Liet. matem. rink., T. 41, spec. nr., 463–468

Discriminant space for mixtures of elliptical distributions in classification problem

Marius RADAVIČIUS (VGTU)
e-mail: mrad@ktl.mii.lt

1. Introduction

Let a d-dimensional random feature vector X be observed and the problem is to classify it to one of q classes (populations) Ω_i, $i = 1, \ldots, q$. In practice one rather frequently meets a situation where the number of classes q is much less than the dimension of X, $q \ll d$. Therefore it is natural to expect that X contains a lot of redundant information useless for classification. In other words, it is supposed that there exists a mapping $T\colon \mathbf{R}^d \to \mathbf{R}^k$, $T \in \mathcal{T}$, such that $k < d$ and $T(X)$ preserves all statistical information of the observation X about the number of a population it is drawn from. Here $\mathcal{T}$ is some given class of permissible mappings. Further we assume that k is the least dimension for which such a mapping does exist.

For Gaussian mixture model and $\mathcal{T}$ consisting of all linear mappings, the mapping T can be identified with a projection (in the Mahalanobis metric) onto some linear subspace, say H, of the dimensionality $\dim H = k$. This linear subspace is called *a discriminant space* (DS) (a formal definition is given in Definition 2). The assumption that $\mathcal{T}$ is the class of all linear mappings is natural when covariance matrices of Gaussian mixture components are equal but is not completely justified in a general case quadratic forms being a reasonable alternative.

For mixtures of elliptical distributions, a natural generalization of Gaussian mixtures, the class $\mathcal{T}$ must necessarily include *nonlinear* mappings otherwise T is trivial (a proof of this statement, almost obvious from intermediate arguments, is given in Section 3). Nevertheless a reasonable definition of the DS can be given (Radavičius (2001)).

In the paper a modification of the DS definition is proposed and extension of the DS characterization result announced in (Radavičius (2001)) is obtained.

2. Discriminant space definition

Let Y_i be a d-dimensional random vector representing the distribution of the feature vector X in the population Ω_i, $i = 1, \ldots, q$. Then one can write

$$X = Y_\nu$$

where ν is a discrete random variable, inependent of Y_i, $i = 1, \ldots, q$, taking on values $1, \ldots, q$ with respective probabilities p_i, $i = 1, \ldots, q$. The probability p_i is *a priori* probability of drawing X from the population Ω_i, $i = 1, \ldots, q$.

Aposterior probabilities

$$\pi_i(x) = \mathbf{P}\{\nu = i | X = x\}, \quad i = 1, \ldots, q,$$

play a crutial role in classification. According to the Bayes theorem

$$\pi_i(x) = \frac{p_i f_{Y_i}(x)}{f_X(x)},$$

where $f_X(x)$ and $f_{Y_i}(x)$ are distribution densities of the random vectors X and Y_i, respectively, $i = 1, \ldots, q$. Since ν given that $X = x$ has the multinomial distribution with probabilities $\pi_1(x), \ldots, \pi_q(x)$, i.e.,

$$\nu \mid \{X = x\} \sim Mult\big(1, \pi_1(x), \ldots, \pi_q(x)\big),$$

the aposterior probabilities $\{\pi_i(x),\ i = 1, \ldots, q\}$ contain all statistical information of the observation X about the parameter i (or the random variable ν in the Bayes' setting), the unobservable number of the population the observation X is drawn from. Let $\mathcal{T}$ be a fixed class of transformations of $x \in \mathbf{R}^d$.

DEFINITION 1. A mapping $T\colon \mathbf{R}^d \to \mathbf{R}^k$, $T \in \mathcal{T}$, with the least dimension k such that

$$\pi_i(x) = \mathbf{P}\{\nu = i | T(X) = T(x)\} \quad \forall x \in \mathbf{R}^d,\ i = 1, \ldots, q,$$

is called a minimal sufficient statistic for classification (discrimination) in the class $\mathcal{T}$ or shortly *minimal discriminant statistic.*

In the sequel we assume that X has finite second momemts and is standardized, i.e.,

$$\mathbf{E}X = 0_d, \quad \operatorname{cov}(X, X) = I_d,$$

where 0_d is a d-dimensional null-vector and I_d is a $d \times d$ unit matrix. Given a linear subspace $L \subset \mathbf{R}^d$, the orthogonal projection of $x \in \mathbf{R}^d$ onto L is denoted by x_L.

