Affinity of densities and discriminant analysis

Rachid BOUMAZA
Département de mathématiques et informatique appliquées
Université Paul Valéry. Montpellier. France
boumaza@smr1.univ-montp3.fr

Bernard YCART
UFR de mathématiques et informatique
Université René Descartes. Paris. France
ycart@math-info.univ-paris5.fr

Abstract

An L^2 affinity measure of two densities is defined as their scalar product with respect to a reference measure. In the Gaussian case, when the densities are estimated by replacing their parameters by maximum likelihood estimators, the affinity measure is proved to be asymptotically normal, and the asymptotic variance is explicitly given. Grounded on these theoretical results, four new location criterions for discriminant analysis are defined. As an illustration, they are applied to the dating of a sample of Alsacian castles.

AMS 1991 subject classification: 62H20 (primary) 62H30 (secondary).

Key words and phrases: affinity measure, asymptotic normality, discriminant analysis, location criterions.

1 Introduction

Many statistical techniques, in particular in multivariate analysis, require an evaluation of distances between probability distributions. Very often, one has to evaluate the distance between an estimate of the unknown marginal distribution of a sample and another estimate, or a fixed probability distribution. Many different ways of evaluating those distances exist, but not so many potentially lead to explicit results. The distance that will be studied here is the L^2 distance between density functions. The Euclidean structure of L^2 makes it particularly attractive, although other tools have been prevalent in the past. Many applications could be discussed here, but we have chosen to focus on discriminant analysis of multivariate distributions.

The problem is classical. Let:

$$\mathcal{G} = \bigcup_{j=1}^{q} \mathcal{G}_j$$

be a set of probability distributions on $\mathbb{R}^p$, partitioned into q mutually disjoint subsets. Let $(X_1, \ldots, X_n)$ be an independent sample of some unknown probability distribution F on $\mathbb{R}^p$. The problem is to assign the distribution F to that class $\mathcal{G}_j$ to which it best fits, i.e. for which a certain predefined criterion is maximal. That problem arises in particular with the dating of 3-way data (see the different articles of [7], and in particular [6] as a general reference). In that context, we were asked to treat archaeological data concerning Alsacian castles (cf. [13, 14]), for which only partial dating was available, and we had to estimate building dates for those castles that had insufficient historical records.

Many criteria have been proposed in discriminant analysis. In [10], Matusita introduced a decision rule based on square roots of density functions. There the criterion to be minimized was the L^2 distance between the square root of the estimated marginal density of the sample, and that of an estimated reference density. The L^2 distance between square roots of densities has the obvious advantage that the norm of any density function is 1. The L^2 distance between densities has already been considered in a different context by Qannari [12]. Not replacing densities by their square roots permits to take advantage of linearity in certain density estimates. Although L^2 distances are easy to compute in many families of distributions, and in particular in exponential families, we chose to focus on the multivariate Gaussian case, which is by far the most frequent model in applications. After giving the explicit expression for the L^2 distance between two Gaussian densities (Proposition 2.1), we consider the case of Gaussian samples for which the mean and covariance matrix are estimated by their maximum likelihood estimator, and the density by the Gaussian density with estimated mean and covariance matrix. Our approach can be justified if replacing the true density by its estimate only induces a small error that can be statistically controlled. It is indeed the case: our main results (Theorems 2.1, 2.2, 2.3 and 2.4) show that when densities are replaced by estimates, the corresponding random L^2 scalar products are asymptotically normally distributed, and the asymptotic variances are explicitly computed.

The L^2 distance between square roots of densities that was used by Matusita [10] derived from Bhattacharyya's measure of affinity (see McLachlan [11]). In the Gaussian case with equal covariance matrices the decision rule is equivalent to the rule based on the Mahalanobis distance between means. Still in the Gaussian case, the asymptotic distribution of Matusita's distance has been studied by Bar-Hen and Daudin [1].

In the application of our approach to discriminant analysis, we introduced four different location criterions. The first two are based on L^2 distance between density functions and the two others are Bayesian, assuming that some prior probability π_j of belonging to $\mathcal{G}_j$, $j = 1, \ldots, q$, is known.

The paper is organized as follows. Section 2 contains theoretical developments about scalar products of distributions, and in particular the proof of asymptotic normality of these scalar products in the Gaussian case. Section 3 presents the four location rules that were considered for discriminant analysis of three-way data, and their comparison to more classical rules. The application to the dating of Alsacian castles is presented in section 4.

2 L^2 affinity measures of distributions

We shall use the classical notations $\langle\, . \,, \, . \,\rangle_\nu$ for the scalar product, and $\| \, . \, \|_\nu$ for the corresponding norm in $L^2(\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p}, \nu)$. When ν is the Lebesgue measure, scalar products and norms will simply be denoted by $\langle\, .\, ,\, .\, \rangle$ and $\|\, .\, \|$.

The L^2 affinity measure between two distributions on $I\!R^p$ whose density functions are f and g with respect to a measure ν on $I\!R^p$ was introduced by Qannari [12] as the scalar product $\langle f, g\rangle_\nu$ of f and g in $L^2(I\!R^p, \mathcal{B}_{I\!R^p}, \nu)$.

