Análisis de datos. Técnicas aplicadas a datos de proximidad

Tema 1: Análisis de datos de proximidad.

Introducción

Uno de los problemas más interesantes en muchas disciplinas se plantea cuando necesitamos medir y entender las relaciones entre objetos siendo desconocidas las dimensiones subyacentes de los mismos y especialmente en aquellas situaciones en las que la información disponible se refiere exclusivamente a la semejanza o desemejanza entre los objetos que son motivo de estudio. Multidimensional Scaling es un método que en su origen fue concebido como un procedimiento para la construcción de una configuración de puntos en un espacio de dimensión reducida, conocida una determinada información de proximidad entre cada par de elementos del estudio.

En la actualidad este método está constituido por un conjunto de procedimientos y técnicas que no tratan de representar y también de explicar la configuración obtenida en un entorno multidimensional. Así por ejemplo, dada una matriz de correlaciones entre diversas variables, el MDS permite representar esas variables como puntos de forma que dos puntos se encontrarán tan próximos entre sí como estén de correlacionados los elementos a los que representan. Si esta relación entre correlaciones y distancias es lo suficientemente precisa, conseguiremos una representación que pondrá de manifiesto la estructura intrínseca existente, hecho que de otro modo podría permanecer oculto al investigador puesto que en genera resulta mucho más difícil observar una tabla de coeficientes de correlación que una gráfica en un plano.

Para introducir el concepto de MDS puede emplearse el siguiente ejemplo: Supongamos que se está interesado en el estudio de n ciudades respecto a su posición geográfica. Si se dispone de un mapa donde se encuentran representadas las ciudades la construcción de la correspondiente matriz de distancias entre las ciudades a partir de dicho mapa se reduce a una cuestión gráfica elemental. Supongamos por el contrario que se dispone de una matriz cuadrada $(n \times n)$ formada por las distancias lineales entre cada pareja de ciudades y se pretende reconstruir el mapa partiendo de dicha matriz. Este problema, en dimensión arbitraria, tiene solución exacta en base a los trabajos de Schoenberg (1935) y de Young & Houselholder (1938), que se recogen en el siguiente resultado y cuya demostración puede verse en Mardia et al. [1980], pg(397).

Teorema:

Sea $D_{(n\times n)}$ una matriz de distancias entre n puntos en un espacio de configuración de dimensión K y sea $B_{(n\times n)}$ la matriz dada por $B = HAH$, siendo $H_{(n\times n)}$ dada por $H = I - n^{-1}\mathbf{1}\mathbf{1}^t$ y $A_{(n\times n)}$ la matriz cuyos elementos vienen dados a

través de $a_{rs} = -d_{rs}^2 / 2$. Entonces, D es una matriz de distancias Euclídeas, si y solo si, B es semidefinida positiva. Además se tiene:

1. Si D es la matriz de distancias Euclídeas para una configuración dada por $Z_{(n\times K)} = (z_1,.,z_n)^t$, entonces $B = (HZ)(HZ)^t$, es decir $b_{rs} = (z_r - \overline{z})^t(z_s - \overline{z}), \quad \forall r,s = 1,\ldots,n,$ de donde $B \geq 0.$ B será la matriz centrada de productos escalares de Z.

2. Inversamente, si B es semidefinida positiva de rango K, entonces puede construirse una configuración asociada a B de la siguiente forma: Sean $\lambda_1 > \cdots > \lambda_K$ los K valores propios positivos de B correspondientes a los vectores propios $X_{(n\times K)} = (x_{(1)},.,x_{(K)})$, normalizados según la condición $x'_{(i)}x_{(i)} = \lambda_i, \ \forall i = 1,.,K$. Los puntos de $\square^K$ de coordenadas $x_r = (x_{r1},.,x_{rp})^t$ (donde x_r representa la r-ésima fila de la matriz X), tienen matriz de distancias D. Además esa configuración está centrada en $\overline{x} = 0$ y B es la matriz de productos escalares de esa configuración.

El resultado anterior indica que existe solución única para distancias euclídeas en un espacio de dimensión $K = rang(B)$, salvo isometrías, donde K como máximo será $(n-1)$, ya que **1** es un vector propio de B asociado al valor propio 0. El problema que resuelve MDS surge cuando se pretende una representación en un espacio de dimensión menor que K y/o la información de la que se dispone entre cada par de elementos a representar no viene dada en términos de distancia Euclídea sino en términos de pseudo-distancia o en general de proximidad respecto a algún criterio mediante coeficientes de disimilaridad o similaridad.

