QÜESTIIÓ, Vol 16, 1,2,3 pp. 43-57, 1992

SOME REMARKS ON THE INDIVIDUALS-SCORE DISTANCE AND ITS APPLICATIONS TO STATISTICAL INFERENCE

ANTONIO MIÑARRO and JOSEP M. OLLER

This paper is concerned with the study of some properties of the distance between statistical individuals based on representations on the dual tangent space of a parametric manifold representation of a statistical model. Explicit expressions for distances are obtained for well-known families of distributions. We have also considered applications of the distance to parameter estimation, testing statistical hypotheses and discriminant analysis.

AMS 1980 subjects classifications: 62A99, 62F03, 62F10.

Keywords: informative geometry, statistical individuals, distances between individuals, statistical inference.

–Dept. of Statistics. University of Barcelona. SPAIN.

–Article rebut el desembre de 1991.

–Acceptat el març de 1992.

1. INTRODUCTION

Given a measurable space $(\chi, \mathcal{A})$ let Π be the set of all probability measures on $(\chi, \mathcal{A})$. Let P_M be a subset of Π, $P_M \subset \Pi$. We may define a statistical model as a family of probability spaces

$$M = \{(\chi, \mathcal{A}, P) : P \in P_M\}$$

The statistical model is often defined through an n-dimensional C^∞ real and connected manifold (P_M, ϕ_{P_M}), where $P_M \subset \Pi$ and ϕ_{P_M} is a maximal C^∞ atlas on P_M. If the probability measures of P_M are dominated by a common reference measure μ, taking into account the Radon-Nikodym theorem we are able to represent (P_M, ϕ_{P_M}) by a manifold of measurable functions, or, more precisely, equivalent classes of measurable functions $(D^\lambda, \phi_{D^\lambda})$, where $D^\lambda \subset \mathcal{F}$, through the map

$$\begin{array}{rccc} \Phi_\lambda : & P_M & \longrightarrow & D^\lambda \\ & P & \longrightarrow & \lambda\left(\frac{dP}{d\mu}\right) \end{array}$$

where λ is a strictly monotonous real function, $\mathcal{F}$ stands for the set of all measurable functions on $(\chi, \mathcal{A})$ and the ϕ_{D^λ} is obtained by considering all the local charts of the form

$$(\Phi_\lambda(U), \xi \circ \Phi_\lambda^{-1}) \qquad \forall (U, \xi) \in \phi_{P_M}$$

We may call $(D^\lambda, \phi_{D^\lambda})$ the λ-representation of our statistical model. We shall restrict our study to function manifolds which satisfy some adequate regularity conditions such as for every local chart (U, ξ) and given any point $q \in U \subset D^\lambda$ of coordinates $\theta = \xi(q)$ the functions in x , $\frac{\partial\lambda(p(x;\theta))}{\partial\theta_i}$ $i = 1, \ldots, n$ are linearly independent, and belong to a convenient $\mathcal{L}^\alpha(p(\cdot;\theta)d\mu)$. Also, the partial derivatives $\partial/\partial\theta_i$ and the integration with respect to the measure $p(\cdot;\theta)d\mu$ can always be interchanged.

Let $\lambda(p(\cdot;\theta))$ be a point $q \in U \subset D^\lambda$ with coordinates $\theta = \xi(q)$, we denote by D_q^λ the tangent space to D^λ at the point q. D^λ may be represented by the vectorial space $E_\theta^\lambda \in \mathcal{L}^2(\phi(p(\cdot;\theta))d\mu$ defined by

$$E_\theta^\lambda = < \frac{\partial\lambda(p(\cdot;\theta))}{\partial\theta_1}, \ldots, \frac{\partial\lambda(p(\cdot;\theta))}{\partial\theta_n} >$$

through the map

$$\begin{array}{rccc} \Xi: & D^{\lambda}_{\theta} & \longrightarrow & E^{\lambda}_{\theta} \\ & X_q & \longrightarrow & \Xi(X_q) = \sum_{i=1}^{n} x_i \, \frac{\partial \lambda(p(\cdot;\theta))}{\partial \theta_i} \end{array}$$

where $x_1, \ldots, x_n$ are the coordinates of X_q relative to the usual basis of the tangent space D^{λ}_q corresponding to the local chart (U, ξ), and defined by $(\frac{\partial}{\partial \theta_i})_q f = D_i(f \circ \xi^{-1})(\xi(q)) \quad i = 1, \ldots n$, where the differentiation on the right is the usual in $\mathbb{R}^n$.

