A Dissimilarity based on Relevant Population Features

M. Cubedo, A. Miñarro and J.M. Oller [*]
Departament d'Estadística
Universitat de Barcelona
08028 Barcelona, Spain

May 16, 2012

[*] This work is partially sponsored by MTM 2008–00642 (Ministerio de Ciencia e Innovación), Spain.

Abstract

In this paper a dissimilarity index between statistical populations is proposed without the hypothesis of a specific statistical model. We assume that the studied populations differ on some relevant features which are measured trough convenient parameters of interest. We assume also that we dispose of adequate estimators for these parameters. To measure the differences between populations with respect the parameters of interest, we construct an index inspired on some properties of the information metric which are also presented. Additionally, we consider several examples and compare the obtained dissimilarity index with some other distances, like Mahalanobis or Siegel distances.

Keywords and phrases: Information metric, Mahalanobis distance, Siegel distance, Dissimilarity index.

1 Introduction

The distance concept has been proved to be a very useful tool in data analysis and statistics, in order to study the similarity or dissimilarity between physical objects, usually, in a wide sense, populations. See for instance the classical works of Mahalanobis (1936), Bhattacharyya (1942) or Rao (1945), the foundations of the information metric studied in Atkinson and Mitchell, Burbea and Rao (1982), or Burbea (1986), some applications to statistical inference in Amari (1985), Oller and Corcuera (1995) or Cubedo and Oller (2002), applications to data analysis in Gabriel (1971), Huber (1985), Friedman (1987), Gower and Harding (1988), Greenacre(1993), Cook, Buja, Cabrera, and Hurley (1995) or Gower and Hand (1996) and more recent papers on foundations of distances in statistics like Lindsay, Markatou, Ray, Yang and Chen (2008) among many others.

In the present paper we are interested to define a distance between p different statistical populations $\Omega_1, \cdots, \Omega_p$. We assume that the studied populations differ on some relevant features which are measured trough convenient parameters of interest $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_k)^t$, although we shall not suppose a particular parametric statistical model for the distribution of data. After obtaining for every statistical population a data matrix X_i of order $n_i \times m$, we define certain convenient vectorial statistic $\boldsymbol{T} = (T_1, \ldots, T_k)^t$, and we assume that it is an unbiased and consistent estimator of $\boldsymbol{\xi}$ in each population. From this *relevant statistic* $\boldsymbol{T}$ we shall construct a dissimilarity index between statistical populations, which we shall refer hereafter as *Relevant Features Dissimilarity* (RFD). This index is inspired on some properties of the information metric. Additionally, a Multidimensional Scaling has been realized to compare the differences between our dissimilarity with Mahalanobis and Siegel distances.

2 Some remarks on the information metric

We first introduce some notation. Let χ be a sample space, $\mathcal{a}$ a σ–algebra of subsets of χ and μ a positive measure on the measurable space $(\chi, \mathcal{a})$. In the present paper, a *parametric statistical model* is defined as the *triple* $\{(\chi, \mathcal{a}, \mu)\,,\, \Theta\,,\, f\}$, where $(\chi, \mathcal{a}, \mu)$ is a measure space, Θ, also called the *parameter space*, is a manifold, and f is a measurable map, $f : \chi \times \Theta \longrightarrow \mathbb{R}$ such that $f(x, \theta) \geq 0$ and $dP_\theta = f(\cdot, \theta) d\mu$ is a probability measure on $(\chi, \mathcal{a})$, $\forall \theta \in \Theta$. We shall refer μ as the *reference measure* of the model.

Although in general Θ can be any manifold, for many purposes it will be enough to consider the case that Θ is a connected open set of $\mathbb{R}^n$, and, in this case, it is customary to use the same symbol (θ) to denote points and coordinates, notation which we are going to use hereafter. Additionally, we shall assume that function f satisfy certain standard regularity conditions which allow us to define on Θ a Riemannian structure induced by the probability measures, referred as the *information metric*, and given by

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}(\theta) d\theta^i d\theta^j$$

where $g_{ij}(\theta)$ are the components of the Fisher information matrix. For further details, see Amari (1985), Atkinson and Mitchell (1981), Burbea and Rao (1982) or Burbea (1986) among many others.

Let us denote the expectation, the covariance and the variance, with respect the probability measure $dP_\theta = f(\cdot, \theta) d\mu$, as $E_\theta(X)$, $\mathrm{cov}_\theta\,(X, Y)$ and $\mathrm{var}_\theta\,(X)$ respectively.

Let us define now $\mathcal{D}$ as the class of maps $X : \chi \times \Theta \longrightarrow \mathbb{R}$ such that $X(\cdot, \theta)$ is measurable map with finite variance with respect the probability measure dP_θ, $\forall \theta \in \Theta$, which additionally satisfy the condition

$$\frac{\partial}{\partial \theta^i}\, E_\theta(X(\cdot, \theta)) = \int_\chi X(x, \theta)\, \frac{\partial f(x, \theta)}{\partial \theta^i}\, d\mu, \quad \forall \theta \in \Theta, \quad i = 1, \ldots, n$$

Observe that this condition is fulfilled in many situations, for instance when X is not dependent on θ or in most of the cases when $E_\theta(\partial X(\cdot, \theta)/\partial \theta^i) = 0$, like $X(x, \theta) = \log f(x, \theta)$.

Additionally, let us consider the n-dimensional vector space:

$$H_\theta = < \frac{\partial \log f(\cdot,\theta)}{\partial \theta^1}, \ldots, \frac{\partial \log f(\cdot,\theta)}{\partial \theta^n} > \ \subset \quad L^2(f(\cdot,\theta)\ d\mu)$$

where we have identified, as usual, the functions $\partial \log f(\cdot,\theta)/\partial\theta^i$, which form a basis of H_θ, as elements of $L^2(f(\cdot,\theta)\ d\mu)$. Then H_θ inherits a scalar product, $<,>_{H_\theta} = \operatorname{cov}_\theta(\cdot,\cdot)$, since $E_\theta(\partial \log f(\cdot,\theta)/\partial\theta^i) = 0$. Observe that, with respect the above-mentioned basis, the corresponding scalar product matrix is the Fisher information matrix $G = (g_{ij}(\theta))$.

