

# Tratamiento basado en distancias para datos longitudinales

*Anna Esteve<sup>1</sup>, Josep Fortiana<sup>2</sup>*

<sup>1</sup>Centre d'Estudis Epidemiològics sobre VIH/SIDA de Catalunya (CEESCAT). aesteve@ceescat.hugtip.scs.es

<sup>2</sup>Departament d'Estadística, Facultat de Matemàtiques Universitat de Barcelona. fortiana@ub.edu

## Resumen

Proponemos un tratamiento basado en distancias para datos longitudinales con variables observadas en escalas de medida de naturaleza heterogénea.

**Palabras Clave:** Datos longitudinales, Predicción basada en distancias.

**AMS:** 62H25, 62P10, 62P20, 62P25.

## 1. Introducción

En múltiples contextos estadísticos surge la necesidad de tratar datos de carácter longitudinal descritos por variables observadas en escalas de medida de naturaleza heterogénea. Por poner un ejemplo concreto, en el seguimiento clínico de pacientes seropositivos para la infección por el VIH, las variables de estudio pueden clasificarse, según su afinidad conceptual, en:

- Demográficas: edad, sexo, provincia de nacimiento, . . .
- Socioeconómicas: situación laboral, formación académica, . . .
- Epidemiológicas: grupo de transmisión, fecha de infección por el VIH, . .
- Clínicas: enfermedades definitivas de SIDA y relacionadas
- Tratamientos antirretrovirales
- Biológicas: recuento de CD4, carga viral (CV), . . .
- Pérdidas de seguimiento: muerte, baja del estudio, . . .

La metodología genérica basada en distancias fue propuesta inicialmente por Cuadras [1] y desde entonces, tanto él mismo como otros autores, incluyéndonos a los del presente trabajo, la han aplicado con éxito a diversos problemas de modelado y predicción estadística. En este trabajo proponemos un tratamiento basado en distancias, específico para datos longitudinales con variables de carácter mixto (cuantitativas, cualitativas, booleanas), estructuradas en grupos de procedencia diversa.

## 2. Planteamiento e implementación del modelo

Partimos de una lista ordenada de pares  $\{t_j, W_j; j = 1, \dots, J\}$ , siendo  $t_j$  instantes de tiempo y  $W_j$  matrices de  $m$  filas, con las variables observadas para los  $m$  individuos en los  $t_j$ . Obtenemos  $D$ , matriz de interdistancias entre individuos, por una operación de conjunción de métricas, cfr. [3], a partir de las diversas matrices de interdistancias asociadas a los distintos grupos de variables observadas. En principio  $D$  es  $m \cdot J \times m \cdot J$ , o  $J \times J$ -por bloques- $m \times m$ . Es suficiente fijar  $k \ll J$  y considerarla  $(2k + 1)$ -diagonal por bloques:

$$D = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1k} & 0 & \cdots & & & 0 \\ D_{12} & D_{22} & & D_{2,k} & D_{2,k+1} & 0 & & & \\ \vdots & & \ddots & & & \ddots & & & \\ D_{k1} & D_{k2} & \cdots & D_{kk} & \cdots & & D_{k,2k-1} & 0 & \\ 0 & D_{k+1,2} & & & & & & & \\ 0 & 0 & D_{k+2,2} & & & \ddots & & & \\ \vdots & \vdots & 0 & \ddots & & & & & \end{pmatrix}$$

lo que corresponde a emplear una ventana temporal de anchura  $\pm k$ . En cada intervalo  $I(j_0) = \{t_j, j = j_0 - k, \dots, j_0 + k\}$  el método de proyectores descrito en [2], [3] y [4] da lugar a la matriz de producto interior "entre",  $\hat{G}(j_0)$ , de dimensión  $(2k + 1) \times (2k + 1)$ , y la "dentro de",  $\tilde{G}(j_0)$ , de dimensión  $m \times m$ . De ellas obtenemos las respectivas configuraciones euclídeas,  $\hat{X}(j_0), \tilde{X}(j_0)$ . Finalmente, construimos las dos curvas continuas  $\hat{X}(t), \tilde{X}(t)$ , para  $t \in [t_1, t_J]$ , aplicando un suavizado a las colecciones de puntos  $(t_j, \hat{X}(j)), (t_j, \tilde{X}(j))$ .

## 3. Bibliografía

- [1] Cuadras, C. M. (1989), *Distance Analysis in discrimination and classification using both continuous and categorical variables*, in: Y. Dodge (ed.), *Statistical Data Analysis and Inference*, Amsterdam: North Holland Publishing Co., pp. 459-473.
- [2] Batista-Foguet, J. M., J. Fortiana, C. Currie and J. R. Villalbí (2003), *Socio-economic indexes in surveys for comparisons between countries*. Social Indicators Research **65**, 1-18.
- [3] Esteve, A. (2003), *Distancias estadísticas y relaciones de dependencia entre conjuntos de variables*, Tesis Doctoral, Departament d'Estadística, Facultat de Matemàtiques, Universitat de Barcelona.
- [4] Esteve, A., J. Fortiana and J M. Batista-Foguet (2004), *Analysis of Stratified Data as a Tool in Designing Socio-Economic Indexes*. Working Paper.