If $\mathcal{T}$ is the class of all linear transformations, then a minimal discriminant statistic $T \in \mathcal{T}$ can be identified with a projection onto some linear subspace, and we arrive to the following definition.

DEFINITION 2 (Rudzkis and Radavičius (1997,1999)). A linear subspace $H \subset \mathbf{R}^d$ satisfying the condition

$$\mathbf{P}\{\nu = i | X = x\} = \mathbf{P}\{\nu = i | X_H = x_H\} \quad \forall x \in \mathbf{R}^d,\ i = 1, \ldots, q,$$

and having the minimal dimension is called a *discriminant space* (DS) of X.

This definition is quite natural for mixtures of Gaussian random vectors. In (Rudzkis and Radavičius (1997,1999)) a complete characterization of the DS for Gaussian mixtures is given.

Now suppose that X is a mixture of elliptical (but non-Gaussian) random vectors Y_i, $i = 1, \ldots, q$ (see, for instance, Aivazyan *et al.* (1989), Gupta and Varga (1993)). Then the restriction of T to the class of linear mappings is too rigid since the only possible minimal discriminant statistic in this case is a trivial one, $T(X) \equiv X$. A formal statement this fact is presented in the next section.

DEFINITION 3 (cf. Radavičius (2001)). A linear subspace $H \subset \mathbf{R}^d$ is called a *discriminant space* for a random vector X being a non-trivial mixture of elliptical (non-Gaussian) random vectors Y_i, $i = 1, \ldots, q$, iff

$$T(X) = \left(X_H^T, |X - X_H|\right)^T \tag{1}$$

is a minimal discriminant statistic for classification.

3. Characterization of DS

First we present several well-known facts about elliptical distributios (see, e.g., Gupta and Varga (1993)).

Let Y_i be a d-dimensional elliptical random vector with parameters M_i, R_i, and ψ. A short notation of this statement is $Y_i \sim E_d(M_i, R_i, \psi)$, $i = 1, \ldots, q$. Hence,

$$\mathbf{E} \exp\{i\, t^T Y_i\} = \exp\{i\, t^T M_i\}\, \psi(t^T R_i t), \quad t \in \mathbf{R}^d,$$

and

$$\mathbf{E} Y_i = M_i, \quad \operatorname{cov}(Y_i, Y_i) = -2\psi'(0)\, R_i, \quad i = 1, \ldots, q.$$

If $Y \sim E_d(M, R, \psi)$ is absolute continuous, then distribution density $f_Y(y)$ of Y takes the following form:

$$f_Y(y) = \det(R)^{-1/2}\, h\big((y - M)^T R^{-1} (y - M)\big), \tag{2}$$

where the function h: $[0, \infty) \to [0, \infty)$ is uniquely determined by ψ and d.

REMARK 1. Without loss of generality in the sequel it is supposed that $\psi'(0) = -1/2$. Indeed, this can be always achieved simply by rescaling the function ψ and the matrices R_i, $i = 1, \ldots, q$ (see, e.g., Gupta and Varga (1993), Theorem 2.1.4). Then we have

$$R_i = \operatorname{cov}(Y_i, Y_i), \quad i = 1, \ldots, q.$$

It is said that the DS H is *trivial* if either $H = \{0_d\}$ or $H = \mathbf{R}^d$.

Theorem 1. *Assume that X is a mixture of absolute continuous elliptical random vectors*

$$Y_i \sim E_d(M_i, R_i, \psi), \quad i = 1, \dots, q,$$

with continuous distribution densities. If the DS H of X in the sense of Definition 2 is non-trivial, then X is a Gaussian mixture.

