UTILIZACIÓN DE MÉTRICAS RIEMANNIANAS EN ANALISIS DE DATOS MULTIDIMENSIONALES Y SU APLICACIÓN A LA BIOLOGÍA

JOSE Mª OLLER SALA

BARCELONA, 25 de NOVIEMBRE de 1982.

donde c_1 es una constante de integración. Nótese que $c_1 > 0$ si y sólo si $\mu_B > \mu_A$, $c_1 < 0$ si y sólo si $\mu_B < \mu_A$. Para el caso $\frac{d\mu}{ds} = 0$, μ = cte. es una solución de la primera ecuación de (28) y puede considerarse incluida en (30), haciendo $c_1=0$. Entonces $\mu=\mu_A=\mu_B$.

Sustituyendo (30) en la segunda de (28), obtenemos:

$$\frac{d^2\sigma}{ds^2} - \frac{1}{\sigma}\left(\frac{d\sigma}{ds}\right)^2 + \frac{c_1^2}{2}\sigma^3 = 0 \qquad (31)$$

Si efectuamos el cambio:

$$\frac{d\sigma}{ds} = \pm\sqrt{y} \qquad (32)$$

donde el signo de $\sqrt{y}$ depende del signo de $\frac{d\sigma}{ds}$, resulta:

$$\frac{d^2\sigma}{ds} = \frac{1}{\pm 2\sqrt{y}}\frac{dy}{d\sigma}\frac{d\sigma}{ds} = \frac{1}{2}\frac{dy}{d\sigma} \qquad (33)$$

por tanto obtenemos:

$$\frac{dy}{d\sigma} - \frac{2}{\sigma}y + c_1^2\sigma^3 = 0 \qquad (34)$$

ecuación lineal cuya integral general viene dada por:

$$y = e^{\int 2/\sigma\, d\sigma}\left(-c_1^2\int\sigma^3 e^{-\int 2/\sigma\, d\sigma}\, d\sigma + c_2\right) = \sigma^2\left(c_2 - \frac{c_1^2}{2}\sigma^2\right) \qquad (35)$$

Recordemos ahora que el sistema debe resolverse con la condición adicional que el vector tangente a la curva $\left(\frac{d\mu}{ds}, \frac{d\sigma}{ds}\right)$ sea uni-

tario en un punto, (28), y por tanto, en todos los puntos de la curva (por transporte paralelo). De (28),(30) y (32) se sigue que:

$$c_2 = \frac{1}{2} \tag{36}$$

por tanto (35) se transforma, teniendo en cuenta (32), en:

$$\frac{d\sigma}{ds} = \pm\sigma \sqrt{\frac{1}{2} - \frac{c_1^2}{2}\,\sigma^2} \tag{37}$$

Vamos a resolver (37) distinguiendo varios casos.

<u>Caso a</u> $c_1=0$, ocurre como hemos visto cuando $\mu_A=\mu_B$, resulta:

$$\frac{d\sigma}{ds} = \pm \frac{\sigma}{\sqrt{2}} \tag{38}$$

puesto que $\sigma > 0$ eligiremos la solución positiva si $\sigma_B > \sigma_A$ y la negativa en caso contrario:

$$\int_{\sigma_A}^{\sigma_B} \pm \frac{d\sigma}{\sigma} = \int_{o}^{s_{AB}} \frac{ds}{\sqrt{2}} \tag{39}$$

por tanto:

$$s_{AB} = \sqrt{2} \left| \ln \frac{\sigma_B}{\sigma_A} \right| \tag{40}$$

y las ecuaciones de las geodésicas que pasan por el punto (μ_A,σ_A) son:

$$\begin{aligned} \sigma &= \sigma_A\, e^{\pm s/\sqrt{2}} \\ \mu &= \mu_A \end{aligned} \tag{41}$$

eligiendo el signo en la primera, positivo si $\sigma_B > \sigma_A$ y negativo si $\sigma_B < \sigma_A$. Es claro que siempre existe una geodésica que une (μ_A,σ_A) con otro punto de la forma (μ_B,σ_B), con $\mu_B = \mu_A$.

Caso b Supongamos que $\mu_A \neq \mu_B$ entonces $c_1 \neq 0$ y supongamos además que $\frac{d\sigma}{ds} \neq 0$ en todos los puntos de la curva comprendidos entre (μ_A,σ_A) y (μ_B,σ_B). Entonces $\frac{d\sigma}{ds} > 0$ si $\sigma_B > \sigma_A$ y $\frac{d\sigma}{ds} < 0$ si $\sigma_B < \sigma_A$, habiendo de elegir el signo en (37) de acuerdo con estas consideraciones. Resulta:

$$\left| \int_{\sigma_A}^{\sigma_B} \frac{d\sigma}{\sigma\sqrt{1-c_1^2\sigma^2}} \right| = \int_0^{s_{AB}} \frac{ds}{\sqrt{2}} \tag{42}$$

y por tanto:

$$s_{AB} = \sqrt{2} \; | \; \text{argsech}(|c_1|\sigma_A) - \text{argsech}(|c_1|\sigma_B) | \tag{43}$$

Podemos expresar σ en función de s, de forma que para $s=0$ la curva pase por σ_A:

$$\sigma = \frac{1}{|c_1|} \text{ sech } (\pm \frac{s}{\sqrt{2}} + \text{argsech } (|c_1| \sigma_A)) \tag{44}$$

eligiendo el signo negativo si queremos que $\sigma > \sigma_A$ y positivo en caso contrario.

En cuanto a μ, teniendo en cuenta (30), podemos escribir:

$$\frac{d\mu}{ds} = \frac{1}{c_1} \text{ sech}^2 \; (\pm \frac{s}{\sqrt{2}} + \text{argsech } (|c_1|\sigma_A)) \tag{45}$$

y por tanto:

$$\int_{\mu_A}^{\mu_B} d\mu = \frac{1}{c_1} \int_{o}^{s_{AB}} \operatorname{sech}^2\left(\pm\frac{s}{\sqrt{2}} + \operatorname{argsech}(|c_1|\sigma_A)\right) ds \qquad (46)$$

Teniendo en cuenta que:

$$\operatorname{tgh}\ (\operatorname{arg\ sech}\ (|c_1|\ \sigma)) = \sqrt{1 - c_1^2\ \sigma^2} \qquad (47)$$

y las consideraciones efectuadas acerca del signo en la expresión (44), al integrar (46) obtenemos:

$$\mu_B - \mu_A = \frac{\sqrt{2}}{c_1} \left| \sqrt{1-c_1^2\sigma_B^2} - \sqrt{1-c_1^2\sigma_A^2} \right| \qquad (48)$$

Podemos obtener también μ en función de s, de forma que para $s=0$ la curva pase por μ_A:

$$\mu = \mu_A + \frac{\sqrt{2}}{c_1} \left| \operatorname{tgh}\ \left(\mp\frac{s}{\sqrt{2}} + \operatorname{arg\ sech}(|c_1|\sigma_A)\right) - \sqrt{1-c_1^2\sigma_A^2} \right| \qquad (49)$$

Nótese que el signo de c_1 es positivo si $\mu > \mu_A$ y negativo en caso contrario. Las ecuaciones (44) y (49) son las ecuaciones de las geodésicas, en el caso b, y la distancia entre dos puntos medida sobre la geodésica viene dada por (43).