We can identify this Euclidean vector space with the tangent space at θ, Θ_θ, which also has an Euclidean structure induced by the information metric, through the correspondence:

$$\sum_{k=1}^{n} Z^k \frac{\partial \log f(\cdot,\theta)}{\partial \theta^k} \longleftrightarrow \sum_{k=1}^{n} Z^k \left(\frac{\partial}{\partial \theta^k}\right)$$

which is, clearly, a natural lineal isometry between H_θ and Θ_θ. Then, fixed θ, the tangent vectors may be viewed as random maps, and we shall make use of this identification, hereafter, when necessary.

Notice also that the projection map $\pi_\theta : L^2(f(\cdot,\theta)\ d\mu) \longrightarrow H_\theta$ is well defined, with $X(\cdot,\theta) - \pi_\theta(X(\cdot,\theta)) \in H_\theta^\perp$. The following proposition make explicit some properties of the above–mentioned identification.

Proposition 2.1 *Let $X \in \mathcal{D}$ and define $U(\cdot,\theta) = \pi_\theta(X(\cdot,\theta))$ and $V(\cdot,\theta) = X(\cdot,\theta) - U(\cdot,\theta)$. Then*

a) $U(\cdot,\theta) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} g^{ij}(\theta) \operatorname{cov}_\theta \left(X(\cdot,\theta), \frac{\partial \log f(\cdot,\theta)}{\partial \theta^j} \right) \right) \frac{\partial \log f(\cdot,\theta)}{\partial \theta^i}$

where $g^{ij}(\theta)$ are the elements of the inverse of the Fisher information matrix, $G = (g_{ij}(\theta))$.

Moreover, identifying the tangent vectors with the elements of H_θ, we have

b) $\operatorname{grad}(E_\theta(X(\cdot,\theta))) = U(\cdot,\theta)$

c) $\|\operatorname{grad}(E_\theta(X(\cdot,\theta)))\|_\theta^2 = \operatorname{var}_\theta(U(\cdot,\theta))$

d) $\operatorname{var}_\theta(X(\cdot,\theta)) = \|\operatorname{grad}(E_\theta(X(\cdot,\theta)))\|_\theta^2 + \operatorname{var}_\theta(V(\cdot,\theta))$

Proof:

Observe that $U(\cdot,\theta) = \pi_\theta(X(\cdot,\theta)) \in H_\theta$ then

$$U(\cdot,\theta) = U^1 \frac{\partial \log f(\cdot,\theta)}{\partial\theta^1} + \ldots + U^n \frac{\partial \log f(\cdot,\theta)}{\partial\theta^n}$$

where the coefficients U^i are determined in order to minimize $\|X(\cdot,\theta) - U(\cdot,\theta)\|_\theta^2$. It is straightforward to prove that

$$U^i = \sum_{j=1}^n g^{ij}(\theta)\ \mathrm{cov}_\theta\left(X(\cdot,\theta), \frac{\partial \log f(\cdot,\theta)}{\partial\theta^j}\right) \tag{1}$$

and *a)* follows. On the other hand, since $X \in \mathcal{D}$ and thus $E_\theta\left(\frac{\partial \log f(\cdot,\theta)}{\partial\theta^i}\right) = 0$, we have

$$\begin{aligned} \frac{\partial}{\partial\theta^i}\ E_\theta(X(\cdot,\theta)) &= \int_\chi X(x,\theta)\ \frac{\partial f(x,\theta)}{\partial\theta^i}\ \mu(dx) \\ &= \mathrm{cov}_\theta\left(X(\cdot,\theta), \frac{\partial \log f(\cdot,\theta)}{\partial\theta^i}\right) \end{aligned}$$

Therefore the i-th component of $\mathrm{grad}\,(E_\theta(X(\cdot,\theta)))$, with respect the basis vector field $\partial/\partial\theta^1, \ldots \partial/\partial\theta^n$, at θ, is given by

$$\mathrm{grad}\,(E_\theta(X(\cdot,\theta)))^i = \sum_{\alpha=1}^n g^{i\alpha}(\theta)\ \mathrm{cov}_\theta\left(U(\cdot,\theta), \frac{\partial \log f(\cdot,\theta)}{\partial\theta^\alpha}\right)$$

and taking into account (1) and the natural lineal isometry between H_θ and Θ_θ, we can write

$$\mathrm{grad}\,(E_\theta(X(\cdot,\theta))) = U(\cdot,\theta)$$

the square of its norm is given by

$$\|\mathrm{grad}\,(E_\theta(X(\cdot,\theta)))\,\|_\theta^2 = \mathrm{var}_\theta\,(U(\cdot,\theta))$$

and since $X(\cdot,\theta) = U(\cdot,\theta) + V(\cdot,\theta)$ and $\mathrm{cov}_\theta\,(U(\cdot,\theta), V(\cdot,\theta)) = 0$ we obtain

$$\mathrm{var}_\theta\,(X(\cdot,\theta)) = \|\mathrm{grad}\,(E_\theta(X(\cdot,\theta)))\,\|_\theta^2 + \mathrm{var}_\theta\,(V(\cdot,\theta))$$

completing the proof. ■

Let us define now $\mathcal{S} \subset \mathcal{D}$ as the subclass of maps such that $\mathrm{var}_\theta\,(X(\cdot,\theta)) \leq 1$ $\forall\theta \in \Theta$, then from Proposition 2.1, as an immediate consequence, we have the following corollary:

Corollary 2.2 *Let $X \in \mathcal{S}$ then*

$$\|\mathrm{grad}\,(E_\theta(X(\cdot,\theta)))\,\|_\theta \leq 1 \quad \forall \theta \in \Theta$$

We can prove also the following property

Proposition 2.3 *Let $X \in \mathcal{D}$, and let $\alpha : [a,b] \longrightarrow \Theta$ be a piecewise C^∞ curve parametrized by the arc length, $a, b \in \mathbb{R}$, $a < b$, such that $\nu = \alpha(a)$ and $\xi = \alpha(b)$. Let also define $\Phi_X(\theta) = E_\theta(X(\cdot,\theta))$ and $\Delta\Phi_X = \Phi_X(\xi) - \Phi_X(\nu)$. Then*

$$\left|\frac{d(\Phi_X \circ \alpha)}{ds}\Big|_{s=s^*}\right| \leq \sigma_X(\alpha(s^*)) \tag{2}$$

and

$$|\Delta\Phi_X| \leq \left(\max_{s\in[a,b]} \sigma_X\left(\alpha(s)\right)\right)\ l(\nu,\xi) \tag{3}$$

where $l(\xi,\nu)$ is the curve length between ν and ξ, and $\sigma_X(\theta) = \sqrt{\mathrm{var}_\theta\left(X(\cdot,\theta)\right)}$.