En el ejemplo anterior el MDS daría solución al problema de que quisiésemos reconstruir el mapa (dimensión dos) utilizando las distancias medidas por carretera como disimilaridades o empleando cualquier coeficiente de proximidad entre las ciudades basándonos en algún criterio de tipo económico, social, etc. (por ejemplo, los tiempos de viaje entre las ciudades).

Datos en MDS. Medidas de proximidad

Diferentes tipos de datos conducen por sí mismos al tipo de análisis con MDS que debe emplearse. La terminología empleada en MDS ha sido elaborada fundamentalmente en el ámbito de las ciencias de la conducta y frecuentemente puede no resultar usual en estadística. Desde un punto de vista general el término proximidad indica el concepto de cercanía en espacio, tiempo o cualquier otro contexto. Desde un punto de vista matemático, ese término hace referencia al concepto de disimilaridad o similaridad entre dos elementos.

Sea O un conjunto finito o infinito de elementos (individuos, estímulos sujetos u objetos) sobre los que queremos definir una proximidad.

Definición:
Dados dos puntos $o_i, o_j \in O$ y δ una función real de $O \times O \to \square$, con $\delta_{ij} = \delta(o_i, o_j)$. Se dirá que δ es una disimilaridad si verifica:

1. $\delta_{ij} = \delta_{ji}, \forall i, j$
2. $\delta_{ii} \leq \delta_{ij}, \forall i, j$
3. $\delta_{ii} = \delta_0, \forall i.$

La primera condición podría eliminarse aunque resulta necesaria si se desea comparar con una distancia. No obstante esa condición suele violarse cuando las disimilaridades provienen de juicios emitidos por sujetos, ya que éstos no siempre califican igual al par (i, j) que al par (j, i). Las condiciones segunda y tercera suelen establecerse igualmente para $\delta_0 = 0$, aunque también es conocido que cuando a un individuo le son presentados dos estímulos idénticos, éste tiende a asignarles algún valor de disimilaridad no nulo y generalmente positivo, y además no siempre se define $\delta_0 \geq 0$ ya que si por ejemplo las disimilaridades provienen de una transformación, éstas podrían ser negativas.

Definición:
Una función real s de $O \times O \to \square$, se dirá que es una *similaridad* si verifica:

1. $s_{ij} = s_{ji}, \ \forall i, j$
2. $s_{ii} \geq s_{ij}, \ \forall i, j$
3. $s_{ij} > s_0, \ \forall i, j.$

Algunos autores consideran $s_0 = 0$ y además suponen que $0 \leq s_{ij} \leq 1$ ya que una similaridad es un término opuesto al de disimilaridad por lo que deberá existir alguna transformación monótona t tal que $t(s) = \delta$. Una transformación de ese tipo podría ser $\delta = 1 - s$ si $0 \leq s \leq 1$, aunque otra utilizada en INDSCAL podría ser $\delta = -s$, sin que s deba estar acotada.

Puesto que la idea fundamental sobre la que se basa el MDS es la de asociar disimilaridades a distancias, hemos de verificar, entre otras, que se cumpla la desigualdad triangular. No obstante, si se cumplen los demás axiomas salvo éste, es posible transformar los datos para que ésta también sea verificada, tomando $c = \max_{i,j,h}\{\delta_{hj} - \delta_{hi} - \delta_{ij}\}$, de forma que, $\gamma_{ii} = 0; \quad \gamma_{ij} = \delta_{ij} + c, \quad \forall i \neq j.$

Existen diferentes medidas para el cálculo de disimilaridades o similaridades entre un par de variables o individuos. Si consideramos una matriz de datos $\{x_{ri}\}_{r,i}$, obtenida de n objetos sobre p variables, algunos ejemplos de medidas son:

1. *Distancia Euclídea Ponderada*.

$$\delta_{rs} = \left\{ \sum_i w_i (x_{ri} - x_{si})^2 \right\}^{1/2}$$

2. *Metrica de Minkowski.*

$$\delta_{rs} = \left\{ \sum_i \mathrm{j} x_{ri} - x_{si} \mathrm{j}^{\lambda} \right\}^{1/\lambda}, \quad \lambda \geq 1$$