Definition 2.1 *Let ν be a non negative Radon measure on $I\!R^p$, and f, g, be two non negative functions such that $\int f^2 d\nu$ and $\int g^2 d\nu$ are finite. We call affinity measure of f and g relative to ν the following scalar product:*

$$\langle f, g\rangle_\nu = \int_{I\!R^p} f(x)\, g(x)\; d\nu(x)\ .$$

With these notations, Bhattacharyya's measure of affinity (McLachlan [11]) is $\langle f^{1/2}, g^{1/2}\rangle_\nu$. Replacing densities by their square roots has two obvious advantages. One is that if $f\, d\nu$ is a probability measure, then $\|f^{1/2}\|_\nu = 1$. The other one is that it measures the affinity between probability distributions, independently of the reference measure ν. Indeed, if ϕ is a strictly positive function on $I\!R^p$ and $d\nu' = \phi\, d\nu$, then the measures $f\, d\nu$ and $g\, d\nu$ have densities f/ϕ and g/ϕ with respect to ν'. But the scalar product $\langle (f/\phi)^{1/2}, (g/\phi)^{1/2}\rangle_{\nu'}$ is equal to $\langle f^{1/2}, g^{1/2}\rangle_\nu$. However bilinearity is an advantage, for instance when densities are estimated by kernel methods (Example 2.4 below), which is lost when using Bhattacharyya's measure.

We begin with some examples.

Example 2.1 *Exponential families of distributions.*

Let ν be a non negative Radon measure on $I\!R^p$, and ϕ be a measurable function from $I\!R^p$ into $I\!R^k$. For each $\theta \in I\!R^k$, denote by $L(\theta)$ the Laplace transform:

$$L(\theta) = \int_{I\!R^p} e^{\langle \theta, \phi(x)\rangle_{I\!R^k}}\; d\nu(x)\ ,$$

and by Θ the set of those θ for which $L(\theta)$ is finite ($\langle\, .\, ,\, .\, \rangle_{I\!R^k}$ denotes the standard scalar product of $I\!R^k$).

The exponential family generated by ν and ϕ is the following family of probability distributions (see Barndorff-Nielsen [2] as a general reference):

$$\mathcal{F}_{\nu,\phi} = \left\{ \frac{1}{L(\theta)} e^{\langle \theta, \phi(x) \rangle_{\mathbb{R}^k}} \, d\nu \, , \; \theta \in \Theta \right\} \, .$$

Let f and g be the densities with respect to ν of two elements of $\mathcal{F}_{\nu,\phi}$, and assume that they are square integrable with respect to ν.

$$f(x) = \frac{1}{L(\theta_f)} e^{\langle \theta_f, \phi(x) \rangle_{\mathbb{R}^p}} \quad \text{and} \quad g(x) = \frac{1}{L(\theta_g)} e^{\langle \theta_g, \phi(x) \rangle_{\mathbb{R}^p}} \, .$$

Then $\theta_f + \theta_g \in \Theta$, and:

$$\langle f, g \rangle_\nu = \frac{L(\theta_f + \theta_g)}{L(\theta_f) \, L(\theta_g)} \, .$$

Example 2.2 *Uniform distributions.*

Let ν be the Lebesgue measure on $\mathbb{R}^p$. Let D_f and D_g be two measurable domains of $\mathbb{R}^p$, with finite, strictly positive volumes. Let

$$f = \frac{1}{\mathrm{Vol}(D_f)} \mathbb{1}_{D_f} \quad \text{and} \quad g = \frac{1}{\mathrm{Vol}(D_g)} \mathbb{1}_{D_g} \, ,$$

be the densities with respect to ν of the uniform distributions on D_f and D_g respectively. Then:

$$\langle f, g \rangle = \frac{\mathrm{Vol}(D_f \bigcap D_g)}{\mathrm{Vol}(D_f) \; \mathrm{Vol}(D_g)} \, .$$

Example 2.3 *Gaussian distributions.*

The reference measure ν still is the Lebesgue measure on $\mathbb{R}^p$. Assume f and g are the density functions with respect to ν of two Gaussian distributions $\mathrm{N}(\mu, \Sigma)$ and $\mathrm{N}(m, V)$, where μ, m are two vectors of $\mathbb{R}^p$, and Σ, V are two positive definite matrices. The affinity measure between f and g with respect to ν has been computed in [3].

Proposition 2.1

$$\langle f, g \rangle = \frac{1}{(2\pi)^{\frac{p}{2}}} \frac{1}{|\Sigma + V|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mu - m)'(\Sigma + V)^{-1}(\mu - m)} \, . \tag{2.1}$$

(The determinant of a square matrix is denoted by $|\,.\,|$).
One could also normalize by $\|f\|$ and $\|g\|$, and compute the cosine of the angle between f and g (see [3]).

$$\langle \frac{f}{\|f\|}, \frac{g}{\|g\|} \rangle = 2^{\frac{p}{2}} \frac{|\Sigma|^{\frac{1}{4}} |V|^{\frac{1}{4}}}{|\Sigma + V|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mu - m)'(\Sigma + V)^{-1}(\mu - m)} \ .$$

This expression is very close to that of the Batthacharyya's affinity measure between f and g ([11], p. 23): the exponential term appears with 1/2 instead of 1/4.

When the means μ and m are equal, the exponential term is null and the L^2 affinity measure (2.1) depends on the determinant of $(\Sigma + V)$, that is the product of the eigenvalues of the sum of the covariance matrices.