If $X \in \mathcal{S}$ and α is a minimal geodesic, then we shall have

$$|\Delta\Phi_X| \leq \rho(\nu,\xi) \tag{4}$$

being $\rho(\nu,\xi)$ the Rao distance between ν and ξ.

Proof:

With a customary notation, let us denote the curve α parametrized by the arc length s as $\alpha(s) = (\theta^1(s), \ldots, \theta^n(s))$, then the components of the tangent vector to α, in each smooth piece, are given by $d\alpha/ds = (d\theta^1/ds, \ldots, d\theta^n/ds)$, and, by the chain rule, using classical notation, we have

$$\frac{d(\Phi_X \circ \alpha)}{ds} =< \mathrm{grad}\,(\Phi_X)\,, \frac{d\theta}{ds} >_\theta$$

where $<,>_\theta$ is the scalar product either in Θ_θ, where the point θ is on the curve α.

From proposition (2.1) and Cauchy-Schwarz inequality, we have

$$\left|\frac{d(\Phi_X \circ \alpha)}{ds}\right| = \left| < \mathrm{grad}\,(\Phi_X)\,, \frac{d\theta}{ds} >_\theta \right| \leq \left\|\mathrm{grad}\,(\Phi_X)\right\|_\theta \left\|\frac{d\theta}{ds}\right\|_\theta$$

where $\theta = \alpha(s)$ and taking into account that since the curve is parametrized by the arc-length the norm of the tangent vector is equal to one and Proposition 2.1 we have

$$\left|\frac{d(\Phi_X \circ \alpha)}{ds}\right| \leq \sigma \circ \alpha$$

which prove (2), and therefore

$$\begin{aligned} |\Delta\Phi_X| &= \left|\int_a^b \frac{d(\Phi_X \circ \alpha)}{ds}\, ds\right| \leq \int_a^b \left|\frac{d(\Phi_X \circ \alpha)}{ds}\right| ds \\ &\leq \int_a^b \sigma(\alpha(s))\; ds \leq \left(\max_{s\in[a,b]} \sigma_X\left(\alpha(s)\right)\right)\; l(\nu,\xi) \end{aligned}$$

which prove (3). Additionally, if α is a minimal geodesic and $\operatorname{var}_\theta\left(X(\cdot,\theta)\right) \leq 1$ we obtain

$$|\Delta\Phi_X| \leq \rho(\nu,\xi)$$

completing the proof. ■

The following proposition shall supply an interpretation of the information metric.

Proposition 2.4 *Let $\theta_0 \in \Theta$ and $Z \in \Theta_{\theta_0}$ with $Z \neq 0$. Let $X \in \mathcal{S}$ and $\Phi_X(\theta) = E_\theta(X(\cdot,\theta))$. Moreover, let us define the map $w : \mathcal{S} \longrightarrow \mathbb{R}$ such that*

$$w(X) = Z(\Phi_X)$$

then

$$\max_{X\in\,\mathcal{S}} w(X) = \|Z\|_{\theta_0}$$

where the norm in Θ_{θ_0} is the norm corresponding to the information metric and a random map such that maximizes the above-mentioned expression is given by

$$X(\cdot,\theta_0) = Z^1\,\frac{\partial \log f(\cdot,\theta)}{\partial\theta^1}\bigg|_{\theta=\theta_0} + \ldots + Z^n\,\frac{\partial \log f(\cdot,\theta)}{\partial\theta^n}\bigg|_{\theta=\theta_0} + C$$

where $Z^1,\ldots,Z^n$ are the components of Z with respect the basis vector field $\partial/\partial\theta^1,\ldots\partial/\partial\theta^n$ at θ_0, and C is almost surely a constant (dependent only on θ_0).

Proof:

Taking into account Proposition 2.1 and the natural lineal isometry between H_θ and Θ_θ, we have that $w(X) = Z(\Phi_X) = <\mathrm{grad}\,(\Phi_X)\,, Z>_{\Theta_{\theta_0}} = < U, Z >_{H_{\theta_0}}$ where $U = U(\cdot,\theta) = \pi_\theta(X(\cdot,\theta))$ and Z is identified with the corresponding vector in H_{θ_0}. For a fixed norm of U, by Cauchy-Schwarz inequality $w(X)$ is maximized if we choose X such that $U(\cdot,\theta_0) = \lambda Z$ for a convenient λ but, since $X \in \mathcal{S}$, we also have that

$$|\lambda|^2 \|Z\|^2_{H_{\theta_0}} = \|U\|^2_{H_{\theta_0}} = \mathrm{var}_{\theta_0}\,(U(\cdot,\theta_0)) \leq \mathrm{var}_{\theta_0}\,(X(\cdot,\theta_0)) \leq 1$$

then $\lambda < 1/\|Z\|_{H_{\theta_0}}$. Combining these results, we obtain that $w(X) = \lambda \|Z\|^2_{H_{\theta_0}} \leq \|Z\|_{H_{\theta_0}}$ and therefore

$$\max_{X \in \mathcal{S}} w(X) = \|Z\|_{\theta_0}$$

maximum which is attained when we choose $\lambda = 1/\|Z\|_{H_{\theta_0}}$. Observe that when the maximum is attained we have $\mathrm{var}_{\theta_0}\,(U(\cdot,\theta_0)) = \mathrm{var}_{\theta_0}\,(X(\cdot,\theta_0)) = 1$ and thus $X(\cdot,\theta_0) - U(\cdot,\theta_0)$ must be almost surely a constant C, and then