We may now define an inner product on D^{λ}_q by

$$\begin{aligned} < X_q, Y_q > \;\; &\equiv \;\; < \Xi(X_q), \Xi(Y_q) >= \\ & \sum_{i,j=1}^{n} x_i \, y_j \int_{\mathcal{X}} \frac{\partial \lambda(p(\cdot;\theta))}{\partial \theta_i} \frac{\partial \lambda(p(\cdot;\theta))}{\partial \theta_j} \phi(p(x;\theta)) \, d\mu(x) \end{aligned}$$

where if we require that this inner product has to be invariant under reference measure changes, we obtain the inner product matrix

$$g_{ij}(\theta) = k \mathrm{E} \left(\frac{\partial \ln p(\mathrm{x};\theta)}{\partial \theta_i} \frac{\partial \ln p(\mathrm{x};\theta)}{\partial \theta_j} \right) \qquad i, j = 1, \cdots, n$$

which is, up to a proportionality constant, the Fisher information matrix, Rao (1945), see Oller(1989) for more details.

From now on we shall consider $\lambda(x) = \ln(x)$ and we shall denote by E^{l}_{θ} the representation of the tangent space D^{l}_{θ}.

Let us consider the manifold (D^l, ϕ_{D^l}) representation of a statistical model. Given a point coordinated by θ there is a natural way to represent statistical individuals as linear forms on E^{l}_{θ}, through the map:

$$\text{(1.1)} \qquad \begin{array}{rccc} \delta: & \chi & \longrightarrow & E^{l*}_{\theta} \\ & \mathrm{x} & \longrightarrow & \delta(\mathrm{x}) = Y^* \end{array}$$

in such a way that $Y^*(Y) = Y(\mathrm{x}) \quad \forall Y \in E^{l}_{\theta}$. Every statistical individual can be represented as an element of E^{l*}_{θ} of coordinates

$$\mathrm{x} \longrightarrow (\partial_1 \ln \; p(\mathrm{x};\theta), \cdots, \partial_n \ln \; p(\mathrm{x};\theta))$$

where $\partial_i \ln \; p(\mathrm{x};\theta) \equiv \frac{\partial \ln p(\mathrm{x};\theta)}{\partial \theta_i}$.

Notice that on the whole manifold, whitout considering a given point θ, every statistical individual can be identifed with a first order covariant tensor field.

Through representation (1.1), we may define a pseudodistance on χ by

$$(1.2) \qquad d_\chi^2(\tilde{x},x) = d_{E_\theta^*}^2(\tilde{Y^*},Y^*) = < \tilde{Y^*} - Y^*, \tilde{Y^*} - Y^* >_{E_\theta^*} =$$

$$= (\partial_\theta \ln p(\tilde{x};\theta) - \partial_\theta \ln p(x;\theta))' G^{-1}(\theta)(\partial_\theta \ln p(\tilde{x};\theta) - \partial_\theta \ln p(x;\theta))$$

where

$$\partial_\theta \ln p(\tilde{x};\theta) \equiv \left(\frac{\partial \ln p(\tilde{x};\theta)}{\partial\theta_1}, \dots, \frac{\partial \ln p(\tilde{x};\theta)}{\partial\theta_n}\right)'$$

For this pseudodistance to be a proper distace we may define an equivalence relation on χ as $x \sim \tilde{x} \Longleftrightarrow d_\chi^2(x,\tilde{x}) = 0$, and extending the pseudodistance d_χ to the equivalence classes of the quotien set $\chi/\sim$.

Notice that the defined distance is not an intrinsic distance between individuals in the sense of Rao(1982), since it depends on the statistical model and on the population to which individuals belong, characterized by the coordinates θ . Additional details may be found in Cuadras (1989a,b), Oller (1989) and Miñarro (1991).

2. SOME PROPERTIES OF THE SCORE DISTANCE BETWEEN STATISTICAL INDIVIDUALS

Proposition 2.1

The distance (1.2) is invariant under reference measure changes $\mu \rightarrow \nu$ such that $\mu << \nu$.