Example 2.4 *Kernel estimates.*

Assume that the densities f and g have been estimated by a linear combination of uniform or Gaussian kernels (see Silverman [15]):

$$f = \sum_i \alpha_i \, K_i^f \quad \text{and} \quad g = \sum_j \beta_j \, K_j^g \ .$$

The affinity measure between f and g is a linear combination of affinities between kernels, directly deduced from one of the two previous examples.

$$\langle f, g \rangle = \sum_{i,j} \alpha_i \, \beta_j \ \langle K_i^f, K_j^g \rangle \ .$$

For the rest of this section, we shall focus on another type of density estimates in the Gaussian case. Let $(X_1, \ldots, X_{n_x})$ be a sample of the distribution $\mathrm{N}(\mu, \Sigma)$, with density function f. Denote by $\bar{X}$ and S_x the maximum likelihood estimators of μ and Σ respectively. The random function $f^{(n_x)}$ defined by:

$$f^{(n_x)}(z) = \frac{1}{(2\pi)^{\frac{p}{2}}} \frac{1}{|S_x|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(z - \bar{X})' S_x^{-1} (z - \bar{X})} \ ,$$

will be referred to as the density of the Gaussian distribution $\mathrm{N}(\bar{X}, S_x)$. Let g be the density function of the distribution $\mathrm{N}(m, V)$. By the law of large numbers, the affinity measure $\langle f^{(n_x)}, g\rangle$ converges almost surely to $\langle f, g\rangle$, as n_x tends to infinity. Symmetrically, let $(Y_1, \ldots, Y_{n_y})$ be a sample of the distribution $\mathrm{N}(m, V)$, independent from $(X_1, \ldots, X_{n_x})$. Let $\bar{Y}$ and S_y be the maximum likelihood estimators of m and V respectively, and $g^{(n_y)}$ be the density of the Gaussian distribution $\mathrm{N}(\bar{Y}, S_y)$. As n_x and n_y both tend to infinity, the affinity measure $\langle f^{(n_x)}, g^{(n_y)}\rangle$ converges almost surely to $\langle f, g\rangle$. Theorems 2.1 and 2.2 below show that the differences are asymptotically Gaussian. The case of a single estimate has already been treated in [5]. For sake of completeness and simplification of notations in other results, we recall it below.

Theorem 2.1 *As n_x tends to infinity, the random variable*

$$\sqrt{n_x}\,(\langle f^{(n_x)}, g\rangle - \langle f, g\rangle) \tag{2.2}$$

converges in distribution to the $\mathrm{N}(0, a[\mu, \Sigma, m, V])$, *with*

$$a[\mu, \Sigma, m, V] = \frac{|\Gamma| e^{-\delta'\Gamma\delta}}{(2\pi)^p} \left(\delta'\Gamma\Sigma\Gamma\delta \;+\; 2\,\mathrm{tr}\left[\left((-\Gamma + \Gamma\delta\delta'\Gamma + \frac{|\Gamma|}{2}\,E - \frac{1}{2}\,D^2)\Sigma \right)^2 \right] \right) \tag{2.3}$$

where $\Gamma = (\Sigma + V)^{-1}$, $\delta = (\mu - m)$, $E = \mathrm{diag}\{[\Gamma^{-1}]^c_{ii}\}$ *is the diagonal matrix of the cofactors of order* (i, i) *of* Γ^{-1} *and* $D = \mathrm{diag}\{\Gamma\delta\}$ *is the diagonal matrix of components of the vector* $\Gamma\delta$.

Proof.

We only give the main steps of the proof, referring to [4] for details. Let $\mathcal{M}_p$ be the set of positive definite matrices of order p. If $\nu \in I\!R^p$ and $K \in \mathcal{M}_p$, denote by $\psi(\nu, K)$ the affinity measure of the density of the Gaussian distribution $\mathrm{N}(\nu, K)$ with g. Since $\mathcal{M}_p$ is open, the function ψ is defined and indefinitely differentiable in a neighborhood of (μ, Σ). By the law of large numbers, for n_x large enough, $(\bar{X}, S_x)$ almost surely belongs to that neighborhood. We need to study the asymptotic normality of $\sqrt{n_x}(\psi(\bar{X}, S_x) - \psi(\mu, \Sigma))$. A Taylor expansion of order one in the neighborhood of (μ, Σ) gives:

$$\begin{array}{rcl} \psi(\nu, K) - \psi(\mu, \Sigma) & = & \sum_i \frac{\partial\psi}{\partial\nu_i}(\mu, \Sigma)\ (\nu_i - \mu_i) \\ & + & \sum_i \sum_j \frac{\partial\psi}{\partial K_{ij}}(\mu, \Sigma)\ (K_{ij} - \Sigma_{ij}) \\ & + & o(\|\nu - \mu\|_p + \|K - \Sigma\|_{p\times p})\ , \end{array}$$

where $\| \|_p$ and $\| \|_{p\times p}$ are any norm in the vector spaces of p-dimensional vectors and square matrices respectively, and o is a function such that $o(x)/x$ tends to 0 as x tends to 0. Thus the quantity of interest splits into three terms:

$$\begin{aligned}
\alpha_1(\bar{X}) &= \sqrt{n}\sum_i \frac{\partial\psi}{\partial\nu_i}(\mu,\Sigma)\ (\bar{X}-\mu)_i\ ,\\
\alpha_2(S_x) &= \sqrt{n}\ \sum_i\sum_j \frac{\partial\psi}{\partial K_{ij}}(\mu,\Sigma)\ (S_x-\Sigma)_{ij}\ ,\\
\alpha_3(\bar{X},S_x) &= \sqrt{n}\ o(\|\bar{X}-\mu\|_p\ +\ \|S_x-\Sigma\|_{p\times p})\ .
\end{aligned}$$

The random variable $\alpha_1(\bar{X})$ is Gaussian. Its mean is 0, and its variance is $(\nabla_\mu\psi)'\Sigma(\nabla_\mu\psi)$, where $\nabla_\mu\psi$ is the gradient of ψ with respect to ν, evaluated at (μ,Σ).

