ and the probability density function (pdf) $f_U^{(a)}$ are derived below. First, an invariance argument is used. Let Q^* be the orthogonal matrix satisfying $Q^* a = e_1 = (1, 0, \dots, 0)^T$ and observe that $X^* = Q^* X \sim \mathcal{M}_p(Q^* \mu, \kappa)$.

Lemma 1. *Let Θ^* be the angular version of $X^* = Q^* X$, with $X_1^* = \cos \Theta_1^*$. Then, for $-1 \leq t \leq 1$*

$$F_U^{(a)}(t) = P(U_a \leq t) = P(X_1^* \leq t) = 1 - P(\Theta_1^* < \arccos t). \quad (3)$$

Proof. The proof follows because $U_a = a^T X = (Q^* a)^T (Q^* X) = e_1^T X^* = X_1^*$. $\square$

The pdf of U_a is given below. Write $\alpha = (\alpha_1, \dots, \alpha_{p-1})^T$ for the angular version of μ and note that $Q^* \mu = (\mu_1^*, \dots, \mu_p^*)^T$, with $\mu_1^* = a^T \mu = \cos \alpha_1^*$.

Theorem 1. *Put $\kappa_1(a) = \kappa \cos \alpha_1^* = \kappa a^T \mu$, $\kappa_2(a) = \kappa \sin \alpha_1^* = \kappa(1 - (a^T \mu)^2)^{1/2}$.*

i. If $\cos \alpha_1^ \neq \pm 1$, i. e., $a \neq \mp \mu$, for $-1 \leq t \leq 1$*

$$f_U^{(a)}(t) = \left(\frac{\kappa}{2\pi} \right)^{1/2} \frac{I_{(p-1)/2-1}(\kappa_2(a)(1-t^2)^{1/2})}{I_{p/2-1}(\kappa)} \sin^{-(p-3)/2} \alpha_1^* \quad (4)$$

$$\times (1-t^2)^{(p-3)/4} \exp(\kappa_1(a)t).$$

ii. If $\cos \alpha_1^ = \pm 1$, i. e., $a = \mp \mu$, for $-1 \leq t \leq 1$*

$$f_U^{(a)}(t) = \left(\frac{\kappa^{p-2}}{2^{p-2}\pi} \right)^{1/2} \frac{1}{I_{p/2-1}(\kappa)\Gamma((p-1)/2)} (1-t^2)^{(p-3)/2} \exp(\kappa_1(a)t) \quad (5)$$

Proof. Switch to Θ^* , the angular version of X^* , and obtain the distribution of the first marginal component Θ_1^* . The proof then follows by differentiation of $F_U^{(a)}$ in Lemma 1. $\square$

The following remark gives a geometrical interpretation of U_a .

Remark 1. Suppose $p = 3$. For $x \in \mathbb{R}^3$ and $t \in \mathbb{R}$, the equation $a^T x = t$ describes the family of parallel hyperplanes orthogonal to a . If $-1 < t < 1$, they intersect the sphere $\mathcal{S}_3$ in a set of parallel circles and if $t = \pm 1$ they are tangent to $\mathcal{S}_3$ at $\pm a$, the intersection points of the line λa , $\lambda \in \mathbb{R}$, with $\mathcal{S}_3$. It follows that $A_t = \{x \in \mathcal{S}_3 : a^T x \leq t\}$ is the spherical cap including $-a$ whose boundary is the circle $\{x \in \mathcal{S}_3 :$

$a^T x = t\}$. Rotating the sphere according to Q^* amounts to rotate the coordinate axes so as a coincides with $e_1 = (1, 0, 0)^T$, the direction from 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.

3 Discriminant analysis

Suppose the distribution of the random vector $X = (X_1, \dots, X_p)^T$ to depend on a partition of the sample space in $G \geq 2$ classes C_g , $1 \leq g \leq G$. The conditional random vector X , given C_g , is denoted with $X^{(g)}$ and is assumed to belong to vMF family for all $1 \leq g \leq G$, with parameters dependent on the class, i. e., $X^{(g)} \sim \mathcal{M}_p(\mu^{(g)}, \kappa_g)$. Let $x \in \mathcal{S}_p$ be the data vector observed on a unit to be allocated to one of the classes. Two popular allocation criteria are Bayes' and maximum likelihood rules stating that x must be assigned to the class providing maximum posterior probability or maximum likelihood, respectively. In terms of maximum log - likelihood rule, we have to maximize the function

$$d_g(x) = \ln c_p(\kappa_g) + \kappa_g \mu^{(g)T} x, \quad (6)$$

with respect to $1 \leq g \leq G$. It is clear that, conditional on C_g , $1 \leq g \leq G$, d_g functions are affine transformations of U_a -type random variables, with $a \equiv \mu^{(g)}$.

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 $d_1(x) \geq d_2(x)$, that is, $\tilde{\mu}^T x \geq \gamma$, where

$$\tilde{\mu} = \frac{\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}}{\|\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}\|}, \quad (7)$$

and

$$\gamma = \frac{\ln(c_p(\kappa_2)/c_p(\kappa_1))}{\|\kappa_1 \mu^{(1)} - \kappa_2 \mu^{(2)}\|}. \quad (8)$$

Once again, the discriminant variable $D_{12} = \tilde{\mu}^T X$ is a U_a -type random variable, with known class-conditional distributions. This allows an easy evaluation of several summaries, including misclassification probabilities. The probability of erroneously allocating a unit to C_1 , given that it belongs to C_2 , is $P_{1|2} = P(\tilde{\mu}^T X^{(2)} \geq \gamma)$ and the probability of erroneously allocating a unit to C_2 , given that it belongs to C_1 , is $P_{2|1} = P(\tilde{\mu}^T X^{(1)} < \gamma)$.