Proof:

It follows from the invariance of Fisher information matrix and of the coordinates $(\partial_1 \ln\ p(x/\theta), \cdots, \partial_n \ln\ p(x/\theta))$ ■

Proposition 2.2

Let T be a measurable map from $(\chi, \mathcal{A})$ into $(\chi', \mathcal{A}')$, with $P'(B) = \int_B g(\mathbf{t};\theta)\, d\nu(t)$ $\forall B \in \mathcal{A}'$. Where $P' = PT^{-1}$ and ν is a σ-finite reference measure on $\mathcal{A}'$. If T is a sufficient statistic then:

$$d^2_\chi(\tilde{\mathbf{x}}, \mathbf{x}) = d^2_{\chi'}(\tilde{\mathbf{t}}, \mathbf{t}) \tag{2.1}$$

where $\mathbf{t} = T(\mathbf{x})$ and $\tilde{\mathbf{t}} = T(\tilde{\mathbf{x}})$.

Proof:

It follows from the invariance of Fisher information matrix and from the Neyman-Fisher factorization criterion, since then:

$$p(\mathbf{x};\theta) = g(\mathbf{t};\theta)\, h(\mathbf{x}) \tag{2.2}$$

and from (1.2) it follows immediately (2.1) ■

Proposition 2.3

The distance (1.2) is not decreasing if the number of parameters increases.

Proof:

(1.2) may be represented by the quadratic form $u'G^{-1}u$ where $u = (\partial_\theta \ln p(\tilde{\mathbf{x}};\theta) - \partial_\theta \ln p(\mathbf{x};\theta))$. Let us remember that G is a symmetric and positive definite matrix. Let us consider now $u = (u_1, u_2)$ and

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

The proposition will be proved if we show that $u'G^{-1}u - u_1'G_{11}^{-1}u_1 \geq 0$, where

$$G^{-1} = \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix}$$

Let us break down $u = v + w$ where $v = (G_{11}G_{11}^{-1}u_1, G_{21}G_{11}^{-1}u_1)'$ and $w = (0, u_2 - G_{21}G_{11}^{-1}u_1)'$, then $u'G^{-1}u = w'G^{-1}w + v'G^{-1}v + 2v'G^{-1}w$. Immediately $w'G^{-1}w \geq 0$, since G^{-1} is also positive definite, and we can easily see that $v'G^{-1}v = u_1'G_{11}^{-1}u_1$ and also that $v'G^{-1}w = 0$. So $u'G^{-1}u - u_1'G_{11}^{-1}u_1 = w'G^{-1}w \geq 0$ ■

Proposition 2.4

Let $\theta_1^1, \ldots, \theta_{n_1}^1, \theta_1^2, \ldots, \theta_{n_2}^2, \ldots, \theta_1^k, \ldots, \theta_{n_k}^k$ be the parameters of the density corresponding to a statistical model, where $n_1 + \cdots + n_k = n$. If the Fisher information matrix is of the form

$$G(\boldsymbol{\theta}) = \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{pmatrix}$$

where A_i are $n_i \times n_i$ matrices. The squared distance between two statistical individuals in the dual tangent space to the manifold coordinated by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is equal to the sum of the k squared distances between the individuals in the k dual tangent spaces of the k n_i-dimensional manifolds coordinated by $\boldsymbol{\theta}_i = (\theta_1^i, \ldots, \theta_{n_i}^i)$ $i = 1, \ldots, k$. As a particular case if $p_1(\mathbf{x}_1; \boldsymbol{\theta}_1), \ldots, p_k(\mathbf{x}_k; \boldsymbol{\theta}_k)$ are representations of k independent statistical models, then the distance between two individuals on the model $p_{1\cdots k}(\mathbf{x}_1, \ldots, \mathbf{x}_k; \boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_k)$ is obtained from

$$(2.3) \quad d_{1\cdots k}^2((\mathbf{x}_1, \ldots, \mathbf{x}_k), (\mathbf{y}_1, \ldots, \mathbf{y}_k)) = d_1^2(\mathbf{x}_1, \mathbf{y}_1) + \cdots + d_k^2(\mathbf{x}_k, \mathbf{y}_k)$$

where $d_i^2(\mathbf{x}_i, \mathbf{y}_i)$ is the distance on the model $p_i(\mathbf{x}_i; \boldsymbol{\theta}_i)$.