The term $\alpha_2(S_x)$ can be rewritten as

$$\alpha_2(S_x)\ =\ (\mathrm{vec}(\nabla_\Sigma\psi))'\ (\sqrt{n}\,\mathrm{vec}(S_x-\Sigma))\ ,$$

where 'vec' denotes the vectorization of matrices, and $\nabla_\Sigma\psi$ denotes the gradient matrix of ψ with respect to K, evaluated at (μ,Σ). The asymptotic normality of $\sqrt{n}\,\mathrm{vec}(S_x-\Sigma)$ is a standard result. It yields that of $\alpha_2(S_x)$. The asymptotic variance can be computed by standard linear algebra, using the identities given in Fang and Zhang [8], p.11 *ff.* It can be expressed as $2\,\mathrm{tr}\left[((\nabla_\Sigma\psi)\Sigma)^2\right]$.

Standard arguments show that $\alpha_3(\bar{X},S)$ converges in probability to 0. Due to the Gaussian hypothesis, the random variables $\bar{X}$ and S are independent and so are $\alpha_1(\bar{X})$ and $\alpha_2(S)$. The sum $\alpha_1(\bar{X})+\alpha_2(S)+\alpha_3(\bar{X},S)$ converges to a centered Gaussian distribution, the variance being the sum of variances of α_1 and α_2. In order to obtain the announced result, one has to check that:

$$(\nabla_\mu\psi)'\Sigma(\nabla_\mu\psi)\ +\ 2\,\mathrm{tr}\left[((\nabla_\Sigma\psi)\Sigma)^2\right]\ =\ a[\mu,\Sigma,m,V].$$

This is done using the results of Fang and Zhang [8], after some tedious linear algebra calculations that will not be reproduced here (see [4]).

□

When both densities are estimated, the result is the following.

Theorem 2.2 *Let α and β be two positive reals, $n_x = n_x(n)$ and $n_y = n_y(n)$ be two sequences of integers such that:*

$$\lim_{n\to\infty} \frac{n_x(n)}{n} = \alpha \quad \textit{and} \quad \lim_{n\to\infty} \frac{n_y(n)}{n} = \beta \; .$$

As n tends to infinity, the random variable

$$\sqrt{n}\,(\langle f^{(n_x)}, g^{(n_y)}\rangle - \langle f, g\rangle) \tag{2.4}$$

converges in distribution to the $\mathrm{N}(0, b[\mu, \Sigma, m, V])$*, with*

$$b[\mu, \Sigma, m, V] = \frac{1}{\alpha}\, a[\mu, \Sigma, m, V] + \frac{1}{\beta}\, a[m, V, \mu, \Sigma] \; . \tag{2.5}$$

Proof.

We rewrite $\sqrt{n}\,(\langle f^{(n_x)}, g^{(n_y)}\rangle - \langle f, g\rangle)$ as the sum:

$$\sqrt{n}\,(\langle f^{(n_x)}, g\rangle - \langle f, g\rangle) + \sqrt{n}\,(\langle f, g^{(n_y)}\rangle - \langle f, g\rangle) + \sqrt{n}\,\langle f^{(n_x)} - f, g^{(n_y)} - g\rangle \; .$$

The first two terms are independent. By Theorem 2.1, they are asymptotically Gaussian and their asymptotic variances are respectively $a[\mu, \Sigma, m, V]/\alpha$ and $a[m, V, \mu, \Sigma]/\beta$. There remains to prove that the third term converges to 0 in probability, which is easy, using the Schwarz inequality.

□

For two of the location criterions to be introduced in the next section, we shall need a multivariate Gaussian approximation. Theorems 2.3 and 2.4 below extend Theorems 2.1 and 2.2 to a vector of scalar products. Let $g_1, \ldots, g_q$ be the density functions of the Gaussian distributions $\mathrm{N}(m_1, V_1), \ldots, \mathrm{N}(m_q, V_q)$ on $\mathbb{R}^p$.

Theorem 2.3 *As n_x tends to infinity, the random vector*

$$\sqrt{n_x}\,\Big((\langle f^{(n_x)}, g_1\rangle, \ldots, \langle f^{(n_x)}, g_q\rangle) - (\langle f, g_1\rangle, \ldots, \langle f, g_q\rangle)\Big) \; . \tag{2.6}$$

converges in distribution to the q-dimensional $\mathrm{N}(0, A[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q])$*. For all $j, k = 1, \ldots, q$, the coefficient of order j, k of the covariance matrix A is:*

$$A_{jk} = \frac{|\Gamma_j \Gamma_k|^{\frac{1}{2}}}{(2\pi)^p} e^{-\frac{1}{2}(\delta_j' \Gamma_j \delta_j + \delta_k' \Gamma_k \delta_k)} \Big(\delta_j' \Gamma_j \Sigma \Gamma_k \delta_k + 2\,\mathrm{tr}[\Delta_j \Sigma \Delta_k \Sigma]\Big) \; , \tag{2.7}$$

where for $l = 1, \ldots, q$:

$$\begin{array}{rcl} \Gamma_l & = & (\Sigma + V_l)^{-1} \ , \\ \delta_l & = & \mu - m_l \ , \\ E_l & = & \mathrm{diag}\{[\Gamma_l^{-1}]^c_{ii}\} \ , \\ D_l & = & \mathrm{diag}\{\Gamma_l \delta_l\} \ , \\ \Delta_l & = & -\Gamma_l + \Gamma_l \delta_l \delta_l' \Gamma_l + \frac{|\Gamma_l|}{2} E_l - \frac{1}{2} D_l^2 \ . \end{array}$$

Proof.