A continuación intentemos encontrar c_1 y a la vez caracterizar el caso b, es decir, dados dos puntos de la variedad paramétrica (μ_A,σ_A), (μ_B,σ_B) que condiciones deben cumplir para que exista una geodésica que los una, cumpliendo las hipótesis del caso b. Elevando al cuadrado (48) obtenemos:

$$c_1^2 \left(\frac{(\mu_B-\mu_A)^2}{2} + \sigma_A^2 + \sigma_B^2 \right) - 2 = -2 \sqrt{1-c_1^2\sigma_B^2} \quad \sqrt{1-c_1^2\sigma_A^2} \tag{50}$$

para que exista solución, bajo las hipótesis consideradas, debe cumplirse:

$$c_1^2 \left(\frac{(\mu_B-\mu_A)^2}{2} + \sigma_A^2 + \sigma_B^2 \right) \leq 2 \tag{51}$$

Hallemos ahora c_1. Elevando al cuadrado (50) y reordenando términos, obtenemos:

$$c_1^2 \left(\frac{(\mu_B-\mu_A)^4}{4} + (\sigma_A^2-\sigma_B^2)^2 + (\sigma_A^2+\sigma_B^2)(\mu_B-\mu_A)^2 \right) = 2(\mu_B-\mu_A)^2 \tag{52}$$

y finalmente, como $c_1 > 0$ si $\mu_B-\mu_A > 0$, resulta:

$$c_1 = \frac{\sqrt{2}\,(\mu_B - \mu_A)}{\sqrt{\frac{(\mu_B-\mu_A)^4}{4} + (\sigma_A^2-\sigma_B^2)^2 + (\sigma_A^2+\sigma_B^2)(\mu_B-\mu_A)^2}} \tag{53}$$

y por tanto la condición (51) puede reescribirse, al sustituir c_1 por (53), como:

$$(\mu_B-\mu_A)^2 \left(\frac{(\mu_B-\mu_A)^4}{2} + \sigma_A^2 + \sigma_B^2 \right) \leq \frac{(\mu_B-\mu_A)^4}{4} + (\sigma_A^2-\sigma_B^2)^2 + (\sigma_A^2+\sigma_B^2)(\mu_B-\mu_A)^2 \tag{54}$$

equivalente a:

$$(\mu_B-\mu_A)^2 \leq 2\left|\sigma_A^2-\sigma_B^2\right| \tag{55}$$

La expresión (55) nos da la condición que deben verificar dos puntos de la variedad para que pueda existir una geodésica que los una cum

pliendo las hipótesis del caso b.

Caso c Consideremos ahora el caso $\mu_A \neq \mu_B$ y $\frac{d\sigma}{ds} > 0$ al inicio de la curva y $\frac{d\sigma}{ds} < 0$ al final. Por razones de continuidad existirá un punto intermedio de la curva con $\frac{d\sigma}{ds} = 0$, en este punto σ toma el valor máximo. Nótese que no es posible que $\frac{d\sigma}{ds} < 0$ al inicio de la curva y $\frac{d\sigma}{ds} > 0$ al final, ya que en el punto donde se anularía la derivada habría un mínimo, para σ, hecho que estaría en contradicción con (31) que implica $\frac{d^2\sigma}{ds^2} < 0$.

Dividiremos la geodésica en dos trozos. Es claro que $\sigma_{max} = \frac{1}{|c_1|}$, por tanto:

$$\int_{\sigma_A}^{1/|c_1|} \frac{d\sigma}{\sigma\sqrt{1-c_1^2\sigma^2}} - \int_{1/|c_1|}^{\sigma_B} \frac{d\sigma}{\sigma\sqrt{1-c_1^2\sigma^2}} = \int_o^{s_M} \frac{ds}{\sqrt{2}} + \int_{s_M}^{s_{AB}} \frac{ds}{\sqrt{2}} \qquad (56)$$

integrando resulta:

$$s_{AB} = \sqrt{2}\,(\text{arg sech}(|c_1|\,\sigma_A) + \text{arg sech}(|c_1|\,\sigma_B)) \qquad (57)$$

Podemos expresar σ en función de s, de forma que para $s=0$ la curva pase por σ_A:

$$\sigma = \frac{1}{|c_1|}\,\text{sech}\,(|\frac{s}{\sqrt{2}} - \text{arg sech}\,(|c_1|\sigma_A)|) \qquad (58)$$

En cuanto a μ, podemos escribir:

$$\int_{\mu_A}^{\mu_M} d\mu + \int_{\mu_M}^{\mu_B} d\mu = \frac{1}{c_1} \int_o^{s_M} \text{sech}^2(-\frac{s}{\sqrt{2}} + \text{argsech}(|c_1|\sigma_A))\,ds +$$

$$+ \frac{1}{c_1} \int_{s_M}^{s_{AB}} \operatorname{sech}^2 \left(\frac{s}{\sqrt{2}} - \operatorname{argsech} \left(|c_1| \sigma_A \right) \right) ds \qquad (59)$$

teniendo en cuenta (47) y que para s_M, σ alcanza $\frac{1}{|c_1|}$, integrando obtenemos:

$$\mu_B - \mu_A = \frac{\sqrt{2}}{c_1} \left(\sqrt{1-c_1^2\sigma_B^2} + \sqrt{1-c_1^2\sigma_A^2} \right) \qquad (60)$$

Podemos también obtener μ en función de s, de forma que para s=0, la curva pase por μ_A:

$$\mu = \mu_A + \frac{\sqrt{2}}{c_1} \left| \operatorname{tgh} \left(- \frac{s}{\sqrt{2}} + \operatorname{argsech} \left(|c_1| \sigma_A \right) \right) - \sqrt{1-c_1^2\sigma_A^2} \right| \qquad (61)$$

La expresión (58) junto con (61) constituyen las ecuaciones de las geodésicas, en el caso c, y la distancia entre dos puntos medida sobre la geodésica viene dada por (57).