Outline of the proof. Since H is non-trivial we have $1 < k < d$ and $q > 1$. Definition 2 and the factorization theorem yield

$$f_{Y_i}(y) = A(y)\, B\big(y_H|i\big) \quad \forall y \in \mathbf{R}^d, \tag{3}$$

where $A\colon \mathbf{R}^d \to [0, \infty)$ and $B(\cdot|i)\colon H \to [0, \infty)$ are some measurable functions, $i = 1, \dots, q$. From (3) and (2) we have

$$h\big((x - M_i)^T R_i^{-1}(x - M_i)\big) = A(x)\, B_i\big(x_H\big) \quad \forall x \in \mathbf{R}^d, \tag{4}$$

for some measurable function B_i, $i = 1, \dots, q$. It is not difficult to make sure that this implies the positivity of the function h.

Set $y = y(t) = M_i + t\bar{x}$, $\bar{x} \in H^{\perp}$, $t \in \mathbf{R}$, where $H^{\perp}$ denotes the orthogonal complement of H in the space $\mathbf{R}^d$, and $\Delta_{ij} = M_j - M_i$, $i, j = 1, \dots, q$. Then, from (4) it follows that

$$h_1(t) \stackrel{\text{def}}{=} h\big((t\bar{x} + \Delta_{ij})^T R_j^{-1}(t\bar{x} + \Delta_{ij})\big) = c_{ij}\, h\big(t^2\, \bar{x}^T R_i^{-1} \bar{x}\big), \tag{5}$$

for some positive constants c_{ij} and for all $\bar{x} \in H^{\perp}$. This implies that h_1 is symmetric with respect to the two different centers and hence periodic unless

$$\bar{x}^T R_j^{-1} \Delta_{ij} = 0 \quad \forall i, j = 1, \dots, q. \tag{6}$$

The periodicity of h_1 contradicts to integrability of h with respect to the Lebesque measure. Thus, (6) holds. Since the random vector X is centered, we get

$$\bar{x}^T R_j^{-1} M_i = 0 \quad \forall \bar{x} \in H^{\perp},\ i, j = 1, \dots, q. \tag{7}$$

Let $\alpha_k \stackrel{\text{def}}{=} \bar{x}^T R_k^{-1} \bar{x}$, $k = 1, \dots, q$. Now suppose that $\alpha_i \neq \alpha_j$ for some i and j. Without loss of generality one can assume that $\alpha_i > \alpha_j$. If $c_{ij} = 1$, (5) shows that h is non-integrable. On the other hand, if $c_{ij} \neq 1$, h either is not continuous or is not positive. Thus,

$$\bar{x}^T R_i^{-1} \bar{x} = \bar{x}^T R_1^{-1} \bar{x} \quad \forall \bar{x} \in H^{\perp},\ i = 2, \dots, q. \tag{8}$$

Further, fix some $x = x_H + \bar{x}$, $\bar{x} \in H^{\perp}$, and set $y = y(t) = x_H + t\,\bar{x}$, $t \in \mathbf{R}$. From (4) we obtain that

$$h_2(t) \stackrel{\text{def}}{=} h\big(\alpha_1 t^2 + \beta_j t + \gamma_j\big) = c_{ij}\, h\big(\alpha_1 t^2 + \beta_i t + \gamma_i\big), \quad \forall t \in \mathbf{R}, \tag{9}$$

where $c_{ij}, \gamma_j, \gamma_i$ are certain functions of x_H and $\beta_k \stackrel{\text{def}}{=} \bar{x}^T R_k^{-1} x_H$, $k = 1, \ldots, q$. Again, (9) imlies $\beta_i = \beta_j$ otherwise h_2 being periodic and hence non-integrable. Since $x = x_H + \bar{x}$ is arbitrary and X is standardized we conclude that

$$y^T R_k^{-1} y = 1, \quad x^T R_k^{-1} y = 0$$

for all $x \in H$, $y \in H^{\perp}$, $k = 1, \ldots, q$. From the last equalities, (4), (8), and (7), it follows that

$$h\big(w + q_1(z)\big)\, h\big(q_2(z)\big) = h\big(w + q_2(z)\big)\, h\big(q_1(z)\big) \quad \forall w \geqslant 0, z \in \mathbf{R}, \tag{10}$$

where q_1 and q_2 are two different nonegative second order polynomials such that there exists real solutions of the equations $q_1(z) = 0$ and $q_1(z) = q_2(z)$. From (10) it is not difficult to derive that h is the exponential function with the negative exponent, i.e., $f_{Y_i}(y)$, $i = 1, \ldots, q$, are Gaussian distribution densities.