Let $u = (u_j)$, $j = 1, \ldots, q$ be any vector of $I\!\!R^q$. Let

$$g = \sum_{j=1}^{q} u_j g_j \ .$$

We need to prove that the random variable $\sqrt{n_x}\,(\langle f^{(n_x)}, g\rangle - \langle f, g\rangle)$ converges in distribution to the $\mathrm{N}(0, u'\,A[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q]\,u)$. Re-defining ψ as the function that associates to (ν, K) the scalar product between the density of the $\mathrm{N}(\nu, K)$ and g, the arguments given in the proof of Theorem 2.1 remain valid, up to the computation of the asymptotic variance. Here, one has to prove that

$$(\nabla_\mu \psi)' \Sigma (\nabla_\mu \psi) \ + \ 2\,\mathrm{tr}\left[((\nabla_\Sigma \psi)\Sigma)^2\right] \ = \ u'\,A[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q]\,u \ .$$

Let ψ_j denote the function that associates to (ν, K) the scalar product between the density of the $\mathrm{N}(\nu, K)$ and g_j. One has $\psi = \sum_j u_j \psi_j$, and thus $\nabla_\mu \psi = \sum_j u_j \nabla_\mu \psi_j$ and $\nabla_\Sigma \psi = \sum_j u_j \nabla_\Sigma \psi_j$. The calculations follow exactly the same lines as those of the variance in Theorem 2.1. Notice that the diagonal coefficient $A[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q]_{jj}$ is $a[\mu, \Sigma, m_j, V_j]$, as was to be expected.

□

When the densities g_j have to be estimated, we shall assume that for each of them a sample $(Y_1^{(j)}, \ldots, Y_{n_y^j}^{(j)})$ is available, and that all the samples (including $(X_1, \ldots, X_{n_x})$) are independent. The estimator of the density g_j is denoted by $g_j^{(n_y^j)}$.

Theorem 2.4 *Let α, $\beta_1, \ldots, \beta_q$ be $q+1$ positive reals, $n_x = n_x(n)$, $n_y^1 = n_y^1(n), \ldots, n_y^q = n_y^q(n)$ be $q+1$ sequences of integers such that:*

$$\lim_{n\to\infty} \frac{n_x(n)}{n} = \alpha \quad \textit{and for all } j \quad \lim_{n\to\infty} \frac{n_y^j(n)}{n} = \beta_j \ .$$

As n tends to infinity, the random vector

$$\sqrt{n}\left((\langle f^{(n_x)}, g_1^{(n_y^1)}\rangle, \ldots, \langle f^{(n_x)}, g_q^{(n_y^q)}\rangle) - (\langle f, g_1\rangle, \ldots, \langle f, g_q\rangle)\right) \tag{2.8}$$

converges in distribution to the q-dimensional $\mathrm{N}(0, B[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q])$*, with*

$$B[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q] = \frac{1}{\alpha} A[\mu, \Sigma, m_1, V_1, \ldots, m_q, V_q] + \mathrm{diag}\{\frac{1}{\beta_j}\, a[m_j, V_j, \mu, \Sigma]\} \ . \tag{2.9}$$

Proof.

We proceed as in the proof of Theorem 2.2, rewriting (2.8) as the sum:

$$\begin{aligned} &\sqrt{n}\left((\langle f^{(n_x)}, g_1\rangle, \ldots, \langle f^{(n_x)}, g_q\rangle) - (\langle f, g_1\rangle, \ldots, \langle f, g_q\rangle)\right) && (2.10)\\ +&\sqrt{n}\left((\langle f, g_1^{(n_y^1)}\rangle, \ldots, \langle f, g_q^{(n_y^q)}\rangle) - (\langle f, g_1\rangle, \ldots, \langle f, g_q\rangle)\right) && (2.11)\\ +&\sqrt{n}\left((\langle f^{(n_x)} - f, g_1^{(n_y^1)} - g_1\rangle, \ldots, \langle f^{(n_x)} - f, g_q^{(n_y^q)} - g_q\rangle)\right) && (2.12) \end{aligned}$$

The first two terms are independent. The asymptotic normality of (2.10) is given by Theorem 2.3. That of each coordinate in the vector (2.11) follows from Theorem 2.1. Moreover these coordinates are independent and thus the vector converges to the centered Gaussian distribution with covariance matrix $\mathrm{diag}\{\frac{1}{\beta_j}\, a[m_j, V_j, \mu, \Sigma]\}$. Having checked that (2.12) tends to 0 in probability, the result follows.