Intentemos determinar, a continuación, la condición que deben verificar los puntos (μ_A,σ_A) y (μ_B,σ_B) para que se de el caso c. Elevando al cuadrado (60) obtenemos:

$$c_1^2 \left(\frac{(\mu_B-\mu_A)^2}{2} + \sigma_B^2 + \sigma_A^2 \right) - 2 = 2 \sqrt{1-c_1^2\sigma_B^2} \sqrt{1-c_1^2\sigma_A^2} \qquad (62)$$

por tanto debe verificarse:

$$c_1^2 \left[\frac{(\mu_B-\mu_A)^2}{2} + \sigma_B^2 + \sigma_A^2 \right] > 2 \qquad (63)$$

Podemos hallar c_1 a partir de (62), elevando al cuadrado y reordenando términos obtenemos finalmente el mismo valor que en el caso anterior,

en otras palabras podemos calcular c_1 a partir de (53). Al sustituir en (63) y operar obtenemos finalmente:

$$(\mu_B-\mu_A)^2 > 2\,|\sigma_A^2 - \sigma_B^2| \tag{64}$$

La expresión (64) nos da la condición que deben verificar dos puntos de la variedad para que pueda existir una geodésica que los una cumpliendo las hipótesis del caso c.

Observemos que dados dos puntos (μ_A,σ_A) y (μ_B,σ_B) siempre existirá una geodésica que los una. Bastará comprobar que:

$$|c_1|\ \sigma_A \leq 1 \qquad\qquad |c_1|\ \sigma_B \leq 1 \tag{65}$$

será suficiente probarlo para el máximo de σ_A,σ_B. Sea, por ejemplo, σ_A, la primera expresión de (65) es equivalente a:

$$2(\mu_B-\mu_A)^2\ \sigma_A^2 \leq \frac{(\mu_B-\mu_A)^4}{4} + (\sigma_A^2-\sigma_B^2) + (\sigma_A^2+\sigma_B^2)\cdot(\mu_B-\mu_A)^2 \tag{66}$$

sea $\sigma_A^2 = \sigma_B^2 + \varepsilon$, $\varepsilon \geq 0$, entonces, si llamamos $\Delta = \mu_B-\mu_A$, resulta:

$$2\Delta^2\sigma_A^2 \leq \frac{\Delta^4}{4} + \varepsilon^2 + 2\sigma_A^2\Delta^2 - \varepsilon\,\Delta^2 \tag{67}$$

equivalente a:

$$0 \leq \frac{\Delta^4}{4} + \varepsilon^2 - \varepsilon\Delta^2 \tag{68}$$

La igualdad en la expresión anterior sólo se verifica cuando $\varepsilon=\frac{\Delta^2}{2}$ lo que implica que la expresión (68) no puede tomar valores negativos,

quedando por tanto probada (65).

Notemos que las geodésicas verifican, en el caso a, la propiedad de ser rectas perpendiculares al eje de las μ. En los demás casos, de (47), (59) y (61), se deduce:

$$(\mu - \mu_M)^2 = \frac{2}{c_1^2}\,(1 - c_1^2\sigma^2) \tag{69}$$

y como $\sigma_M = \frac{1}{c_1}$ resulta:

$$\frac{(\mu - \mu_M)^2}{2\sigma_M^2} + \frac{\sigma^2}{\sigma_M^2} = 1 \tag{70}$$

es decir, las geodésicas son arcos de elipse con centro en $(\mu_M, 0)$. Es fácil comprobar que:

$$\mu_M = \frac{\mu_A+\mu_B}{2} - \frac{\sigma_B^2-\sigma_A^2}{\mu_B-\mu_A} \tag{71}$$

Notemos también una transformación admisible de los parámetros que deja constante al tensor métrico excepto en una componente. La transformación viene definida por:

$$\begin{aligned} (\mu,\sigma) &\longrightarrow (\mu,v) \\ 2\sigma^2 &= e^{-\sqrt{2}\,v} \end{aligned} \tag{72}$$

Bajo las nuevas coordenadas el tensor métrico viene definido por:

$$\bar{g}_{11} = 2e^{\sqrt{2}\,v} \qquad \bar{g}_{12} = \bar{g}_{21} = 0 \qquad \bar{g}_{22} = 1 \tag{73}$$

Notemos además que como las geodésicas están definidas para todo valor de su parámetro, estamos en condiciones de afirmar que la variedad Riemanniana E definida en (17) y (20) es completa, por lo que podemos afirmar que la distancia entre dos puntos de la variedad vendrá dada por integración del elemento de linea a lo largo de una geodésica. Resumiendo los resultados anteriores, la distancia entre dos puntos (μ_A,σ_A) y (μ_B,σ_B), d_{AB}, viene dada por:

$$\mu_A=\mu_B \quad \Rightarrow \quad d_{AB} = \sqrt{2}\left|\ln\left(\frac{\sigma_B}{\sigma_A}\right)\right|$$

$$\left.\begin{array}{l}\mu_A\neq\mu_B \\ (\mu_B-\mu_A)^2 \leq 2|\sigma_A^2-\sigma_B^2|\end{array}\right\} \Rightarrow d_{AB} = \sqrt{2}|\operatorname{argsech}(|c_1|\sigma_A)-\operatorname{argsech}(|c_1|\sigma_B)| \qquad (74)$$

$$\left.\begin{array}{l}\mu_A\neq\mu_B \\ (\mu_B-\mu_A)^2 > 2|\sigma_A^2-\sigma_B^2|\end{array}\right\} \Rightarrow d_{AB} = \sqrt{2}(\operatorname{argsech}(|c_1|\sigma_A)+\operatorname{argsech}(|c_1|\sigma_B))$$

donde $|c_1|$ viene dado por (53).

Una vez resuelto el caso normal univariante, por el resultado 3.1.1 queda resuelto el caso de la distribución log-normal, y en general de el caso de todas aquellas variables aleatorias que puedan transformarse mediante una transformación admisible, en una variable aleatoria con distribución normal.

5.3. DISTRIBUCION NORMAL MULTIVARIANTE

Sea $X = (X_1,\ldots,X_n)$ un vector aleatorio cuya función de densidad conjunta viene definida por:

$$f(X,\mu,\Sigma) = (2\pi)^{-\frac{1}{2}n} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right) \tag{75}$$

donde μ es el vector formado con las esperanzas de las X_i y Σ es la matriz de varianzas y covarianzas entre las X_i. El espacio paramétrico vendrá definido por:

$$E = \{(\mu_1,\ldots,\mu_n,\sigma_{11},\ldots,\sigma_{nn}) \in \mathbb{R}^{n+\frac{n(n+1)}{2}} / \Sigma = (\sigma_{ij}) \text{ es simétrica y definida positiva}\} \tag{76}$$

Se comprueba facilmente que E es una variedad $n+\frac{n(n+1)}{2}$ dimensional arco-conexa.