Proof:

It follows from definition (1.2) and from the fact that the inverse Fisher information matrix is of the form

$$G^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} A_1^{-1} & 0 & \cdots & 0 \\ 0 & A_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k^{-1} \end{pmatrix}$$

■

3. DISTANCES FOR SOME WELL-KNOWN DISTRIBUTIONS

In this section we show the resulting expressions of (1.2) for some well-known familes of distributions. The two individuals are considered to be samples of size m, $\mathrm{x} = (x_1, \dots x_m)$ and $\mathrm{y} = (y_1, \dots y_m)$.

3.1 One-parameter distributions

Distribution	$d^2(\mathrm{x}\ ,\ \mathrm{y}\ )$
Poisson (Mean λ)	$\frac{m}{\lambda}(\bar{\mathrm{x}} - \bar{\mathrm{y}})^2$
Weibull (Mean $\frac{\Gamma(1+\frac{1}{r})}{\lambda}$ (r known))	$\frac{\lambda^2}{m}\,(\sum_{i=1}^{m}(x_i^r - y_i^r))^2$
Gamma (Mean $\frac{p}{\alpha}$ (p known))	$\frac{m\ \alpha^2}{p}(\bar{\mathrm{x}} - \bar{\mathrm{y}})^2$
Exponential (Mean $\frac{1}{\alpha}$)	$m\ \alpha^2(\bar{\mathrm{x}} - \bar{\mathrm{y}})^2$
Binomial (Mean mp)	$\frac{(x-y)^2}{m\,p\,(1-p)}$
Negative Binomial (Mean $\frac{k\,(1-p)}{p}$ (k known))	$\frac{(x-y)^2\,p^2}{k\,(1-p)}$
N $(\mu, \sigma_0\ )$ (σ_0 known)	$\frac{m}{\sigma_0^2}\,(\bar{\mathrm{x}} - \bar{\mathrm{y}})^2$
N $(\mu_0, \sigma\ )$ (μ_0 known)	$\frac{m}{2\,\sigma^4}(S^2(\mathrm{x}) - S^2(\mathrm{y}))^2$

where $\bar{\mathrm{x}} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $S^2(\mathrm{x}\ ) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_0)^2$.

Note that most of the cases above are particular cases of the exponential family $p(\mathrm{x};\theta) = \exp\{Q(\theta)T(\mathrm{x}) + D(\theta) + S(\mathrm{x})\}$ where the distance takes the form

$$d^2(\mathrm{x},\mathrm{y}) = \{Q''(\theta)\mathrm{E}[T(\mathrm{x})] - D''(\theta)\}^{-1}\{Q'(\theta)\}^2\,\{T(\mathrm{x}) - T(\mathrm{y})\}^2.$$

3.2 Multiparameter distributions

We now consider some multiparameter examples.

- **Multinomial distribution**

$$p(x_1,\cdots,x_{n+1};p_1,\cdots,p_n)=\frac{m!}{x_1!\cdots x_{n+1}!}(p_1)^{x_1}\cdots(p_{n+1})^{x_{n+1}}$$

The squared distance is:

$$d^2(\mathrm{x},\mathrm{y})=\frac{1}{m}\sum_{i=1}^{n+1}\frac{(x_i-y_i)^2}{p_i} \tag{3.1}$$

- **Negative Multinomial distribution**

$$p(x_1,\cdots,x_n;p_1,\cdots,p_n)=\frac{(x_1+\cdots+x_n+r-1)!}{x_1!\cdots x_{n+1}!(r-1)!}(p_1)^{x_1}\cdots$$
$$\cdots(p_n)^{x_n}(1-p_1-\cdots-p_n)^r$$

The squared distance is:

$$d^2(\mathrm{x},\mathrm{y})=\frac{1-p_1-\cdots-p_n}{r}\left(\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{p_i}-(\sum_{i=1}^{n}(x_i-y_i))^2\right) \tag{3.2}$$

- **Distributions of the form $p(x_1,\cdots,x_m;\mu,\beta)=\prod_{i=1}^m\frac{1}{\beta}F\left[\left(\frac{x_i-\mu}{\beta}\right)^2\right]$.**