□

3 Location rules

In discriminant analysis, one is given a set:

$$\mathcal{G} = \bigcup_{j=1}^{q} \mathcal{G}_j$$

of (Gaussian) distributions on $I\!\!R^p$, partitioned into q subsets, and a sample $(X_1, \ldots, X_n)$ of some unknown (Gaussian) distribution $\mathrm{N}(\mu, \Sigma)$, with density f. As in the previous section, we shall denote by $f^{(n)}$ the density of the Gaussian distribution $\mathrm{N}(\bar{X}, S_x)$, where $\bar{X}$ and S_x are the maximum likelihood estimators of μ and Σ respectively. The goal is to allocate the unknown density f to one of the classes $\mathcal{G}_j$, by maximizing a certain criterion:

$$\hat{\jmath} = \arg \max_{1 \leq j \leq q} C((X_1, \ldots, X_n), j) \ .$$

We describe below four possible criterions.

Criterion 1: *Affinity to class representative.*

Here we assume that each $\mathcal{G}_j$ can be represented by a Gaussian distribution $\mathrm{N}(m_j, V_j)$, with density g_j. This situation arises of course when each class contains a single distribution, but also when possibly different distributions in the class can only be estimated through a global sample. Our first criterion is the following:

$$C_1((X_1, \ldots, X_n), j) = \langle f^{(n)}, g_j \rangle \ . \tag{3.1}$$

Obviously, maximizing C_1 is equivalent to minimizing the L^2 distance between $f^{(n)}$ and g_j.

$$\hat{\jmath} = \arg \max_{1 \leq j \leq q} C((X_1, \ldots, X_n), j) = \arg \min_{1 \leq j \leq q} \| f^{(n)} - g_j \| \ .$$

Notice that Matusita's procedure ([10]) is the same, up to replacing $f^{(n)}$ and g_j by their square roots. In the particular case where all variances V_j are equal to a fixed V, formula (2.1) shows that maximizing C_1 is equivalent to minimizing the Mahalanobis distance of means:

$$\hat{\jmath} = \arg \min_{1 \leq j \leq q} \| \bar{X} - m_j \|_{(S_x + V)^{-1}} \ .$$

The convergence results of section 2 allow a statistical study of the estimator $\hat{\jmath}$. Assume that each $\mathcal{G}_j$ is reduced to a single known distribution, and that f is equal to g_{j_0}, with j_0 unknown. Then $\hat{\jmath}$ is a convergent estimator of j_0. Moreover Theorem 2.1 permits to compute the asymptotic probability that $\hat{\jmath} = j_0$, and also the asymptotic level and power of a test for "$H_0 : f = g_{j_0}$" against "$H_1 : f = g_{j_1}$". In the case where the g_j's are estimated through samples $(Y_1^{(j)}, \ldots, Y_{n_j}^{(j)})$, Theorem 2.2 can be applied with the same effect.

Criterion 2: *Affinity to class center.*

In the general case, each class $\mathcal{G}_j$ contains $T_j \geq 2$ distributions. Denote by $g_j^{(1)}, \ldots, g_j^{(T_j)}$ their densities. It seems natural to summarize the class by some convex combination of the $g_j^{(t)}$'s:

$$g'_j = \alpha_j^{(1)} g_j^{(1)} + \cdots + \alpha_j^{(T_j)} g_j^{(T_j)} , \tag{3.2}$$

with $\alpha_j^{(t)} \geq 0$ and $\alpha_j^{(1)} + \cdots + \alpha_j^{(T_j)} = 1$. Of course g'_j is not Gaussian anymore, but the advantage of the L^2 affinity measure over Bhattacharyya's is bilinearity. So minimizing the L^2 distance between $f^{(n)}$ and g'_j is equivalent to maximizing a convex combination of L^2 affinity measures:

$$C_2((X_1, \ldots, X_n), j) = \sum_{t=1}^{T_j} \alpha_j^{(t)} \langle f^{(n)}, g_j^{(t)} \rangle . \tag{3.3}$$

Obviously, C_2 and C_1 coincide when all T_j's are equal to 1. What had been said of the previous criterion regarding the application of Theorem 2.1 to the statistical study, can be extended to this one, using the multivariate convergence results instead. By Theorem 2.3, the asymptotic distribution of $C_2(X_1, \ldots, X_n, j)$ is Gaussian, with explicitly computable variance. Thus the asymptotic probability that $\hat{\jmath} = j$ can be computed for any f and j. Denoting by j_{max} the index j for which $\sum_{t=1}^{T_j} \alpha_j^{(t)} \langle f, g_j^{(t)} \rangle$ is maximum, one can thus compute the P-value and power of a test for "$H_0 : j_{max} = j_0$", against "$H_1 : j_{max} = j_1$".

Criterion 3: *Conditional joint likelihood.*

As in the first case, each class $\mathcal{G}_j$ is represented by a Gaussian density g_j. Denote by $Z = (Z_1, \ldots, Z_q)$ the vector of estimated affinities between f and the g_j's:

$$Z = (\langle f^{(n)}, g_1 \rangle, \ldots, \langle f^{(n)}, g_q \rangle) .$$

For all $j = 1, \ldots, q$, denote by $L_j(z_1, \ldots, z_q)$ the conditional density of Z, knowing that "$f = g_j$". Assume some prior distribution $\pi = (\pi_j)$ on the classes. For $j = 1, \ldots, q$, π_j is interpreted as the probability that f is equal to g_j. Our third criterion is:

$$C_3((X_1,\ldots,X_n),j) = \pi_j \, L_j(\langle f^{(n)}, g_1\rangle, \ldots, \langle f^{(n)}, g_q\rangle) \; . \tag{3.4}$$

The idea is the following. If $C_3(X_1,\ldots,X_n,j)$ is maximal, then so is the conditional probability that $f = g_j$, knowing the observed value of Z. Of course, the conditional densities $L_j(z_1,\ldots,z_q)$ cannot be explicitly computed. However, Theorem 2.3 proves that the conditional distribution of Z knowing "$f = g_j$" is asymptotically normal. So it is reasonable to replace $L_j(z_1,\ldots,z_q)$ by a multivariate Gaussian density, with mean $(\langle g_j, g_1\rangle, \ldots, \langle g_j, g_q\rangle)$ and covariance matrix determined by formula (2.7).