Como el logaritmo de la función de densidad viene dado por:

$$\ln f(X,\mu,\Sigma) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu) \tag{77}$$

al derivar respecto algún parámetro α, (μ_i ó σ_{ij}) resulta:

$$\frac{\partial \ln f}{\partial \alpha} = -\frac{1}{2}\left[\frac{1}{|\Sigma|}\frac{\partial|\Sigma|}{\partial\alpha} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial\sigma^{ij}}{\partial\alpha}(x_i-\mu_i)(x_j-\mu_j) - 2\sigma^{ij}\frac{\partial\mu_i}{\partial\alpha}(x_j-\mu_j)\right)\right] \tag{78}$$

siendo los σ^{ij} los elementos de Σ^{-1}, matriz también simétrica. Volviendo a derivar, ahora respecto a β, y cambiando el signo, resultará:

$$-\frac{\partial^2 \ln f}{\partial\beta\,\partial\alpha} = \frac{1}{2}\left[-\frac{1}{|\Sigma|^2}\frac{\partial|\Sigma|}{\partial\beta}\frac{\partial|\Sigma|}{\partial\alpha} + \frac{1}{|\Sigma|}\frac{\partial^2|\Sigma|}{\partial\beta\partial\alpha} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2\sigma^{ij}}{\partial\beta\partial\alpha}(x_i-\mu_i)(x_j-\mu_j) - \right.\right.$$

$$\left.\left. - 2\frac{\partial\sigma^{ij}}{\partial\alpha}\frac{\partial\mu_i}{\partial\beta}(x_j-\mu_j) - 2\frac{\partial\sigma^{ij}}{\partial\beta}\frac{\partial\mu_i}{\partial\alpha}(x_j-\mu_j) + 2\sigma^{ij}\frac{\partial\mu_i}{\partial\alpha}\frac{\partial\mu_j}{\partial\beta}\right)\right] \tag{79}$$

Hallando la esperanza y teniendo en cuenta que $E(x_i-\mu_i)=0$, resulta:

$$-E\left(\frac{\partial^2 \ln f}{\partial\beta\partial\alpha}\right) = \frac{1}{2}\left[-\frac{1}{|\Sigma|^2}\frac{\partial|\Sigma|}{\partial\beta}\frac{\partial|\Sigma|}{\partial\alpha} + \frac{1}{|\Sigma|}\frac{\partial^2|\Sigma|}{\partial\beta\partial\alpha} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2\sigma^{ij}}{\partial\beta\ \partial\alpha}\ \sigma_{ij} + 2\sigma^{ij}\frac{\partial\mu_i}{\partial\alpha}\frac{\partial\mu_j}{\partial\beta}\right)\right] \tag{80}$$

Consideremos los siguientes casos:

Caso a $\quad \alpha = \mu_k \qquad \beta = \mu_h \qquad k,h = 1,\ldots,n$

$$-E\left(\frac{\partial^2 \ln f}{\partial\mu_k\partial\mu_h}\right) = \sigma^{kh} \tag{81}$$

Caso b $\quad \alpha = \mu_k \qquad \beta = \sigma_{hm} \quad$ (ó $\beta=\mu_k \quad \alpha=\sigma_{hm}$) $\quad k,h,m=1,\ldots,n$

$$-E\left(\frac{\partial^2 \ln f}{\partial\mu_k\partial\sigma_{hm}}\right) = 0 \tag{82}$$

Caso c $\quad \alpha=\sigma_{kh} \qquad \beta=\sigma_{pq} \qquad k,h,p,q = 1,\ldots,n$

$$-E\left(\frac{\partial^2 \ln f}{\partial\sigma_{kh}\partial\sigma_{pq}}\right) = \frac{1}{2}\left[-\frac{1}{|\Sigma|^2}\frac{\partial|\Sigma|}{\partial\sigma_{kh}}\frac{\partial|\Sigma|}{\partial\sigma_{pq}} + \frac{1}{|\Sigma|}\frac{\partial^2|\Sigma|}{\partial\sigma_{kh}\partial\sigma_{pq}} + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2\sigma^{ij}}{\partial\sigma_{kh}\partial\sigma_{pq}}\ \sigma_{ij}\right] \tag{83}$$

Observemos que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^{ij}\ \sigma_{ij} = \sum_{i=1}^{n}\delta_{ii} = n \tag{84}$$

luego su derivada segunda, respecto de σ_{kh} ó σ_{pq}, se anulará:

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial\sigma_{kh}\partial\sigma_{pq}}(\sigma^{ij}\sigma_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2\sigma^{ij}}{\partial\sigma_{kh}\partial\sigma_{pq}}\ \sigma_{ij} + \frac{\partial\sigma^{ij}}{\partial\sigma_{pq}}\frac{\partial\sigma_{ij}}{\partial\sigma_{kh}} + \frac{\partial\sigma_{ij}}{\partial\sigma_{kh}}\cdot\frac{\partial\sigma_{ij}}{\partial\sigma_{pq}}\right) \tag{85}$$

y por tanto resulta:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2\sigma^{ij}}{\partial\sigma_{kh}\partial\sigma_{pq}}\sigma_{ij} = -\frac{\partial\sigma^{kh}}{\partial\sigma_{pq}}(2-\delta_{kh}) - \frac{\partial\sigma^{pq}}{\partial\sigma_{kh}}(2-\delta_{pq}) \tag{86}$$

y teniendo en cuenta que:

$$\sigma^{ij} = \frac{\Sigma^{ji}}{|\Sigma|} \tag{87}$$

siendo Σ^{ji} el adjunto del elemento σ_{ji}. Por tanto podemos escribir:

$$\frac{\partial\sigma^{ij}}{\partial\sigma_{pq}} = -\frac{1}{|\Sigma|^2}\frac{\partial|\Sigma|}{\partial\sigma_{pq}}\Sigma^{ji} + \frac{1}{|\Sigma|}\frac{\partial\Sigma^{ji}}{\partial\sigma_{pq}} \tag{88}$$

y debido a que:

$$\frac{\partial|\Sigma|}{\partial\sigma_{pq}} = \Sigma^{qp}(2-\delta_{pq}) \qquad \frac{\partial^2|\Sigma|}{\partial\sigma_{kh}\partial\sigma_{pq}} = \frac{\partial\Sigma^{qp}}{\partial\sigma_{kh}}(2-\delta_{pq}) \tag{89}$$

resulta:

$$-E\left(\frac{\partial^2 \ln f}{\partial\sigma_{kh}\partial\sigma_{pq}}\right) = \frac{1}{2|\Sigma|^2}\frac{\partial|\Sigma|}{\partial\sigma_{kh}}\frac{\partial|\Sigma|}{\partial\sigma_{pq}} - \frac{1}{2|\Sigma|}\frac{\partial^2|\Sigma|}{\partial\sigma_{kh}\partial\sigma_{pq}} = -\frac{1}{2}\frac{\partial^2\ln|\Sigma|}{\partial\sigma_{kh}\partial\sigma_{pq}} \tag{90}$$

por tanto el tensor métrico asociado a la variedad E, definida en (76) matricialmente podrá expresarse como:

$$G = \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & A \end{pmatrix} \tag{91}$$

siendo Σ^{-1} la matriz inversa de la matriz de varianzas y covarianzas, y A una matriz cuadrada de orden $n(n+1)/2$ y cuyos elementos son de la

forma:

$$a_{\mu\nu} = -\frac{1}{2}\frac{\partial^2 \ln|\Sigma|}{\partial\sigma_{kh}\partial\sigma_{pq}} \qquad \begin{matrix} k \leq h = 1,\ldots,n \\ p \leq q = 1,\ldots,n \end{matrix} \tag{92}$$

los índices μ y ν dependen de k,h,p y q, según ordenamos la parte triangular superior (o inferior) de la matriz de varianzas y covarianzas, al considerar a sus elementos como las coordenadas $n+1,\ldots,n+\frac{n(n+1)}{2}$ de la variedad E.

Hasta el momento no se ha logrado invertir explícitamente la matriz A, por lo que no se ha procedido al cálculo de los símbolos de Christoffel de segunda especie, necesarios para obtener las ecuaciones de las geodésicas de la forma habitual. Para obtener una expresión de las ecuaciones diferenciales de las geodésicas puede plantearse el problema variacional asociado al cálculo de las mismas. El cuadrado del elemento de linea será:

$$ds^2 = \sum_{p,q} \sigma^{pq}\, d\mu_p d\mu_q - \frac{1}{2}\sum_{\alpha\leq\beta}\sum_{r\leq s}\frac{\partial^2 \ln|\Sigma|}{\partial\sigma_{\alpha\beta}\partial\sigma_{rs}}\, d\sigma_{\alpha\beta} d\sigma_{rs} \tag{93}$$

por tanto habrá que resolver:

$$\delta\int_o^s \sqrt{\sum_{p,q}\sigma^{pq}\,\dot{\mu}_p\,\dot{\mu}_q - \frac{1}{2}\sum_{\alpha\leq\beta}\sum_{r\leq s}\frac{\partial^2 \ln|\Sigma|}{\partial\sigma_{\alpha\beta}\partial\sigma_{rs}}\,\dot{\sigma}_{\alpha\beta}\dot{\sigma}_{rs}}\; ds = 0 \tag{94}$$

donde $\dot{\mu}_p$ y $\dot{\sigma}_{rs}$ son las derivadas de μ_p y σ_{rs} respecto s.

Si llamamos G a la expresión interior a la raiz, planteando las ecuaciones de Euler, respecto μ_γ, resulta:

$$-\frac{d}{ds}\left(\frac{1}{2\sqrt{G}}\,\frac{\partial G}{\partial\dot{\mu}_\gamma}\right) = 0 \tag{95}$$

y si s es la longitud medida sobre el arco, como $\sqrt{G} = 1$, tendremos, debido a que $\sigma^{p\gamma} = \sigma^{\gamma p}$,

$$\sum_{p=1}^{n} \sigma^{p\gamma} \dot{\mu}_{\gamma} = A_{\gamma} \tag{96}$$

siendo los A_{γ} constantes de integración. Equivalentemente podremos escribir:

$$\dot{\mu}_{q} = \sum_{\gamma=1}^{n} \sigma_{q\gamma} A_{\gamma} \tag{97}$$

Al plantear las ecuaciones de Euler respecto las coordenadas $\sigma_{\tau\nu}$, con $\tau \leq \nu$, resulta:

$$\frac{\partial\sqrt{G}}{\partial\sigma_{\tau\nu}} - \frac{d}{ds}\left(\frac{\partial\sqrt{G}}{\partial\dot{\sigma}_{\tau\nu}}\right) = \frac{1}{2\sqrt{G}}\frac{\partial G}{\partial\sigma_{\tau\nu}} - \frac{d}{ds}\left(\frac{1}{2\sqrt{G}}\frac{\partial G}{\partial\dot{\sigma}_{\tau\nu}}\right) = 0 \tag{98}$$

pero como $\sqrt{G} = 1$, podemos escribir:

$$\frac{\partial G}{\partial\sigma_{\tau\nu}} - \frac{d}{ds}\left(\frac{\partial G}{\partial\dot{\sigma}_{\tau\nu}}\right) = 0 \tag{99}$$

Como resulta que:

$$\frac{\partial G}{\partial\sigma_{\tau\nu}} = \sum_{p,q} \frac{\partial\sigma^{pq}}{\partial\sigma_{\tau\nu}} \dot{\mu}_{p}\dot{\mu}_{q} - \frac{1}{2}\sum_{\alpha\leq\beta}\sum_{r\leq s} \frac{\partial^{3} \ln|\Sigma|}{\partial\sigma_{\tau\nu}\partial\sigma_{\alpha\beta}\partial\sigma_{rs}} \dot{\sigma}_{\alpha\beta}\dot{\sigma}_{rs} \tag{100}$$

y además:

$$\frac{\partial G}{\partial\dot{\sigma}_{\tau\nu}} = -\sum_{\alpha\leq\beta} \frac{\partial^{2}\ln|\Sigma|}{\partial\sigma_{\alpha\beta}\partial\sigma_{\tau\nu}} \dot{\sigma}_{\alpha\beta} \tag{101}$$

y

$$-\frac{d}{ds}\left(\frac{\partial G}{\partial\dot{\sigma}_{\tau\nu}}\right) = \sum_{\alpha\leq\beta}\left(\left(\sum_{r\leq s} \frac{\partial^{3}\ln|\Sigma|}{\partial\sigma_{rs}\partial\sigma_{\alpha\beta}\partial\sigma_{\tau\nu}} \dot{\sigma}_{\alpha\beta}\dot{\sigma}_{rs}\right) + \frac{\partial^{2}\ln|\Sigma|}{\partial\sigma_{\alpha\beta}\partial\sigma_{\tau\nu}} \ddot{\sigma}_{\alpha\beta}\right) \tag{102}$$

podemos escribir:

$$\sum_{p,q} \frac{\partial \sigma^{pq}}{\partial \sigma_{\tau\nu}} \dot{\mu}_p \dot{\mu}_q + \frac{1}{2} \sum_{\alpha \leq \beta} \sum_{r \leq s} \frac{\partial^3 \ln|\Sigma|}{\partial \sigma_{rs} \partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \dot{\sigma}_{\alpha\beta} \dot{\sigma}_{rs} + \sum_{\alpha \leq \beta} \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \ddot{\sigma}_{\alpha\beta} = 0$$

$$\tau \leq \nu = 1, \ldots, n \qquad (103)$$

Teniendo en cuenta que al cumplirse:

$$\sum_{q=1}^{n} \sigma^{pq} \sigma_{qr} = \delta^p_r \qquad (104)$$

se verifica que:

$$\sum_{q=1}^{n} \frac{\partial \sigma^{pq}}{\partial \sigma_{\tau\nu}} \sigma_{qr} = - \sum_{q=1}^{n} \sigma^{pq} \frac{\partial \sigma_{qr}}{\partial \sigma_{\tau\nu}} \qquad (105)$$

si sustituimos (97) en (103) y simplificamos, resulta finalmente:

$$-(2-\delta_{\tau\nu}) A_\tau A_\nu + \frac{1}{2} \sum_{r \leq s} \sum_{\alpha \leq \beta} \frac{\partial^3 \ln|\Sigma|}{\partial \sigma_{rs} \partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \dot{\sigma}_{\alpha\beta} \dot{\sigma}_{rs} + \sum_{\alpha \leq \beta} \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \ddot{\sigma}_{\alpha\beta} = 0$$

$$\tau \leq \nu = 1, \ldots, n \qquad (106)$$

El sistema de ecuaciones diferenciales de las geodésicas viene dado pues por (97) y (106), y hay que resolverlo con la condición suplementaria:

$$\sum_{p,q} \sigma^{pq} \dot{\mu}_p \dot{\mu}_q - \frac{1}{2} \sum_{\alpha \leq \beta} \sum_{r \leq s} \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{rs}} \dot{\sigma}_{\alpha\beta} \dot{\sigma}_{rs} = 1 \qquad (107)$$

para que la norma del vector tangente, respecto al parámetro natural, sea unitaria. La expresión (107), teniendo en cuenta (97), puede ser escrita como:

$$\sum_{pq} \sigma_{pq} A_p A_q - \frac{1}{2} \sum_{\alpha \leq \beta} \sum_{r \leq s} \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{rs}} \dot{\sigma}_{\alpha\beta} \dot{\sigma}_{rs} = 1 \qquad (108)$$

Es posible expresar de otra forma a (106), teniendo en cuenta que:

$$\frac{d^2}{ds^2} \left(\frac{\partial \ln|\Sigma|}{\partial \sigma_{\tau\nu}} \right) = \sum_{\alpha \leq \beta} \left(\left(\sum_{r \leq s} \frac{\partial^3 \ln|\Sigma|}{\partial \sigma_{rs} \partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \dot{\sigma}_{\alpha\beta} \dot{\sigma}_{rs} \right) + \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \ddot{\sigma}_{\alpha\beta} \right) \qquad (109)$$

podemos escribir:

$$-(2-\delta_{\tau\nu}) A_\tau A_\nu + \frac{1}{2} \frac{d^2}{ds^2} \left(\frac{\partial \ln|\Sigma|}{\partial \sigma_{\tau\nu}} \right) + \frac{1}{2} \sum_{\alpha \leq \beta} \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} \ddot{\sigma}_{\alpha\beta} = 0 \qquad (110)$$

$$\tau \leq \nu = 1, \ldots, n$$

y como resulta:

$$\frac{\partial \ln|\Sigma|}{\partial \sigma_{\tau\nu}} = (2-\delta_{\tau\nu}) \sigma^{\tau\nu} \qquad \frac{\partial^2 \ln|\Sigma|}{\partial \sigma_{\alpha\beta} \partial \sigma_{\tau\nu}} = (2-\delta_{\tau\nu}) \frac{\partial \sigma^{\tau\nu}}{\partial \sigma_{\alpha\beta}} \qquad (111)$$

obtenemos:

$$-2A_\tau A_\nu + \ddot{\sigma}^{\tau\nu} + \sum_{\alpha \leq \beta} \frac{\partial \sigma^{\tau\nu}}{\partial \sigma_{\alpha\beta}} \ddot{\sigma}_{\alpha\beta} = 0 \qquad (112)$$

$$\tau \leq \nu = 1, \ldots, n$$

multiplicando por $\sigma_{\nu\lambda}$, $\lambda \geq \tau$, y sumando respecto a ν, resulta:

$$-2 \sum_{\nu=1}^{n} A_\tau A_\nu \sigma_{\nu\lambda} + \sum_{\nu=1}^{n} \ddot{\sigma}^{\tau\nu} \sigma_{\nu\lambda} + \sum_{\alpha \leq \beta} \left(\sum_{\nu=1}^{n} \frac{\partial \sigma^{\tau\nu}}{\partial \sigma_{\alpha\beta}} \sigma_{\nu\lambda} \right) \ddot{\sigma}_{\alpha\beta} = 0 \qquad (113)$$

$$\tau \leq \lambda = 1, \ldots, n$$

pero como se cumple (105) se obtiene:

$$-2 \sum_{\nu=1}^{n} A_\tau A_\nu \sigma_{\nu\lambda} + \sum_{\nu=1}^{n} \ddot{\sigma}^{\tau\nu} \sigma_{\nu\lambda} - \sum_{\nu=1}^{n} \sigma^{\tau\nu} \ddot{\sigma}_{\nu\lambda} = 0 \qquad (114)$$

$$\tau \leq \lambda = 1, \ldots, n$$

Además,teniendo en cuenta (104), se cumple:

$$\sum_{\nu=1}^{n} A_\tau A_\nu \sigma_{\nu\lambda} + \sum_{\nu=1}^{n} \dot{\sigma}^{\tau\nu} \dot{\sigma}_{\nu\lambda} + \sum_{\nu=1}^{n} \sigma^{\tau\nu} \ddot{\sigma}_{\nu\lambda} = 0 \qquad (115)$$

$$\tau \leq \lambda = 1,\ldots,n$$

o equivalentemente:

$$\sum_{\nu=1}^{n} A_\tau A_\nu \sigma_{\nu\lambda} + \frac{d}{ds} \left(\sum_{\nu=1}^{n} \sigma^{\tau\nu} \dot{\sigma}_{\nu\lambda} \right) = 0 \qquad (116)$$

$$\tau \leq \lambda = 1,\ldots,n$$

que puede a su vez ser expresada matricialmente, junto con (97):

$$\begin{aligned} \dot{\mu} &= \Sigma A \\ AA^t\Sigma + \frac{d}{ds}(\Sigma^{-1}\dot{\Sigma}) &= 0 \end{aligned} \qquad (117)$$

sistema de ecuaciones diferenciales que definen las geodésicas para el caso de la distribución normal multivariante. Hasta el momento, dicho sistema no ha sido integrado por lo que no disponemos de una expresión que permita encontrar la distancia entre dos puntos medida sobre una geodésica. Se ha obtenido sin embargo, una acotación de la misma.