$$X(\cdot,\theta_0) = U(\cdot,\theta_0) + C = \sum_{i=1}^{n} Z^i \left.\frac{\partial \log f(\cdot,\theta)}{\partial \theta^i}\right|_{\theta=\theta_0} + C$$

completing the proof. ■

This result can be interpreted as follows. Given two close statistical populations, corresponding to parameters θ_0 and $\theta_1 = \theta_0 + \Delta\theta$, and a random variable X, a measure, based on X, to quantify the difference between both populations can be the mean value change when we move from one population to the other. It suggests to define a mesure, independent of X, of the difference between populations, as the maximum of the previously mentioned dependent on X measures, with the restriction that $\mathrm{var}_{\theta_0}\,(X) = 1$. Observe that from Propositions 2.1 and 2.4 we have

$$\Delta\Phi_X = \Phi_X(\theta_1) - \Phi_X(\theta_0) \approx \sum_{i=1}^{n} \mathrm{cov}_{\theta_0}\left(X, \left.\frac{\partial \log f(\cdot,\theta)}{\partial \theta^i}\right|_{\theta=\theta_0} \right) \Delta\theta^i$$

Varying $X \in \mathcal{S}$, this quantity is maximized when we chose

$$X \approx \frac{1}{\Delta s} \sum_{i=1}^{n} \left.\frac{\partial \log f(\cdot,\theta)}{\partial \theta^i}\right|_{\theta=\theta_0} \Delta\theta^i$$

where

$$\Delta s = \sqrt{\sum_{i,j=1}^{n} g_{ij}(\theta) \Delta\theta^i \Delta\theta^j}$$

and the maximum is precisely equal to ΔS. In other words, the information metric is a local mesure of the maximum mean value change of standardized random variables, in a neighborhood of θ_0. This interpretation suggest a simple method to define a dissimilarity index to discriminate populations even when we do not have parametric statistical model as we shall see in the previous section.

3 Distances between statistical populations

Let us consider now the case that we are interested to define a distance between p different objects on which we have performed measurements obtaining results which may be identified as samples of a m-dimensional random vector X over p different statistical populations $\Omega_1, \ldots, \Omega_p$, obtaining, for every statistical population, a data matrix which rows are the X values over each particular individual of the corresponding sample, i.e. for the population Ω_i we obtain the data matrix $\boldsymbol{X}_i$ of order $n_i \times m$.

As we have said in the introduction, we shall assume that the statistical populations differ in some relevant features which are measured trough a vectorial parameter of interest $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^t$, although we shall not assume a particular parametric statistical model for the distribution of X. We may try to define a distance, or at least a dissimilarity index, between the populations $\Omega_1, \ldots, \Omega_p$, inspiring us in the results of the previous section, see proposition (2.4).

First of all we need to define certain convenient statistic $\boldsymbol{T} = (T_1, \ldots, T_k)^t$, where $\boldsymbol{T}$ is a vector valued measurable function of the samples $\boldsymbol{X}_i$, and we assume that it is an unbiased and consistent estimator of the parameter of interest $\boldsymbol{\xi}$ and it has covariance finite for each population. Therefore we shall write

$$\text{(5)} \qquad \boldsymbol{\xi}_i = \mathrm{E}(\boldsymbol{T}|\Omega_i) \quad \text{and} \quad \boldsymbol{\Psi}_i = \mathrm{cov}(\boldsymbol{T}|\Omega_i) \quad i = 1, \ldots, p$$

We shall also require that the covariance of $\boldsymbol{T}$ at each statistical population satisfy

$$\text{(6)} \qquad \lim_{n_i \to \infty} n_i \boldsymbol{\Psi}_i = \boldsymbol{\Xi}_i$$

where $\boldsymbol{\Xi}_i$ is a $k \times k$ symmetric, regular and positive definite matrix, which is, essentially, a standardized version of the $\boldsymbol{T}$ covariance matrix independent of sample size.

The values of the statistic $\boldsymbol{T}$ can be evaluated for each sample matrix $\boldsymbol{X}_i$, obtaining $\widehat{\boldsymbol{T}_i} \equiv \boldsymbol{T}(\boldsymbol{X}_i)$, and, in the absence of model assumptions, we can obtain also an unbiased estimation of $\boldsymbol{\Psi}_i$, $\widehat{\boldsymbol{\Psi}}_i$, through Bootstrap methods, by re-sampling the rows of the data matrix $\boldsymbol{X}_i$, and also an estimation of $\boldsymbol{\Xi}_i$, $\widehat{\boldsymbol{\Xi}}_i = n_i \widehat{\boldsymbol{\Psi}}_i$.

Under this framework, we can define a non necessarily symmetric dissimilarity index between the statistical populations Ω_i and Ω_j, δ_{ij}. We define this index as the supremum of the differences of the estimated expected values of any linear combination of $T_1, \ldots, T_k$, $\boldsymbol{\alpha}^t\boldsymbol{T}$, where α is a $k \times 1$ vector, between both populations under the constraint that the estimated standardized variance of this linear combination at Ω_i is less equal to one, i.e.

$$\delta_{ij} = \sup_{\boldsymbol{\alpha}^t\widehat{\boldsymbol{\Xi}}_i\boldsymbol{\alpha} \leq 1} \boldsymbol{\alpha}^t(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i) = \sup_{n_i\boldsymbol{\alpha}^t\widehat{\boldsymbol{\Psi}}_i\boldsymbol{\alpha} \leq 1} \boldsymbol{\alpha}^t(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i) \tag{7}$$

In order to find this supremum, we consider the following proposition

Proposition 3.1 *Given a $k \times 1$ vector $\boldsymbol{v}$, the supremum of any linear combination of its components $v_1, \cdots, v_k$, that is, $\boldsymbol{\alpha}^t\boldsymbol{v}$, where $\boldsymbol{\alpha}$ is a $k \times 1$ vector of real values, under the constraint $\boldsymbol{\alpha}^t\boldsymbol{A}\boldsymbol{\alpha} \leq 1$, where $\boldsymbol{A}$ is a $k \times k$ symmetric and positive definite matrix is*

$$\sup_{\boldsymbol{\alpha}^t\boldsymbol{A}\boldsymbol{\alpha} \leq 1} \boldsymbol{\alpha}^t\boldsymbol{v} = \sqrt{\boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{v}}$$