Where $\mu\in\mathbb{R}$, $\beta>0$ and $F:\mathbb{R}^+\cup\{0\}\longrightarrow\mathbb{R}^+\cup\{0\}$ with

$$\int_{-\infty}^{\infty}F(u^2)\,du\;=\;1$$

Assuming that

$$a=4\int_0^{\infty}t^{1/2}(\mathcal{L}F)^2(t)F(t)dt<\infty$$

and

$$b=\int_{-\infty}^{\infty}(1+2(\mathcal{L}F)(u^2))u^2)^2f(u^2)du<\infty$$

where $\mathcal{L}F \equiv \frac{F'}{F}$. The squared distance is:

$$d^2(\mathrm{x},\mathrm{y}){=}\frac{4}{mb}\left[\sum_{i=1}^{m}(\mathcal{L}F)((\frac{x_i-\mu}{\beta})^2)(\frac{x_i-\mu}{\beta})-\sum_{i=1}^{m}(\mathcal{L}F)((\frac{y_i-\mu}{\beta})^2)(\frac{y_i-\mu}{\beta})\right]^2+$$
$$+\frac{4}{ma}\left[\sum_{i=1}^{m}(\mathcal{L}F)((\frac{x_i-\mu}{\beta})^2)(\frac{x_i-\mu}{\beta})^2-\sum_{i=1}^{m}(\mathcal{L}F)((\frac{y_i-\mu}{\beta})^2)(\frac{y_i-\mu}{\beta})^2\right]^2 \tag{3.3}$$

Some results on this family may be found in Mitchell (1988). Some of the following are particular cases of the mentioned above.

- **Normal distribution** $N(\mu,\sigma)$.

The squared distance is:

$$d^2(\mathrm{x},\mathrm{y}) = \frac{m}{2\,\sigma^4}\left(2\,\sigma^2(\bar{\mathrm{x}}-\bar{\mathrm{y}})^2+(S^2(\mathrm{x})-S^2(\mathrm{y}))^2\right) \tag{3.4}$$

where $\bar{\mathrm{x}} = \frac{1}{m}\sum_{i=1}^{m} x_i$ y $S^2(\mathrm{x}) = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_0)^2$.

- **Logistic distribution**

$$p(x_1,\cdots,x_m;\alpha,\beta) = \prod_{i=1}^{m}\frac{1}{4\beta}\mathrm{sech}^2\left(\frac{x_i-\alpha}{2\beta}\right)$$

The squared distance is:

$$d^2(\mathrm{x},\mathrm{y}) = 3\left(\sum_{i=1}^{m}(T(x_i)-T(y_i))\right)^2+$$

$$\frac{9}{\beta^2(\pi^2+3)}\left(\sum_{i=1}^{m}(x_iT(x_i)-y_iT(y_i)+\alpha(T(x_i)-T(y_i)))\right)^2 \tag{3.5}$$

where $T(x_i) = \tanh\left(\frac{x_i-\alpha}{2\beta}\right)$.

- **Wald distribution**

$$p(x_1,\cdots,x_m;\lambda,\mu) = \prod_{i=1}^{m}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2}\exp\left(-\left(\frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)\right)$$

The squared distance is:

$$(3.6)\quad d^2(\mathrm{x},\mathrm{y}) = \frac{\lambda^2}{2m\mu^4}\left(m(\bar{\mathrm{x}}-\bar{\mathrm{y}}) + \mu^2\sum_{i=1}^{m}\frac{x_i - y_i}{x_i y_i}\right)^2 + \frac{\lambda m}{\mu^3}(\bar{\mathrm{x}}-\bar{\mathrm{y}})^2$$

where $\bar{\mathrm{x}} = \frac{1}{m}\sum_{i=1}^{m} x_i$.

- **Multivariate normal distribution** $N(M,\Sigma_0)$ (Σ_0 known).

  The squared distance between two individuals X and Y is:

$$(3.7)\qquad d^2(X,Y) = (X-Y)'\Sigma_0^{-1}(X-Y)$$

Notice that (3.7) is identical to Mahalanobis distance, although the latter between populations, Mahalanobis (1936).

- **Multivariate normal distribution** $N(M_0,\Sigma)$ (M_0 known).