Sean (μ_A,Σ_A) y (μ_B,Σ_B) dos puntos del espacio paramétrico. Es bien sabido que siempre existe una transformación lineal tal que reduce simultáneamente a Σ_A y Σ_B a la forma diagonal, Mardia (1979):

$$\begin{aligned} \bar{\mu}_A &= T^t \mu_A & \bar{\Sigma}_A &= T^t \Sigma_A T = I \\ \bar{\mu}_B &= T^t \mu_B & \bar{\Sigma}_B &= T^t \Sigma_B T = D \end{aligned} \qquad (118)$$

donde la matriz T es una matriz cuyas columnas son los vectores propios de Σ_B respecto Σ_A y:

$$D = \text{diag}\ (\lambda_1, \ldots, \lambda_n) \tag{119}$$

siendo los λ_i los valores propios.

Intentemos ahora resolver el problema con la restricción suplementaria de que las variables aleatorias sean independientes en todo el trayecto de la curva que une ambos puntos. En otras palabras, que la matriz de varianzas y covarianzas sea diagonal.

El caso de n variables aleatorias normales independientes ya está resuelto, por haber solucionado el caso de una distribución normal univariante y por el resultado 3.3.5. Por tanto una cota superior de la distancia entre (μ_A, Σ_A) y (μ_B, Σ_B) calculando las distancias entre cada una de las funciones de densidad marginales (normales univariantes), y sacando la raíz cuadrada de la suma de sus cuadrados. La distancia d_j entre la j-ésima marginal de A y de B es la distancia calculable a partir de la expresión (74) entre los puntos $(\bar{\mu}_j, 1)_A$ y $(\bar{\mu}_j, \lambda_j)_B$, y por tanto una cota superior de la distancia vendrá dada por:

$$d^* = \sqrt{d_1^2 + \ldots + d_n^2} \tag{120}$$

Nótese que en el caso $\mu_A = \mu_B$ (y por tanto $\bar{\mu}_A = \bar{\mu}_B$), si existe una solución del sistema (117) con $\dot{\sigma}_{\alpha\beta} = 0$ si $\alpha \neq \beta$. En efecto, en este caso basta coger A=0, y resulta:

$$\frac{d^2 \ln \sigma_{ii}}{ds^2} = 0 \tag{121}$$

por tanto:

$$\ln \sigma_{ii} = A_i s + B_i \tag{122}$$

y al resolverlo con la condición que para $s=0$ $\sigma_{ii}=1$ y para un cierto d, $\sigma_{ii} = \lambda_i$, resulta

$$\sum_{i=1}^{n} \ln^2 \sigma_{ii} = \left(\sum_{i=1}^{n} A_i^2\right) s^2 \tag{123}$$

y al ser por (107):

$$\sum_{i=1}^{n} A_i^2 = 2 \tag{124}$$

resulta:

$$d = \sqrt{\frac{1}{2} \sum_{i=1}^{n} \ln^2 \lambda_i} \tag{125}$$

por tanto, bajo la condición $\mu_A=\mu_B$, la distancia entre ambos puntos no puede ser mayor que d definida en (125).

Algunas aproximaciones a la distancia buscada pueden ser encontradas en el siquiente capítulo.

6. REPRESENTACION EN DIMENSION REDUCIDA.

Resumen:

En este capítulo se examina el problema de representar k puntos de una variedad riemanniana, en un espacio euclideo de dimensión q, generalmente q=2, de forma que la distancia euclidea entre ellos se asemeje a la distancia riemanniana original.

Vamos a efectuar tal estudio considerando varias situaciones posibles, atendiendo en primer lugar a la naturaleza de la geometría riemanniana definida en el espacio paramétrico y en segundo lugar al hecho de disponer, o no, de una expresión analítica de la distancia entre dos puntos de la variedad paramétrica, así como a criterios de tipo práctico.

Sumario:

6.1. ESPACIO PARAMETRICO EUCLIDEO

Corresponde al caso más sencillo. En estas condiciones siempre existe un sistema de coordenadas bajo el cual el tensor métrico es un tensor constante, en particular la identidad. Esto se logra resolviendo el sistema (61) del capítulo 2, eligiendo las constantes de integración de forma adecuada para que el tensor métrico, en el nuevo sistema de coordenadas, sea la identidad. Ejemplo de ello puede verse en el capítulo 4,(91), (137) y (265) y en el capítulo 5, (8).

Si tenemos k puntos de coordenadas $P_i:(x_{i1},\ldots,x_{in})$ $i=1,\ldots,k$, el primer paso será efectuar la transformación de coordenadas antes indicada, obteniendo las nuevas coordenadas de los puntos, $P_i:(y_{i1},\ldots,y_{in})$ $i=1,\ldots,k$. El problema que se plantea ahora es el siguiente: dados k puntos de un espacio afín euclídeo, cuyas coordenadas están referidas a una base ortonormal, hallar una variedad lineal, de dimensión q, tal que la suma de los cuadrados de las distancias de los puntos (vectores) a dicha variedad sea mínima. Dicho en otras palabras, hallar el hiperplano, q dimensional, que "mejor se ajuste" a la nube de puntos $P_1,\ldots,P_k$. Este problema ha sido ampliamente estudiado dentro del Análisis Multivariante, Mardia (1979), y se resuelve con la técnica conocida como Análisis de Componentes Principales. Si con las nuevas coordenadas de los P_i formamos una matriz Y de k filas y n columnas:

$$Y = (y_{ij})_{k\times n} \tag{1}$$

y definimos el vector columna, de n filas Z:

$$z_j = \frac{1}{k} \sum_{h=1}^{k} y_{hj} \tag{2}$$

entonces es bien sabido que el problema se resuelve diagonalizando la matriz:

$$A = Y^t Y - ZZ^t \tag{3}$$

respecto la matriz identidad I:

$$Av_i = \lambda_i I v_i = \lambda_i v_i \tag{4}$$

La variedad lineal buscada será de la forma:

$$V_q = \{x/\ x = \alpha_1 v_1 + \ldots + \alpha_q v_q + z \quad \alpha_i \in \mathbb{R}\ \} \tag{5}$$

siendo z el vector definido en (2), y $v_1,\ldots,v_q$ q vectores propios, ortonormales, asociados a los valores propios $\lambda_1,\ldots,\lambda_q$ con tal de que éstos verifiquen:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q \tag{6}$$