Proof:

Taking into account that $\boldsymbol{A}$ is symmetric and positive definite matrix, if we let $\boldsymbol{\beta} = \boldsymbol{A}^{1/2}\boldsymbol{\alpha}$ and $\boldsymbol{\omega} = \boldsymbol{A}^{-1/2}\boldsymbol{v}$, where $\boldsymbol{A}^{1/2}$ is the symmetric square root of $\boldsymbol{A}$, our optimization problem is equivalent to find a vector $\boldsymbol{\beta}$ such that maximices $\boldsymbol{\beta}^t\boldsymbol{\omega}$ subject to the constraint $\boldsymbol{\beta}^t\boldsymbol{\beta} \leq 1$. Clearly, by Cauchy-Schwarz inequality, the maximum is attained when $\boldsymbol{\beta} = \lambda\boldsymbol{\omega}$ and $\boldsymbol{\beta}^t\boldsymbol{\beta} = 1$. Therefore λ must be equal to $1/\sqrt{\boldsymbol{\omega}^t\boldsymbol{\omega}}$ and

$$\sup_{\boldsymbol{\beta}^t\boldsymbol{\beta} \leq 1} \boldsymbol{\beta}^t\boldsymbol{\omega} = \sqrt{\boldsymbol{\omega}^t\boldsymbol{\omega}}$$

and, in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{v}$, we have

$$\sup_{\boldsymbol{\alpha}^t\boldsymbol{A}\boldsymbol{\alpha} \leq 1} \boldsymbol{\alpha}^t\boldsymbol{v} = \sqrt{\boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{v}}$$

completing the proof. ■

Taking into account the previous proposition, we can define a dissimilarity index δ_{ij} through:

$$\delta_{ij} = \sqrt{(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)^t\widehat{\boldsymbol{\Xi}}_i^{-1}(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)} = \sqrt{\frac{1}{n_i}(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)^t\widehat{\boldsymbol{\Psi}}_i^{-1}(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)} \tag{8}$$

Thus, the matrix $\boldsymbol{\Delta} = (\delta_{ij})$ is a $p \times p$ non symmetric dissimilarity matrix between the populations $\Omega_1, \ldots, \Omega_p$.

It is possible to approximate $\boldsymbol{\Delta}$ by the symmetric matrix $\boldsymbol{D}$, using the next proposition:

Proposition 3.2 *The closest symmetric matrix to the dissimilarity matrix $\boldsymbol{\Delta}$, using the trace norm, is $\boldsymbol{D} = (\boldsymbol{\Delta} + \boldsymbol{\Delta}^t)/2$.*

*Proof:*Observe that with the scalar product defined, in the square matrix vector space, by $< \boldsymbol{A}, \boldsymbol{B} >= \mathrm{tr}(\boldsymbol{A}^t \boldsymbol{B})$ it is straightforward to prove that symmetric and antisymmetric matrices are orthogonal. On the other hand it is well known that any square matrix can be uniquely expressed as the sum of a symmetric plus an antisymmetric matrices, namely $\boldsymbol{\Delta} = (\boldsymbol{\Delta} + \boldsymbol{\Delta}^t)/2 + (\boldsymbol{\Delta} - \boldsymbol{\Delta}^t)/2$. Therefore $(\boldsymbol{\Delta} + \boldsymbol{\Delta}^t)/2$ is the orthogonal projection of $\boldsymbol{\Delta}$ into the subspace of symmetric matrices, concluding the proof. ■

4 Simulation results

In order to assess the performance of the dissimilarity we have used simulations where several samples from known distributions have been generated. Our choose has been for multivariate normal populations since in addition to Mahalanobis distance with the assumption of homogeneity of variances, we can compare our dissimilarity with Siegel distance, as introduced by Calvo et al.(2002), without assumptions about the covariance matrices. Of course our dissimilarity could also be applied to any data without restrictions about the parametric model.

A total of 9 trivariate normal populations have been simulated with three different mean vectors: $M_1 = (0,0,0), M_2 = (1,-1,1)$ and $M_3 = (1,2,3)$ and covariance matrices of the form

$$\Sigma_i = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

The parameters for simulations are summarized below

For each population a sample of size n=500 has been simulated. The most plain choice of statistics $\boldsymbol{T}$ in this situation are the mean values and the different

Pop.	1	2	3	4	5	6	7	8	9
mean	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
ρ	0	0	0	1/3	1/3	1/3	-1/3	-1/3	-1/3

Table 1: Parameters for simulations.

coefficients of covariance matrices, so each population is characterized by nine parameters $(\mu_1, \mu_2, \mu_3, \sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})$.

Once obtained the samples and computed the statistics we have obtained the matrix of coordinates for each population shown in Table 2.

	$\widehat{\mu_1}$	$\widehat{\mu_2}$	$\widehat{\mu_3}$	$\widehat{\sigma_{11}}$	$\widehat{\sigma_{12}}$	$\widehat{\sigma_{13}}$	$\widehat{\sigma_{22}}$	$\widehat{\sigma_{23}}$	$\widehat{\sigma_{33}}$
$\widehat{\boldsymbol{T}_1}$	-0.041	-0.018	0.005	0.991	0.019	0.010	1.069	0.019	0.880
$\widehat{\boldsymbol{T}_2}$	1.053	-0.960	1.024	0.991	0.069	-0.060	0.902	0.065	1.051
$\widehat{\boldsymbol{T}_3}$	0.982	1.964	2.961	1.024	0.053	-0.048	0.972	0.048	0.978
$\widehat{\boldsymbol{T}_4}$	-0.033	-0.024	-0.020	1.107	0.533	0.393	1.113	0.409	1.081
$\widehat{\boldsymbol{T}_5}$	0.978	-0.959	1.043	1.022	0.328	0.241	1.074	0.334	0.999
$\widehat{\boldsymbol{T}_6}$	0.948	1.953	2.952	1.095	0.355	0.303	0.964	0.298	0.938
$\widehat{\boldsymbol{T}_7}$	0.013	-0.080	-0.012	0.967	-0.271	-0.348	1.025	-0.429	1.079
$\widehat{\boldsymbol{T}_8}$	0.994	-0.995	0.914	0.951	-0.300	-0.301	0.983	-0.314	0.980
$\widehat{\boldsymbol{T}_9}$	1.069	1.899	3.005	1.119	-0.367	-0.388	0.974	-0.361	1.039

Table 2: Final coordinates of populations.