  The squared distance between X and Y is now:

$$\begin{aligned} d^2(X,Y) &= \frac{1}{2}((\Delta Y'\Sigma^{-1}\Delta Y)^2 + \\ (3.8)\qquad & (\Delta X'\Sigma^{-1}\Delta X)^2 - 2(\Delta X'\Sigma^{-1}\Delta Y)^2) \end{aligned}$$

where $\Delta X = X - M_0$.

- **Multivariate normal distribution** $N(M,\Sigma)$.

  The squared distance between X and Y is:

$$\begin{aligned} d^2(X,Y) &= (X-Y)'\Sigma_0^{-1}(X-Y) + \frac{1}{2}((\Delta Y'\Sigma^{-1}\Delta Y)^2 + \\ (3.9)\qquad & (\Delta X'\Sigma^{-1}\Delta X)^2 - 2(\Delta X'\Sigma^{-1}\Delta Y)^2) \end{aligned}$$

that can also be written in the alternative form:

$$\begin{aligned} d^2(X,Y) &= (X-Y)'\Sigma_0^{-1}(X-Y) + \\ (3.10)\qquad & \frac{1}{2}((\Delta X' + \Delta Y')W(\Delta X - \Delta Y)) \end{aligned}$$

where $\Delta X = X - M_0$ and $W = \Sigma^{-1}(\Delta X\Delta X' - \Delta Y\Delta Y')\Sigma^{-1}$.

4. SOME APPLICATIONS TO STATISTICAL INFERENCE

4.1 Parameter estimation

We may define an estimation procedure based on geometric considerations by using the distance between statistical individuals. Given a sample, from which we wish to estimate the coordinates of the density, we require that the expected distance between our actual sample and any other possible sample from the population be minimal. That is, we wish determine the value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, provided that exists, which minimizes the function:

$$\mathrm{E}_{\boldsymbol{\theta}}(d^2(\tilde{\mathrm{x}},\mathrm{x})) = \int_{\mathcal{X}} d^2(\tilde{\mathrm{x}},\mathrm{x})p(\mathrm{x};\boldsymbol{\theta})d\mu(\mathrm{x}) \tag{1.1}$$

If this value exists, it is the *minimum expected squared distance estimator* (MESD) of $\boldsymbol{\theta}$. Taking into account definition (1.2) and that $\mathrm{E}_{\boldsymbol{\theta}}(\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta})) = 0$ and

$$\mathrm{E}_{\boldsymbol{\theta}}((\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))\,(\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))') = G(\boldsymbol{\theta})$$

we obtain the following:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}}(d^2(\tilde{\mathrm{x}},\mathrm{x})) \; = \; & (\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta}))'\, G^{-1}(\boldsymbol{\theta})\,(\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta})) - \\
& 2(\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta}))'\, G^{-1}(\boldsymbol{\theta})\, \mathrm{E}_{\boldsymbol{\theta}}(\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta})) + \\
& \mathrm{E}_{\boldsymbol{\theta}}((\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))'\, G^{-1}(\boldsymbol{\theta})\,(\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))) = \\
& (\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta}))'\, G^{-1}(\boldsymbol{\theta})\,(\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta})) + \\
& \mathrm{tr}(G^{-1}(\boldsymbol{\theta})\, \mathrm{E}_{\boldsymbol{\theta}}((\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))\,(\partial_{\boldsymbol{\theta}} \ln p(\mathrm{x};\boldsymbol{\theta}))')) = \\
& (\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta}))'\, G^{-1}(\boldsymbol{\theta})\,(\partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta})) + n
\end{aligned} \tag{1.2}$$

which is the sum of the number of parameters and the squared norm of the vector of coordiantes of the sample $\tilde{\mathrm{x}}$ in the dual tangent space.

As we can see any consistent root of the likelihood equations defines the MESD estimator, since then

$$\| \, \partial_{\boldsymbol{\theta}} \ln p(\tilde{\mathrm{x}};\boldsymbol{\theta}) \, \| = 0 \tag{1.3}$$

However, note that the MESD dos not coincides necessarly with the maximum likelihood estimator (MLE) since we do not require the solution to be a maximum for the likelihood. If we consider, for example, the estimation of the parameters of a normal distribution from a sample of size one, the MLE leads to $\hat{\mu} = \tilde{x}$ and $\hat{\sigma} = 0$. If ,on the other hand, we develope the expected squared distance we obtain:

$$(1.4) \qquad \mathrm{E}\left(d^2(\tilde{x},x)\right) = \left(\frac{\tilde{x}-\mu}{\sigma}\right)^2 + \frac{1}{8}\left(\left(\frac{\tilde{x}-\mu}{\sigma}\right)^2 - 1\right)^2 + 2$$

which reaches the minimum at $\hat{\mu} = \tilde{x}$ being σ^2 arbitrary. We consider more reasonable the result provided by MESD.