Nótese que por ser A simétrica es siempre diagonalizable, por tanto la expresión (5) tendrá sentido para cualquier q, $1 \leq q \leq n$. En cuanto a las coordenadas de los puntos proyectados en la variedad lineal, pueden obtenerse a través de la ecuación matricial:

$$U = (Y - E\ Z^t)\ V \tag{7}$$

donde E es el vector $(k \times 1)$ cuyas componentes son todas iguales a 1, y V una matriz $(n \times q)$ cuyas columnas son las componentes de los vectores $v_1,\ldots,v_q$. La matriz U será por tanto una matriz $(k \times q)$, cada una de sus filas define las coordenadas de un punto proyectado en V_q. Así la proyección del punto P_i en la variedad lineal V_q, puede expresarse como:

$$\hat{p}_i = u_{i1}v_1 + \ldots + u_{iq}v_q + z \qquad i = 1,\ldots,k \tag{8}$$

y cara las representaciones gráficas, podemos prescindir del vector constante z, ya que no afecta para nada las distancias entre los pun-- tos proyectados en V_q (equivale esta supresión a una traslación del sistema de coordenadas), y asociar a cada punto P_i las coordenadas $(u_{i1},\ldots,u_{iq})$ de un sistema de referencia cartesiano.

También se verifica que la suma de los cuadrados de las distan-- cias entre las proyecciones de los P_i, P_j, extendida a todo par (i,j), es máxima dentro de la clase formada por todas las variedades lineales de dimensión q, Cuadras (1981).

6.2. ESPACIO PARAMETRICO NO EUCLIDEO

En este caso habrá que distinguir varias situaciones, atendiendo en primer lugar si disponemos o no de una expresión analítica para la distancia.

6.2.1. Caso de disponer de una expresión analítica de la distancia

En estas condiciones, dados k puntos de la variedad riemanniana, $P_1,\ldots,P_k$, pueden darse dos situaciones: pueden existir $Q_1,\ldots,Q_k$ puntos de un espacio euclídeo, de dimensión $m \leq k-1$, tales que la distancia euclídea entre ellos coincida con la distancia riemanniana original, o bien puede darse el caso contrario.

Si definimos:

$$D = (d_{ij})_{k \times k} \qquad d_{ij} = d(P_i,P_j) \tag{9}$$

$$A = (a_{ij})_{k \times k} \qquad a_{ij} = -\frac{1}{2} d_{ij}^2 \tag{10}$$

$$H = (h_{ij})_{k \times k} \qquad H = I - \frac{1}{k} EE^t \tag{11}$$

$$B = H\,A\,H \tag{12}$$

donde $E = (1,\ldots,1)^t$.

Entonces es bien sabido, Cuadras (1981), que si la matriz B es semidefinida positiva, de rango m, $m \leq k-1$, es posible encontrar k puntos de un espacio euclídeo, $Q_1,\ldots,Q_k$, tales que la distancia euclidea entre ellos, verifique:

$$d_E(Q_i,Q_j) = d_{ij} \tag{13}$$

Además, es posible obtener explicitamente una configuración de k puntos de $\mathbb{R}^m$ que verifiquen (13), mediante la diagonalización de B.

En efecto, por ser B simétrica, siempre diagonaliza ortogonalmente, por tanto:

$$B = T\, D_\lambda T^t = (T\, D_\lambda^{1/2})(D_\lambda^{1/2} T^t) = (TD_\lambda^{1/2})(TD_\lambda^{1/2})^t \qquad (14)$$

con

$$D_\lambda = \text{diag}(\lambda_1,\ldots,\lambda_m) \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0 \qquad (15)$$

siendo T una matriz $k \times m$ cuyas columnas son los vectores propios asociados a $\lambda_1,\ldots,\lambda_m$ respectivamente y ortonormales. Si definimos:

$$X = T\, D_\lambda^{1/2} \qquad (16)$$

entonces resulta que las filas de X definen k puntos $Q_1,\ldots,Q_k$ de $\mathbb{R}^m$, al considerar cada fila como las coordenadas de un punto respecto la base canónica de $\mathbb{R}^m$, y estos k puntos de $\mathbb{R}^m$ verifican (13).

Además, con el cálculo de X resolvemos a la vez el problema de la representación de los puntos $P_1,\ldots,P_k$ en un subespacio de $\mathbb{R}^m$ de dimensión q, con q generalmente igual a 2, de forma que la suma de los cuadrados de las distancias de los k puntos al subespacio sea mínima ó bien que la suma de los cuadrados de las interdistancias entre los puntos proyectados en dicho subespacio sea máxima. Bastará escoger las q primeras columnas de la matriz X.

En el caso que B no sea semidefinida positva, existirán algunos valores propios negativos:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0 > \lambda_{r+1} \ldots \geq \lambda_m \qquad (17)$$

en estas condiciones no será posible encontrar una configuración de puntos de un espacio euclídeo que verifique (13). Si procedieramos de forma análoga al caso anterior, al hallar $D_\lambda^{1/2}$ aparecerían números imaginarios puros y las filas de X serían de la forma:

$$(x_{j1},\ldots,x_{jr}, i\, x_{jr+1},\ldots,ix_{jm}) \qquad (18)$$

siendo $i = \sqrt{-1}$. Cada punto estaría representado en un espacio con ejes reales y con ejes imaginarios puros. Representaciones en dimensión reducida pueden obtenerse escogiendo, por separado, los ejes reales asociados a los mayores valores propios y los ejes imaginarios asociados a los menores valores propios. Sin embargo, la representación respecto los ejes imaginarios presenta problemas de interpretación: cuanto más separados dos puntos, más negativa, y por tanto menor, es su distancia.

Para solventar esta dificultad, se utilizan otros métodos que se basan en transformar la matriz de interdistancias D, definida en (9), en otra matriz $\hat{D}$ tal que al repetir el proceso anterior con ésta última, obtengamos que exista una configuración de puntos en un espacio euclídeo tales que sus interdistancias coincidan con los correspondientes elementos de $\hat{D}$, en otras palabras los valores propios de la matriz B, calculada a partir de $\hat{D}$, sean no negativos. Para que todo el proceso tenga utilidad la matriz $\hat{D}$ debe ser "parecida" a la matriz D, en un sentido que vamos a precisar, introduciendo el concepto de preordenación asociada a una distancia.