Unbiased estimations of covariance matrices $\widehat{\boldsymbol{\Psi}}_i$ for these statistics in each population have been obtained through 10000 bootstrap samples from each population sample. Finally using (8) we obtain the dissimilarity matrix D between the nine populations. As mentioned before and since we are working with multivariate normal distributions we can compare our dissimilarity not only with Mahalanobis distance under the assumption of a common covariance matrix, but also with Siegel distance between general multivariate normal populations as proposed by Calvo et al.(2002) and which is a lower bound of the Rao distance between multivariate normal distributions that has not been obtained explicitly until now. Siegel distances have been computed following (2.6) in Calvo et al. (2002).

Table 3 shows the dissimilarities computed based on relevant features (RFD) and Siegel distances between the 9 simulated populations. In Table 4 we compare RFD dissimilarity with the usual Mahalanobis distance after estimating a pooled covariance matrix.

	Pop.1	Pop.2	Pop.3	Pop.4	Pop.5	Pop.6	Pop.7	Pop.8	Pop.9
Pop.1	0.000	1.876	3.785	0.729	1.966	3.621	1.003	2.045	5.309
Pop.2	1.672	0.000	3.509	2.158	0.468	3.581	2.101	1.039	4.674
Pop.3	2.765	2.651	0.000	3.459	3.404	0.532	5.099	4.487	1.262
Pop.4	0.644	1.888	2.622	0.000	2.158	3.189	2.260	3.105	5.447
Pop.5	1.742	0.439	2.590	1.836	0.000	3.386	2.597	1.833	4.836
Pop.6	2.682	2.679	0.494	2.477	2.569	0.000	5.145	4.787	2.291
Pop.7	0.892	1.826	3.295	1.486	2.048	3.271	0.000	1.659	6.491
Pop.8	1.715	0.824	3.063	2.130	1.215	3.121	1.510	0.000	5.535
Pop.9	3.367	3.134	0.964	3.311	3.134	1.418	3.767	3.459	0.000

Table 3: RFD dissimilarities for the 9 simulated populations in the upper part and Siegel distances in the lower part.

	Pop.1	Pop.2	Pop.3	Pop.4	Pop.5	Pop.6	Pop.7	Pop.8	Pop.9
Pop.1	0.000	1.876	3.785	0.729	1.966	3.621	1.003	2.045	5.309
Pop.2	1.800	0.000	3.509	2.158	0.468	3.581	2.101	1.039	4.674
Pop.3	3.673	3.487	0.000	3.459	3.404	0.532	5.099	4.487	1.262
Pop.4	0.027	1.806	3.693	0.000	2.158	3.189	2.260	3.105	5.447
Pop.5	1.765	0.076	3.473	1.772	0.000	3.386	2.597	1.833	4.836
Pop.6	3.651	3.475	0.035	3.672	3.461	0.000	5.145	4.787	2.291
Pop.7	0.085	1.743	3.705	0.074	1.709	3.684	0.000	1.659	6.491
Pop.8	1.722	0.129	3.575	1.727	0.134	3.562	1.661	0.000	5.535
Pop.9	3.700	3.456	0.119	3.720	3.444	0.144	3.729	3.546	0.000

Table 4: RFD dissimilarities for the 9 simulated populations in the upper part and Mahalanobis distances in the lower part.

We present in Table 5 Spearman's rank correlation coefficient between the three computed measures.

	Dissimilarity (RFD)	Siegel
Siegel	0.98147	
Mahalanobis	0.87979	0.90862

Table 5: Spearman's rank correlation between the three computed measures

Finally in order to obtain a graphical representation of the similarities between the populations we have done a Multidimensional Scaling (MDS) based on Mahalanobis distance and on our dissimilarity (RFD). Fig. (1) shows the result obtained for the RFD dissimilarity in the big frame and the corresponding result for Mahalanobis distance in the upper right frame.

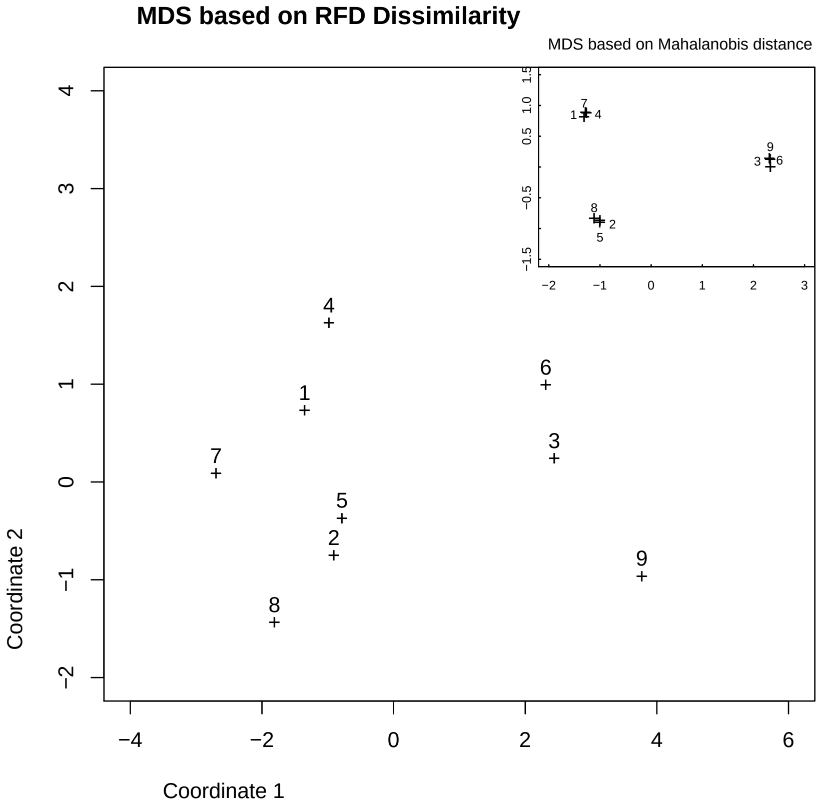

Figure 1: MDS based on RFD dissimilarity (big frame) and Mahalanobis distance (upper right frame) of the 9 simulated populations