The overall performance of a classification rule in the two-group case is conveniently displayed by the ROC curve (Krzanowski and Hand, 2009) $\zeta = \{(\zeta_1(\gamma), \zeta_2(\gamma)), \gamma \in \Gamma\}$, where $\zeta_1(\gamma)$, $\zeta_2(\gamma)$ are the false positive rate (FPR) and the true positive rate (TPR), respectively. In the present context, identifying the *positive (negative)* pop-

Class	$\hat{\mu}$				$\hat{\kappa}$
Men	.580	.626	.398	.336	16.5
Women	.863	.130	.438	.217	22.1

Table 1 Household expenditure. Maximum likelihood estimates of centers ($\hat{\mu}$) and concentration parameters ($\hat{\kappa}$) of class conditional von Mises-Fisher distributions.

ulation with C_1 (C_2), for a given value of γ , $\zeta_1(\gamma) = P_{1|2}$ and $\zeta_2(\gamma) = P_{1|1} = P(\tilde{\mu}^T X^{(1)} \geq \gamma)$. As the domain of γ we take $\Gamma = \{-1 \leq \gamma \leq 1\}$.

The allocation domains of ML rule are described below.

Remark 2. Suppose, for simplicity, $p = 3$. The plane $\tilde{\mu}^T x = \gamma$, $-1 \leq \gamma \leq 1$, intersects the sphere $\mathcal{S}_3$ in a circle. Therefore, the domain of allocation to C_1 , i. e., the set $\{x \in \mathcal{S}_3 : \tilde{\mu}^T x \geq \gamma\}$ is the spherical cap cut by such a plane which includes $\tilde{\mu}$ and the domain of allocation to C_2 is the complementary spherical cap, which includes $-\tilde{\mu}$. The boundary circle is the set of equal density of the two classes. The normal direction of the cutting plane, the $\tilde{\mu}$ -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 $\mu^{(2)}$. This holds for general dimension p .

4 An application

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

As an illustration, we use the household expenditure data set from R package HSAUR2 (Everitt and Hothorn, 2013). Units are 20 single men and 20 single women and variables are expenditures on housing, food, other goods and services (currency is Hong Kong dollars). Our aim is to compare the expenditure profiles of men and women and evaluate class overlap. After normalization by division of each unit expenditure vector by its total expenditure, data were mapped to $\mathcal{S}_4$ by the squared root transformation (Stephens, 1982). This approach allows data to be described by vMF distributions, characterized by a unique center μ and a scalar scatter parameter κ . Two vMF distributions were separately fit to men's class and women's class using the improved ML estimators provided by R package movMF (Hornik and Grun, 2013). The estimates in Table 1 show a tendency for men to spend more on food whereas women more heavily weight housing. Moreover, women expenditure profiles appear comparatively more concentrated around the center. The class-conditional plug-in densities of the discriminant function and the ROC curve are shown in Figure 1. The estimated misclassification probabilities are about .115 for men and .0945 for women and the AUC (area under curve) value of ROC curve is .961.

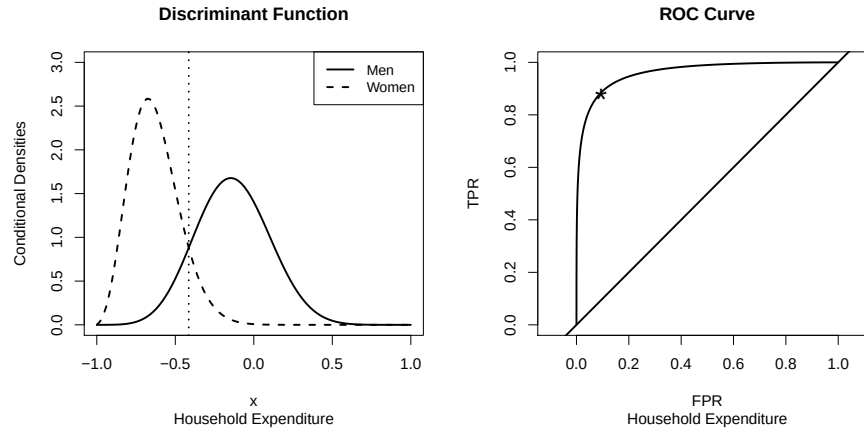


Fig. 1 Household expenditure. Left: Plug-in densities of the discriminant function and γ threshold (vertical line). Right: ROC curve and point corresponding to ML rule (star).

References

- Banerjee A, Dhillon IS, Gosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises - Fisher distributions. *Journal of Machine Learning Research* 6:1345 – 1382
- El Khattabi S, Streit F (1996) Identification analysis in directional statistics. *Computational Statistics & Data Analysis* 23:45 – 63
- Everitt BS, Hothorn T (2013) HSAUR2: A handbook of statistical analyses using R (2nd edition). R package version 1.1-6.4
- Hornik K, Grün B (2013) movMF: An R package for fitting mixtures of von Mises - Fisher distributions. R package version 0.1-2
- Krzanowski WJ, Hand DJ (2009) ROC curves for continuous data. CRC Press, London
- Morris JE, Laycock PJ (1974) Discriminant analysis of directional data. *Biometrika* 61:335–341
- Song H, Liu J, Wang G (2012) High - order parameter approximation for von Mises - Fisher distributions. *Applied Mathematics and Computation* 218:11,880–11,890
- 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
- Stephens M (1982) Use of the von Mises distribution to analyse continuous proportions. *Biometrika* 6:170–176