3. *Separación angular*.

$$\delta_{rs} = 1 - \frac{\sum_i x_{ri} x_{si}}{\left[\sum_i x_{ri}^2 \sum_i x_{si}^2 \right]^{1/2}}.$$

Aunque pueden obtenerse valores de disimilaridad partiendo de variables estadísticas tal y como antes apuntábamos, las medidas más utilizadas se corresponden con datos psicológicos procedentes del campo de aplicación en el que surgió el MDS. El MDS es efectuado sobre datos relativos a objetos, individuos, sujetos o estímulos. Estos cuatro términos suelen usarse indiferentemente aunque generalmente, *objetos* y *estímulos* se refieren a los elementos sobre los que se emite el juicio y *sujetos* e *individuos* a los emisores de dichos juicios. La situación comprende por tanto un conjunto de objetos o estímulos y otro conjunto de sujetos o individuos.

Para ilustrar lo anterior consideremos un conjunto de objetos que podría ser diez botellas de vino tinto correspondientes a distintas bodegas. La disimilaridad δ_{ij} entre cada par de botellas (i, j) podría ser una puntuación entera entre cero y diez correspondiente a la comparación entre la i-ésima y la j-ésima botella por un experto catador de vino. El juicio podría venir dado comparando una copa del vino i con una del vino j, clasificando sus diferencias en una escala en la que el 0 representaría que los vinos son idénticos y el 10 que son completamente diferentes. Si en lugar de uno suponemos que hay varios catadores, entonces los datos serán δ_{ijr} donde los índices i, j, se refieren a los vinos y r, al r-eximo catador.

Uno de los procedimientos más utilizados para la obtención de juicios de disimilaridad es el de *juicios directos de disimilaridad* entre objetos. Para el diseño de un procedimiento de obtención de disimilaridades entre pares de estímulos, deben tenerse en cuenta tres consideraciones: La muestra de estímulos y de sujetos, la elección del procedimiento de obtención de juicios y el diseño del instrumento para presentar el procedimiento de obtención de juicios a los individuos.

Una vez seleccionada una muestra de estímulos y de sujetos, el investigador debe elegir el tipo de juicio de disimilaridad que requerirá al sujeto encuestado. Dos de los procedimientos más empleados son: *estimación de magnitudes* y *clasificación categórica*. En el procedimiento de *estimación de magnitudes* se elige un par de estímulos estándar. Cada uno de los pares restantes se denomina pares de estímulos juzgados y son comparados con el estándar. Se trata de que cada sujeto asigne un valor que indique lo disimilar de cada par respecto a la disimilaridad del par estándar.

El procedimiento de *clasificación categórica* es el método más utilizado para obtener juicios directos de disimilaridad y consiste en que cada para de estímulos sea evaluado en una escala que va de *Muy similar* a *Muy disimilar*. El número de categorías presentadas varía, aunque si se desea analizar los datos mediante procedimientos continuos, debería ser mayor de siete.

Existe otro tipo de datos usualmente analizados en MDS que son los llamados datos de perfil, que son medidas que representan el grado de preferencia que un individuo muestra sobre un objeto de una lista. Estos datos se suelen presentar en una matriz rectangular en la que las filas se corresponden a los objetos y las columnas hacen referencia a las puntuaciones que cada individuo da a cada objeto sobre una característica determinada y en una escala predefinida. Ambos tipos de datos son analizados por las técnicas MDS según diferentes modelos para cuya descripción son necesarios previamente algunos conceptos importantes en la aplicación de estas técnicas:

- **Número de vías y número de modos**. Una clasificación importante de los datos pude hacerse atendiendo al número de vías. El término vía se refiere a cada uno de los índices considerados en los datos para proceder a la clasificación de los objetos analizados. Por otra parte, es interesante también el concepto del número de modos en los datos, término que fue utilizado para referirse a cada uno de los objetos que participan en el análisis.

- **Escala de medida**. Las variables con las que se trabaje en los diferentes análisis pueden clasificarse de acuerdo a su escala de medida, es importante distinguir entre los cuatro tipos habituales: nominal, ordinal, intervalo y razón.

    - **Escala nominal.** Los datos medidos en la escala nominal responden tan solo a una clasificación y sólo es posible distinguir entre las diferentes clases establecidas.