5 Some applications to real data

5.1 Archaeological data

As an example, we consider the classical dataset on Egyptian skulls from five different epochs reported by Thomson and Randall-Maciver (1905) which has been widely used in so many different works. Data consists of four measurements of 150 male Egyptian skulls from five time periods: Early Predynastic, Late Predynastic, $12^{th} - 13^{th}$ dynasty, Ptolemaic and Roman. The four variables measured were: MB (Maximal Breadth of Skull), BH (Basibregmatic Height of Skull), BL (Basialveolar Length of Skull) and NH (Nasal Height of Skull).

In Table 6 we show the interdissimilarity matrix obtained by the method developed in the present paper with means and different coefficients of covariance matrices as coordinates. Figure 2a give us a representation in two dimensions through a Multidimensional Scaling of the dissimilarity matrix. Obviously the first axis can be interpreted as time, from most recent on the right to most distant on the left. The second axis has a not so clear interpretation but we observe that central values are occupied by native Egyptian populations before any foreign invasion, while late populations, where native Egyptian rulers had vanished long ago, diverge from the centerline in an almost perpendicular way. The result is quite different of that obtained by the standard Canonical Discriminant Analysis (CDA) as can be seen in Figure 2b.

	Early P.	Late P.	Dyn. 12-13	Ptolemaic	Roman
Early P.	0.000				
Late P.	2.045	0.000			
Dyn. 12-13	2.546	1.785	0.000		
Ptolemaic	3.595	2.934	2.244	0.000	
Roman	3.636	2.155	2.261	2.642	0.000

Table 6: RFD Dissimilarities matrix for the five populations of Egyptian skulls.

5.2 Academic achievement data

The data were collected from N = 382 university students on the number of GCE A-levels taken and the students' average grades (Mardia et al., 1979, p. 294). The students were grouped according to their final degree classification into seven

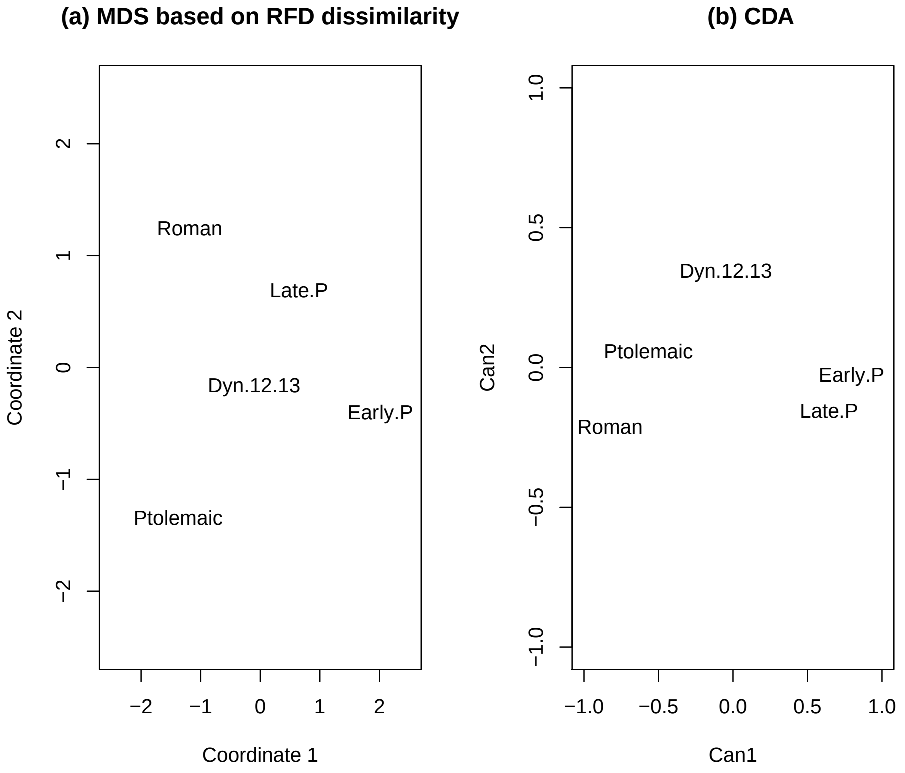

Figure 2: MDS based on RFD dissimilarity (a) and CDA (b) of the five populations of Egyptian skulls

groups: 1-I (students with degree class I), 2-II(i) (students with degree class II(i)), 3-II(ii) (students with degree class II(ii)), 4-III (students with degree class III), 5-Pass (students who obtained a 'Pass'), 6-3(4) (students who took four years over a three-year course) and 7-$\longrightarrow$ (students who left without completing the course). The average A-level grade obtained is a continuous variable (X_1) and the number of A-levels taken is categorized in two variables: X_2 (1 if two A-levels take, 0 otherwise) and X_3 (1 if four A-levels taken, 0 otherwise).

We have considered as statistics the mean and standard deviation of X_1 and the relative frequencies for X_2 and X_3. After obtaining the dissimilarity matrix the two-dimensional plot obtained through a classical MDS is shown in Figure 3. The minimum spanning tree of the dissimilarity matrix has been computed making use of the function `mst` from package **ape** developed in R by Paradis et al. (2009) and the tree has been plotted on the two-dimensional scaling solution. The plot shows no distorsion in the muldimensional scaling solution, so the MDS solution reflects accurately the dissimilarity.