Criterion 4: *Conditional local likelihood.*

The situation and notations are those of criterion 3. Both the joint likelihood $L_j(z_1,\ldots,z_q)$ and its Gaussian approximation are uneasy to deal with. Instead of conditioning by the observed value of the vector Z of scalar products, the idea of criterion 4 is to condition only by its marginals. Thus we denote by $\ell_j(z)$ the conditional density of Z_j knowing "$f = g_j$". Our fourth criterion is:

$$C_4((X_1,\ldots,X_n),j) = \pi_j \, \ell_j(\langle f^{(n)}, g_j\rangle) \; . \tag{3.5}$$

Once again, ℓ_j will not be explicitly computed, but replaced by a Gaussian approximation, using Theorem 2.1. The inconvenient of criterion 4 is to remain local, not taking advantage of the full information contained in the vector Z, contrarily to criterion 3. On the other hand, calculations are much simpler with C_4 than with C_3.

In classical linear discriminant analysis, a single observation X is to be affected to a class and it is often assumed that the covariance matrices of the g_j's and f are all equal to some fixed matrix V. Re-defining $f^{(n)}$ as the density of the Gaussian distribution $\mathrm{N}(X,V)$, to apply criterions 1 or 4 amounts to minimizing the Mahalanobis distance between X and the vectors $m_1,\ldots,m_q$ (see for instance [9] p. 303).

$$\hat{\jmath} \;=\; \arg \max_{1\leq j\leq q} C_1(X,j)$$

$$
\begin{aligned}
&= \arg \max_{1 \leq j \leq q} C_4(X, j) \\
&= \arg \min_{1 \leq j \leq q} \|X - m_j\|_{V^{-1}} \ .
\end{aligned}
$$

4 Application to the dating of Alsacian castles

The benchmark data that we used were collected by J.M. Rudrauf [14] on Alsacian castles. On each castle, he measured the values of 4 structural parameters on a sample of building stones. The building dates of 40 castles were approximately known to historians, and partitioned into 5 consecutive periods ranging from 1140 to 1350, the first 4 periods being 35 years long and the last one 70 years long. Rudrauf's assumption was that the building techniques, and thus the structural parameters of the stones must have changed over the periods, and that the values of these parameters for a given castle should give an indication of the building period of the castle.

Our modelling hypothesis was that the four parameters measured on the stones of a given castle were independent realizations of a 4-dimensional Gaussian random vector, the mean and covariance matrix of which depended only on the castle. The number of stones measured on a given castle varied between 10 and 50. Obviously the normality hypothesis could not be validated on such small samples. Concatenating those small samples over each period gave numbers of stones ranging from 66 to 312 depending on the period. The mean and covariance matrices of those 5 global samples were computed and are presented in Table 1. Except between the first period and the others, these results do not show any striking difference between periods, and it seems difficult to use them for the dating of castles. We proposed to use the criterions defined in section 3 in order to locate the castles into periods. With the notations of the previous sections, if f is the 4-dimensional Gaussian density associated to the sample collected on a given castle, the dating problem consists in affecting f to one of the classes $\mathcal{G}_j = \{g_j^{(1)}, \ldots, g_j^{(T_j)}\}$, where $j = 1, \ldots, 5$ is the rank of the period and the $g_j^{(t)}$'s are the densities associated to the castles dated in the j-th period. In order to validate the method and compare the criterions by cross-validation, we successively applied the four criterions to each one of the 40 dated castles and checked

Table 1: Means and covariance matrices per period.

Period	Number of castles	Number of stones	Mean				Covariance			
1-Between 1140 and 1175	13	312	43.8	63.5	2.9	8.8	97.8			
							81.9	266		
							0.44	-0.17	0.64	
							13.2	13.2	-0.23	13.7
2-Between 1175 and 1210	6	155	34.8	54.7	3.2	7.7	33.8			
							12.1	155		
							1.05	1.81	0.59	
							3.46	2.90	0.05	5.77
3-Between 1210 and 1245	5	108	35.9	57.3	3.9	7.4	55.1			
							29.3	206		
							1.27	0.22	0.55	
							8.79	4.66	0.40	5.56
4-Between 1245 and 1280	13	273	32.6	54.2	3.8	6.2	39.0			
							20.0	171		
							1.90	2.16	0.92	
							1.61	-1.15	0.42	4.64
5-Between 1280 and 1350	3	66	34.9	54.4	4	6.7	26.4			
							24.3	277		
							-1.19	-1.94	0.68	
							5.62	8.03	-0.38	3.24

for misclassifications.

Criterion 1: *Affinity to class representative.*

Considering each period as homogeneous from the point of view of building parameters, we decided to associate to it a single Gaussian density function g_j on $\mathbb{R}^4$. Its mean and covariance matrix had to be estimated over all stones of the period. These estimates are those of Table 1.