    - **Escala ordinal.** En una escala ordinal los datos pueden ser ordenados, pero los valores observados en estas escalas no pueden ser considerados cuantitativamente.

    - **Escala tipo intervalo**. Los objetos se sitúan en una escala de forma que el tamaño de las diferencias entre estos queda reflejado por la escala, en la que no se consideran ceros reales.

- **Escala tipo razón**. Los datos medidos en una escala tipo razón son similares a los determinados mediante una escala tipo intervalo. Sin embargo, en este tipo de escalas, los objetos se sitúan de forma que su posición en ellas representa el valor exacto del atributo que se mide.

En el ejemplo de los vinos, si solo se considera un catador, los datos serían unimodales y a dos vías. Un ejemplo de datos a dos vías y bimodales puede ser cuando cada uno de *n* individuos ordena los *m* estímulos, estando sujetos a un análisis de desdoblamiento (unfolding). Un ejemplo de datos bimodales a tres vías es cuando hay varios catadores para juzgar los vinos.

Generalmente, a la hora de realizar un estudio a partir de un conjunto de datos, los investigadores disponen de los datos dispuestos en una matriz rectangular, donde las filas se corresponden con los individuos encuestados y las columnas hacen referencia a las variables evaluadas en la encuesta. La mayoría de las técnicas estadísticas están orientadas a ese tipo de datos, pero dichas técnicas solo son aplicables cuando se verifican determinadas hipótesis que en muchas ocasiones son imposibles desde el punto de vista aplicado.

Así, en diversas situaciones suele ser recomendable partir de una matriz de disimilaridad obtenida a partir de las variables y proceder al análisis de los datos transformados. Existen diferentes razones para argumentar esta forma de proceder. Una es que diversas técnicas utilizan distancias Euclídeas como disimilaridades, de forma que cualquier otra medida de disimilaridad más apropiada a los datos invalidaría la técnica. Para ciertos tipos de variables como los nominales, la única forma de ser tratadas con el objetivo que nos ocupa es mediante su transformación en matrices de disimilaridad, especialmente cuando el número de variables es muy elevado.

Diversos autores también han puesto de manifiesto *las ventajas* del MDS frente a otras técnicas multivariantes relacionadas, entre las que una de las más próximas sea el Análisis Factorial (AF). Por un lado, los numerosos problemas AF para su ilegítima aplicación cuando existen más variables que casos, no aparecen en MDS al partir de datos de disimilaridad. Por otra parte, el MDS está basado en distancias entre puntos en un espacio generalmente Euclídeo, por lo que resulta mucho más sencillo de interpretar que el AF, que está basado en ángulos entre vectores.

Así mismo, muchos modelos de AF asumen una relación lineal entre las variables, lo que representa una hipótesis demasiado fuerte para datos subjetivos, Schiffman et al.,(1981). También se observa como el AF ofrece generalmente una solución en dimensión mayor que el MDS, el cual usualmente encuentra una solución óptima en dos o tres dimensiones con escasa pérdida de información. En estudios de aplicación de ambas técnicas sobre matrices de correlación simple, también se pone de manifiesto que el MDS no métrico ofrece solución en dimensión uno mientras que el AF necesita un número comprendido entre dos factores y el total menos uno.

Modelos de MDS y medidas de ajuste

Son muchos los algoritmos que se han desarrollado para obtener la configuración que mejor se ajuste a las disimilaridades de partida. Estos algoritmos pueden clasificarse, atendiendo a un enfoque exploratorio o confirmatorio, en dos grandes grupos: modelos descriptivos y modelos probabilísticos, distinguiendo en ambos casos entre modelos métricos, si las relaciones entre distancias y disimilaridades son cuantitativas, y modelos no métricos, si éstas simplemente son ordinales.

Para describir los diferentes modelos de MDS haremos referencia en principio y sin pérdida de generalidad, al tipo de datos más usual que es de tipo disimilaridad, unimodal y a dos vías. Supongamos por tanto un conjunto de n objetos y una matriz de disimilaridades δ_{ij} entre los pares de objetos. El objetivo del MDS será encontrar una configuración de puntos en un espacio de dimensión *K* de forma que las distancias d_{ij}, no necesariamente Euclídeas, entre cada par de puntos (i, j) representen tan bien como sea posible las disimilaridades δ_{ij}. Las diferentes formas de representación serán las que establezcan las diferentes técnicas de MDS, pudiendo clasificarse como antes apuntábamos en dos grandes grupos.