Theorem 2 (cf. Radavičius (2001)). *Assume that X is a mixture of absolute continuous random vectors $Y_i \sim E_d(M_i, R_i, \psi)$, $i = 1, \ldots, q$. Let*

$$H_0 \stackrel{\text{def}}{=} \big\{u \in \mathbf{R}^d\colon\ u^T M_i = 0,\ \alpha_i R_i u = u \ \text{ for some } \ \alpha_i > 0, \forall i = 1, \ldots, q\big\}^{\perp}.$$

Then the discriminant space H is a subset of H_0,

$$H \subset H_0. \tag{11}$$

If the function $h(t)$ defined in (2) is strictly monotonically decreasing for some $t_0 > 0$ and all $t \geqslant t_0$, then $H = H_0$.

REMARK 2. Scale mixtures of Gaussian distributions are typical examples of elliptical distributions with the strictly monotonically decreasing function h.

Outline of the proof. From the definition of H_0 it follows that distribution density of Y_i is of the form

$$f_{Y_i}(y) = \det(R_i)^{-1/2}\, h\big(\alpha_i |y - y_{H_0}|^2 + Q(y_{H_0}|i)\big) \tag{12}$$

where $Q(y_{H_0}|i)$ is a quadratic form of y_{H_0}, $i = 1, \ldots, q$. Consequently, $T_0(X) = (|X - X_{H_0}|, X_{H_0}^T)^T$ is a sufficient statistic for estimating i. Hence, the minimal discriminant statistic $T(X)$ is a measurable function of $T_0(X)$. Relation (11) follows herefrom.

On the other hand, the factorization theorem yields

$$f_{Y_i}(y) = A(y)\, B\big(T(y)|i\big) \quad \forall y \in \mathbf{R}^d, \tag{13}$$

where A and $B(\cdot|i)$ are some measurable functions, $i = 1, \ldots, q$. Since $h(t)$ in (2) is supposed to be strictly monotonically decreasing for $t \geqslant t_0$, (1), (2), and (13) imply (12) with H instead of H_0. This fact is proved by the similar arguments as in Theorem 1. Thus, $H = H_0$.

References

[1] S.A. Aivazyan, V.M. Buchstaber, I.S. Yenyukov, L.D. Meshalkin, Applied statistics, Classification and reduction of dimensionality, *Finansy i Statistika*, Moscow (1989) (in Russian).

[2] A.K. Gupta, T. Varga, *Elliptically Contoured Models in Statistics*, Kluwer, Dordrecht (1993).

[3] M. Radavičius, Discriminant space for mixtures of elliptical distributions, Computer data analysis and modeling, Robustness and computer intensive methods, *Proceedings of the 6th International Conference*, **2**, 172–177, Minsk (2001).

[4] R. Rudzkis, M. Radavičius, Mixtures of Gaussian distributions and projection pursuit preserving information about cluster structure, *Lithuanian Math. J.*. **37**, 416–425 (1997).

[5] R. Rudzkis, M. Radavičius, Characterization and statistical estimation of a discriminant space for Gaussian mixtures, *Acta Applicandae Mathematicae*, **58**, 279–290 (1999).

Diskriminantinė erdvė eliptinių skirstinių mišinių klasifikavimo uždavinyje

M. Radavičius

Darbe pasiūlytas diskriminantinės erdvės apibrėžimas eliptinių skirstinių mišinių klasifikavimo modelyje ir pateikta jos charakterizacija per modelio parametrus.