This is not the first time that solutions of likelihood equations not necessarly an absolute maximum are used to obtain estimates. Duda and Hart (1973) consider a mixture of normal distributions with unknown parameters, the maximum likelihood solution is singular, but they obtain reasonable estimates usign the greatest relative maximum of the likelihood function.

4.2 Testing Statistical Hypotheses

Another possible application of distance between statistical individuals is concerned with testing parametric statistical hypotheses. Let us consider a hypothesis testing defined by

$$\begin{aligned} H_0 &: \ \theta \in \Theta_H \\ H_1 &: \ \theta \in \Theta \end{aligned}$$

where Θ_H is a restriction of the original parameter space defined by the null hypothesis.

Given a sample $\tilde{\mathrm{x}} \in \chi$, let us consider the following statistics:

$$(1.5) \qquad \inf_{\theta \in \Theta_H} \mathrm{E}_\theta(d^2(\tilde{\mathrm{x}},\mathrm{x})) = \mathrm{E}_{\hat{\theta}_1}(d^2(\tilde{\mathrm{x}},\mathrm{x})) = \mathrm{E}_1(\tilde{\mathrm{x}})$$

$$(1.6) \qquad \inf_{\theta \in \Theta} \mathrm{E}_\theta(d^2(\tilde{\mathrm{x}},\mathrm{x})) = \mathrm{E}_{\hat{\theta}_2}(d^2(\tilde{\mathrm{x}},\mathrm{x})) = \mathrm{E}_2(\tilde{\mathrm{x}})$$

Provided that (4.5) and (4.6) exist, we may define a statistical test for solving the previous hypothesis testing by considering a critical region of the form

$$(1.7) \qquad W = \{(\tilde{x}_1, \cdots, \tilde{x}_m) : \frac{\mathrm{E}_1(\tilde{\mathrm{x}})}{\mathrm{E}_2(\tilde{\mathrm{x}})} \geq \lambda_\epsilon\}$$

where ϵ is the significance level of the test and the constant λ_ϵ is chosen in such a way that $P(\lambda \geq \lambda_\epsilon / H_0) \leq \epsilon$.

For a simple hypothesis $H_0 : \theta = \theta_0$ and taking into account that under the null hypothesis

$$(1.8) \qquad \mathrm{X}^2 = (\partial_{\theta} \ln p(\tilde{\mathrm{x}};\theta_0))' G^{-1}(\theta_0)(\partial_{\theta} \ln p(\tilde{\mathrm{x}};\vec{\theta_0})) \xrightarrow{\mathcal{L}} Y \sim \chi^2_n$$

that is, X^2 will be asymptotically distributed as a chi-squared random variable with n degrees of freedom.

From (4.2), the critical region (4.7) may now be expressed as

$$(1.9) \qquad W = \{\tilde{\mathrm{x}} : (\partial_{\theta} \ln p(\tilde{\mathrm{x}};\theta_0))' G^{-1}(\theta_0)(\partial_{\theta} \ln p(\tilde{\mathrm{x}};\vec{\theta_0})) \geq c_{\epsilon}\}$$

and now, c_ϵ can be easily determined.

The test obtained in (4.9) coincides with Lagrange multiplier test, Aitchison and Silvey (1958), or score tests, Tarone (1988), first considered by Fisher (1935). In general, score tests are asymptotically equivalent to Wald tests based on maximum likelihood estimators and to likelihood ratio tests.