Criterion 2: *Affinity to class center.*

Each castle of a given period being associated to a different Gaussian density, we chose to represent the period by the barycenter of those densities, giving the same importance to each density in the class. Thus the coefficients $\alpha_j^{(t)}$ of formula (3.2) were all equal to $1/T_j$.

Criterions 3 and 4: *Conditional joint or local likelihood.*

As for the first criterion each period was represented by a Gaussian density, with mean and covariance matrix given in Table 1. In the absence of any reliable information concerning undated castles, the prior probabilities π_j were supposed equal.

Having applied to a dated castle one of the four criterions, that castle was affected to a period that might or not be its own. Tables 2 (a) to (d) summarize the results obtained with criterions 1 to 4. In each table the entry of row i, column j is the number of castles of period i that were affected to period j by the corresponding criterion.

Table 2: Tables of location by rule and by period.

(a) Criterion 1

Location period				
1	2	3	4	5
10	3			
1	1	3		1
1			2	2
1	2	3	5	2
	1		2	

Misclassification: 24

(b) Criterion 2

Location period				
1	2	3	4	5
12	1			
2	1	3		
1	1	2	1	
3	1	3	6	
1	2			

Misclassification: 19

(c) Criterion 3

Location period				
1	2	3	4	5
9	3		1	
4		1		1
4			1	
4	2	4	2	1
			3	

Misclassification: 29

(d) Criterion 4

Location period				
1	2	3	4	5
10	3			
1	1	3		1
1			2	2
1	2	3	5	2
	1		2	

Misclassification: 24

The numbers of misclassifications are rather important. Criterion 2 seems to perform somewhat better than the other three, with a misclassification ratio a little below one half. These results may look disappointing, but one has to take into account the sparcity of information: small samples sizes, imprecision of the historical dating (each building date is known at best up to ten years), unknown variations in the building techniques. Also, looking at the entries of Table 2, one can observe that misclassifications by more than one period are rather scarce. Indeed the numbers of castles of period i located in a period j with $|j-i|>1$ are respectively 8, 8, 12, and 8 for criterions 1 to 4. In particular the 13 castles of period 1 are globally well located. The worst error rates are observed for the fourth period, probably indicating that the building techniques may not characterize the castles of that period.

Obviously, the results on Alsacian castles are not sufficient to truly validate our method nor to compare the four criterions. They are an example of a real-life application with many inherent difficulties. More tests need to be performed, in particular using the versatility of criterions 2, 3 and 4 that offer many choices of tuning parameters.

References

[1] A. Bar-Hen and J.J. Daudin. Asymptotic distribution of Matusita's distance: Application to the location model. *Biometrika*, 85 (2):477–481, 1998.

[2] O. Barndorff-Nielsen. *Information and exponential families in statistical theory.* Wiley, New York, 1978.

[3] R. Boumaza. Analyse en composantes principales de distributions gaussiennes multidimensionnelles. *Revue de Statis. Appl.*, XLVI (2):5–20, 1998.

[4] R. Boumaza. Analyses factorielles des distributions marginales de processus. *PhD dissertation, Université Joseph Fourier, Grenoble, France*, 1999.

[5] R. Boumaza. Distribution asymptotique de l'affinité L^2 de densités gaussiennes. *C. R. Acad. Sci. Paris*, 328, Série I, 527–529, 1999.

[6] A. Carlier, C. Lavit, M. Pagès, M.O. Pernin and J.C. Turlot. A comparative review of methods which handle a set of indexed data tables. In : R. Coppi and S. Bolasco (Ed.). *Multiway data analysis.* North-Holland, Amsterdam, 1989, 79–101.

[7] R. Coppi and S. Bolasco. *Multiway data analysis.* Proceedings of the International Meeting on the Analysis of Multiway Data Matrices, Rome, 1988. North-Holland, Amsterdam, 1989.

[8] K.T. Fang and Y.T. Zhang. *Generalized multivariate analysis.* Springer Verlag, New York, 1990.

[9] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis.* Academic Press, New York, 1992.

[10] K. Matusita. Classification based on distance in multivariate Gaussian case. In *Proc. 5th Berkeley Symp. (vol.1)*, 299–304. University of California Press, 1967.

[11] G.J. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley, New York, 1992.

[12] E.M. Qannari. *Analyses factorielles de mesures. Applications.* PhD dissertation, Université Paul Sabatier, Toulouse, France, 1983.

[13] J.M. Rudrauf. Petit-Geroldseck : Mise au point sur son origine. *Etudes Médiévales, Centre de recherches archéologiques médiévales de Saverne, France*, IV:89–120, 1987.

[14] J.M. Rudrauf and R. Boumaza. Contribution à l'étude de l'architecture médiévale : Les caractéristiques des pierres à bossage des châteaux alsaciens. Submitted, 1999.

[15] B.W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, London, 1986.

Title: Affinity of densities and discriminant analysis

Running title: Discriminant analysis of densities

Authors: Rachid BOUMAZA and Bernard YCART

Correspondence to:

Bernard YCART
UFR de mathématiques et informatique
Université René Descartes
45 rue des Saints Pères
75270 Paris Cedex 06
France

Tel. + 33 1 44 55 35 28
Fax. + 33 1 44 55 35 35
`ycart@math-info.univ-paris5.fr`

This manuscript was typeset in Latex 2.09 and can be transmitted electronically.