Mardia's first canonical correlation solution gives the following scores to each of the degree results:

1	2	3	4	5	6	7
0	0.488	1.877	2.401	2.971	2.527	3.310

with $r = 0.4$ as first canonical correlation coefficient.

Mardia interpreted the scores as follows: "The scores for I, II(i), II(ii), III and Pass come out in the natural order, but they are not equally spaced. Moreover the 3(4) group comes between III and Pass, while → scores higher than Pass. Note the large gap between II(i) and II(ii)".

All conclusions are endorsed by our analysis as shown in Figure 3, so we think we have obtained a good representation of the seven populations based on the four chosen statistics.

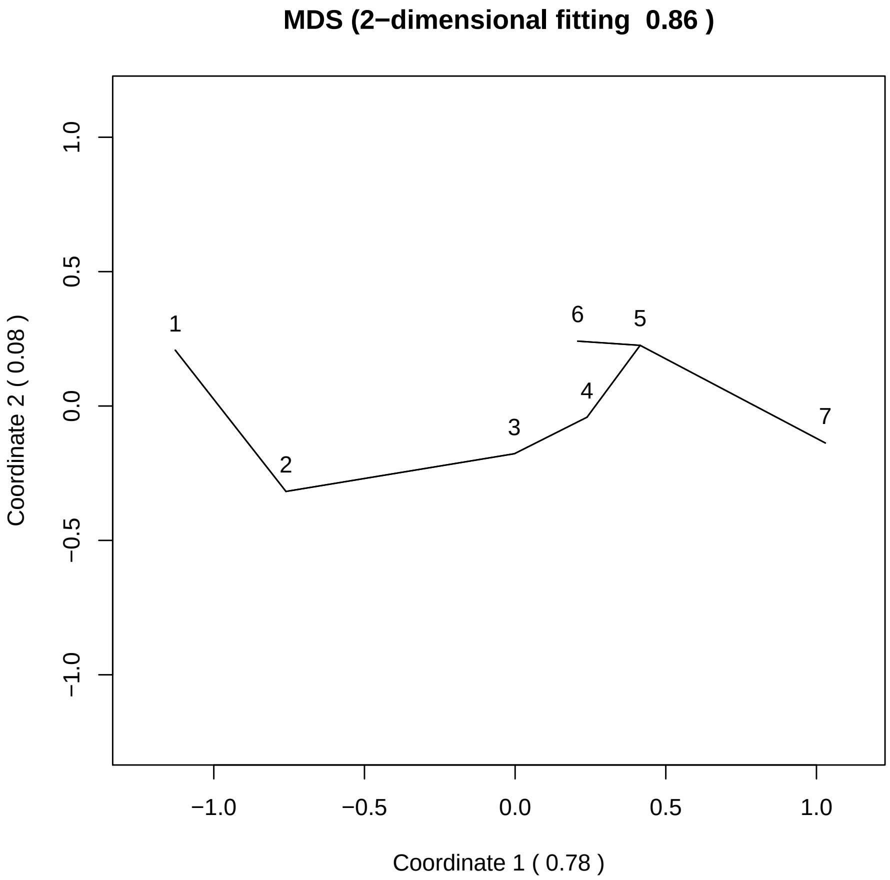

Figure 3: MDS based on RFD dissimilarity of the seven populations of university students

6 Conclusions

The proposed dissimilarity index has a reasonable behavior in all the studied cases, regardless of whether or not an adequate parametric statistical model is available for the data. The index is flexible enough to allow a wide applicability and, at the same time, maintaining a reasonable simplicity from a computational point of view.

7 References

Amari, S. (1985). *Differential-Geometrical Methods in Statistics.* Volume 28 of *Lecture notes in statistics,* Springer Verlag, New York.

Atkinson, C. and Mitchell, A.F.S. (1981). Rao's distance measure. *Sankhyā,* **43 A** , 345–365.

Bhattacharyya, A. (1942). On a measure of divergence between two multinomial populations. *Sankhyā,* **7**, 401–406.

Burbea J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae,* **4**, 347–378.

Burbea J. and Rao C.R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multivariate Analysis,* **12**, 575–596.

Calvo M., Villarroya A. and Oller J.M. (2002). A biplot method for multivariate normal populations with unequal covariance matrices. *Test,* **11**, 143-165

Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand Tour and Projection Pursuit. *Journal of Computational Statistical and Graphical Statistics,* **4** (3), 155–172.

Cubedo, M. and Oller, J.M. (2002). Hypothesis testing: a model selection approach. *Journal of Statistical Planning and Inference,* **108**, 3–21.

Friedman, J.H. (1987) Exploratory Projection Pursuit. *Journal of the American Statistical Association,* **82**, 249–266.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika,* **58**, 453–467.

Gower, J.C. and Hand, D.J. (1996). *BIPLOTS.* Chapman & Hall, London.

Greenacre, M.J (1993). *Correspondence Analysis in Practice,* London: Academic Press.

Huber, P.J. (1985). Projection Pursuit (with discussion). *The Annals of Statistics,* **13**, 435–525.

Lindsay, B.G., Markatou, M., Ray, S., Yang, K. and Chen, S. (2008). Quadratic distances on probabilities: a unified foundation*The Annals of Statistics* **36**(2) 2, 983–1006.

Mahalanobis (1936). On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2** (1) 49–55.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). em Multivariate Analysis, Probability and Mathematical Statistics. Academic Press, London

Miñarro A. and Oller J.M. (1992). Some remarks on the individual-score distance and its applications to statistical inference. *Qüestiió,* **16**, 43–57.

Oller J.M. and Corcuera J.M. (1995), Intrinsic Analysis of the Statistical Estimation, *The Annals of Statistics* **23**(2), 1562–1581.

Paradis, E., Strimmer, K., Claude J., Jobb, G., Opgen-Rhein, R., Dutheil, J., Noel, Y. and Bolker, B. (2009), **ape***: Analyses of Phylogenetics and Evolution,* URL http://CRAN.R-project.org/package=ape, R package version 2.3

Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.,* **37**, 81–91.

Thomson, A. and Randall-Maciver, R. (1905) *Ancient Races of the Thebaid*, Oxford University Press, Oxford.