As an example let us consider the multinomial distribution, as defined in section 3, with the hypothesis test

$$H_0 : \ \mathrm{p} = \mathrm{p}_0$$
$$H_1 : \ \mathrm{p} \neq \mathrm{p}_0$$

It is not difficult to see that the test defined in (4.9) takes the form

$$(1.10) \qquad W = \{\tilde{\mathrm{x}} : \sum_{i=1}^{n+1} \frac{x_i^2}{m p_i} - m \geq c_{\epsilon}\}$$

and under the null hypothesis, the statistic $\sum_{i=1}^{n+1} \frac{x_i^2}{m p_i} - m$ is asymptotically distributed as a chi-squared distribution with n degrees of freedom.

Note that the statistic obtained is the well known Pearson's chi-squared.

4.3 Discriminant Analysis

Let $\tilde{\mathrm{x}}$ be an observation to classify between a set of populations $\Pi_1, \ldots, \Pi_k$. We may define the following discriminant function

$$(1.11) \qquad f_i(\tilde{\mathrm{x}}) = \mathrm{E}_{\Pi_i}(d^2(\tilde{\mathrm{x}}, \mathbf{x}))$$

where $\mathbf{x}$ is any possible sample from the population Π_i. The decision rule is to assign $\tilde{\mathrm{x}}$ to Π_i if

$$(1.12) \qquad f_i(\tilde{\mathrm{x}}) = \min(f_1(\tilde{\mathrm{x}}), \ldots, f_k(\tilde{\mathrm{x}}))$$

Another possible discriminant function proposed by Cuadras (1989a) is

$$f_i(\tilde{\mathbf{x}}) = d^2_{\Pi_i}(\tilde{\mathbf{x}}, 0)$$

which may be considered as the squared distance between $\mathbf{x}$ and the mean individual of the population, since $\mathrm{E}(\partial_i \ln p(\mathbf{x}; \boldsymbol{\theta})) = 0$. Some other comments can be found in Cuadras (1989a) and Sanchez (1989).

BIBLIOGRAPHY

[1] **Aitchison, J. and Silvey, S.D.** (1958). "Maximum likelihood estimation of parameters subject to restraints." *Ann. Math. Statist.*, vol. **29**, pp. 813–828.

[2] **Cuadras, C.M.** (1989a)."Distance analysis in discrimination and classification using both continuous and categorical variables." *Statistical Data Analysis and Inference,* (Y. Dodge, Ed.) North-Holland Pu. Co., Amsterdam.

[3] **Cuadras, C.M.** (1989b)."Distancias estadísticas entre individuos y poblaciones con variables mixtas." *Actas XVIII Reunión Nac. Estad. e I. Oper., Univ. Santiago de Compostela,* 143–148.

[4] **Duda, R.O. and Hart, P.E.** (1973). "Pattern Classification and Scene Analysis." *John Wiley and sons,* New York.

[5] **Fisher, R.A.** (1935). "The Fiducial Argument in Statistical Inference." *Annals of Eugenics,* vol. **6**.

[6] **Mahalanobis, P.C.** (1936). "On the Generalized Distance in Statistics." *Proc. Natl. Inst. Sci. India.,* vol **2**, pp. 49–55.

[7] **Miñarro, A.** (1991). *Aspectos Geométricos de las Poblaciones y los Individuos Estadísticos.* Tesis Doctoral. Universitat de Barcelona.

[8] **Mitchell, A.F.S.** (1988). "Statistical Manifolds of Univariate Elliptic Distributions." *International Statistical Review,* **56**, 1, 1–16.

[9] **Oller, J.M.** (1989). "Some geometrical aspects of data analysis and statistics." in *Statistical Data Analysis and Inference,* (Y. Dodge, Ed.) North-Holland Pu. Co., Amsterdam.

[10] **Rao, C.R.** (1945). "Information and accuracy attainable in the estimation of parameters." *Bull. Calcutta Math. Soc.,* vol. **37**, pp. 81–97.

[11] **Rao, C.R.** (1982). "Diversity and Dissimilarity Coefficients: A Unified Approach." *J. Theoretical Population Biology,* vol. **21**, pp. 24–43.

[12] Sanchez, P. (1989). "Funciones discriminantes basadas en distancias estadísticas." *Actas XVIII Reunión Nacional de Estadística e Investigación Operativa,* Univ. Santiago de Compostela, pp. 468–472.

[13] Tarone, R.E. (1988). "Score Statistics." en *Encyclopedia of Statistical Sciences,* vol 8. (S. Kotz and N.L. Eds.), John Wiley, New York.