- **MDS métrico**. A los modelos de MDS en los que el tipo de relación funcional que se establecen entre las disimilaridades y las distancias tiene en cuenta, además del orden, el valor intrínseco de las disimilaridades, se denominan modelos métricos. El MDS métrico suele estar asociado a datos de tipo intervalo y de razón y trata de encontrar un conjunto de puntos en un espacio de forma que cada punto represente uno de los objetos y las distancias entre los puntos sean aproximadamente iguales cuantitativamente a las correspondientes disimilaridades transformadas mediante la relación,

$$d_{ij} \approx f(\delta_{ij}),$$

donde f es una función continua, paramétrica y monótona. Dentro de estos modelos pueden destacarse principalmente los dos siguientes:

  - *MDS clásico*. Este modelo trata a las disimilaridades directamente como distancias Euclídeas mediante la relación $d_{ij} = \delta_{ij}$, de forma que es utilizada la descomposición espectral de una matriz doblemente centrada de disimilaridades para obtener la matriz de configuración.
  - *MDS mínimo cuadrático*. Este tipo de modelos obtienen la configuración ajustando mediante mínimos cuadrados las distancias $\{d_{ij}\}$ a las disimilaridades transformadas $\{f(\delta_{ij})\}$, siendo f una función del tipo descrito anteriormente.
  - *MDS máximo verosímil*. El primer modelo de este tipo fue debido a Ramsay(1982), de forma que bajo la hipótesis de log-normalidad, permite el ajuste de las disimilaridades a las distancias mediante una función de verosimilitud.

  Alguno de los modelos métricos más importantes es el modelo clásico de Torgerson (1952), o los modelos mínimo cuadráticos ALSCAL (Takane, Young y de Leewu, 1977) o SMACOF (de Leeuw y Heiser, 1980).

- **MDS no métrico**. Si se abandona la naturaleza métrica de la transformación de las disimilaridades por una relación puramente ordinal, se obtienen modelos de MDS no métricos. La transformación f en este caso puede ser arbitraria y solo obedece a la restricción monótona,

$$\delta_{ij} < \delta_{rs} \quad \Rightarrow \quad f(\delta_{ij}) \leq f(\delta_{rs})$$

Por tanto, solo las ordenaciones de las disimilaridades deben preservarse por la transformación y de ahí el término no métrico.
Entre los modelos más importantes de este tipo están:
  - *MDS mínimo cuadrático*. El modelo de Kruskal (1964), aunque la mayoría de los modelos recientes como ALSCAL o SMACOF permiten el análisis tanto métrico como no métrico.
  - *MDS máximo verosímil*. El modelo de Takane (1981) bajo hipótesis de normalidad en los errores permite la estimación por máxima verosimilitud de los parámetros.

Bibliografía

- de Leeuw J, Heiser WJ (1980). Multidimensional Scaling with Restrictions on the con_guration. In P Krishnaiah (ed.), Multivariate Analysis, Volume V, pp. 501{522. North Holland Publishing Company, Amsterdam.

- Kruskal, J. B. (1964).- Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-28, 115-129.

- Mardia, K.V., Kent, J.T. & Bibby, J.M. (1980).- Multivariate Analysis. *Academic Press*.

- Ramsay, J. O. (1982).- Some Statistical Approaches to MDS Data. *J. R. Statist. Soc. A*, 145, 285-312.

- Schiffman, S.S., Reynolds, M.L. & Young, F.W. (1981).- Introduction to Multidimensional Scaling. Theory, Methods and Applications. *Academic Press, Inc*.

- Schönemman, P. H. (1972).- An algebraic solution for a class of subjective metrics models. *Psychometrika*, 37, 441-451.

- Takane, Y., Young, F. & de Leew, J. (1977).- Nonmetric Individual Differences MDS: An Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, 42, 7-67.

- Takane, Y. (1981).- Multidimensional Successive Categories Scaling: A Maximum Likelihood Method. {\it Psychometrika}, 46, 9-28.

- Torgerson, W. S. (1958). Theory and Methods of Scaling. *Wiley, New York.*

- Young, G. & Householder, A. S., (1938).- Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3, 19-22.