Métodos biclustering aplicados a datos antropométricos: Exploración de su posible aplicación en el diseño de indumentaria

Guillermo Vinué Visús

Trabajo Final
Máster en Bioestadística

Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas

Burjassot (Valencia), Junio de 2012

Índice de contenidos

1. Introducción
2. Material y métodos
   - Base de datos
   - Preparación de los datos y propuesta de metodología
   - Software utilizado
3. Biclustering
   - Resumen
   - Definición formal y clasificación
   - Ventajas de utilizar el biclustering
4. Métodos de Biclustering
   - Algoritmo de Cheng & Church
   - Otros algoritmos
5. Conclusiones
6. Referencias básicas

Introducción
Motivación del trabajo (I)

- Proceso de desarrollo de ropa: Definición de un sistema de tallaje que permita un buen ajuste a la mayoría de la población.
- La construcción de tallas se basa en intervalos sobre una única dimensión antropométrica.
- Los estándares actuales utilizan distribuciones bivariantes.
- Las correlaciones entre las medidas antropométricas muestran una gran variabilidad en las proporciones corporales.
- De este modo, los sistemas actuales no encajan con los morfotipos corporales.
- Consecuencias: Falta de ajuste de las prendas de ropa.
    * Alto índice de devolución de ropa.
    * Gran cantidad de ropa sin vender.
    * Problemas en asignar indumentaria en entornos de trabajo.

Introducción
Motivación del trabajo (II)

- Sistemas de tallaje: Clasificar una cierta población en subgrupos homogéneos en base a dimensiones corporales clave.
- Particionar datos $\longrightarrow$ Minería de datos $\longrightarrow$ Clustering
- Clustering convencional: Agrupa por filas o columnas por separado.
- Necesidad de agrupar filas y columnas a la vez $\longrightarrow$ Biclustering
- Campo de aplicación del biclustering: Análisis de genes.
- Sin embargo, puede ser utilizado con cualquier otra base de datos: Estudio Antropométrico Nacional
- Objetivos del trabajo:
  1. Revisión teórica de algunos algoritmos de biclustering de `R`.
  2. Aplicación sobre los datos antropométricos.
  3. Análisis de los resultados en términos de sistemas de tallaje.

Material y métodos
Base de datos

- En 2006 el Ministerio de Sanidad del Gobierno de España llevó a cabo un estudio antropométrico 3D de la población femenina.

- Objetivo: Generar datos antropométricos de las mujeres españolas dirigidos a la industria de ropa.
- Base de datos: 10.415 mujeres entre 12 y 70 años, 95 medidas antropométricas y una encuesta sociológica.

Material y métodos

Preparación de los datos y propuesta de metodología

- Las prendas de ropa inferior dependen no sólo del contorno de cintura, sino de otras dimensiones.
- El biclustering encuentra grupos para un subconjunto de dimensiones.
- Se propone utilizar el biclustering para sistemas de tallaje de ropa inferior.
- Metodología propuesta:
  1. Selección de las mujeres entre 19 y 66 años no embarazadas.
  2. Selección de las variables de la parte inferior del cuerpo.
  3. Transformación de las variables a centímetros.
  4. Segmentación de los datos en tallas de cintura por la Norma UNE.
  5. Aplicación de los métodos biclustering en cada talla.
- Cada mujer es asignada a una talla de cintura. Se agrupan por dimensiones secundarias. Éstas deben determinar el diseño de prendas inferiores en cada grupo.

Material y métodos
Software utilizado

- Métodos biclustering en la red: *Bicat*, `R`.
- Paquetes de `R` para biclustering: `biclust`, `BicARE`, `isa2`, `fabia`.
- Función *biclust* de `biclust`:
  - BCBimax(): algoritmo *Bimax*.
  - BCCC(): algoritmo de Cheng & Church.
  - BCPlaid(): algoritmo correspondiente al *modelo plaid*.
  - BCSpectral(): algoritmo *Spectral*.
  - BCXmotifs(): algoritmo *Xmotifs*.
- Creación del paquete `MBDA` de `R`.

Biclustering
Resumen (I)

- Las filas se agrupan si se expresan igual en un conjunto de columnas.
- Las columnas se agrupan si incluyen filas que se expresan igual.

	c_1	c_2	c_3	c_4	c_5	...
r_1	7	8	6	4	3	⋱
r_2	7	8	6	1	8	⋱
r_3	7	7	6	2	9	⋱
r_4	4	8	8	5	2	⋱
r_5	7	8	6	5	1	⋱
⋮	⋱	⋱	⋱	⋱	⋱	⋱

Bicluster seleccionado			
	c_1	c_2	c_3
r_1	7	8	6
r_2	7	8	6
r_3	7	7	6
r_4	4	8	8
r_5	7	8	6

Biclustering
Resumen (II)

- En un cluster de filas, cada fila se define por todas las columnas.
- En un cluster de columnas, cada columna se define utilizando todas las filas que pertenecen a ese cluster.
- En un bicluster, cada fila se selecciona utilizando sólo un subconjunto de columnas y viceversa.
- Por ejemplo, en el bicluster anterior, la fila r_3 se define por las columnas c_1 y c_2 mientras que las columnas c_1, c_2 y c_3 se definen por las filas r_1, r_2 y r_5.
- El objetivo del biclustering es encontrar subgrupos de filas y columnas que sean los más parecidos entre sí y lo más diferentes al resto, mediante un clustering simultáneo de filas y columnas.

Biclustering
Definición formal y clasificación

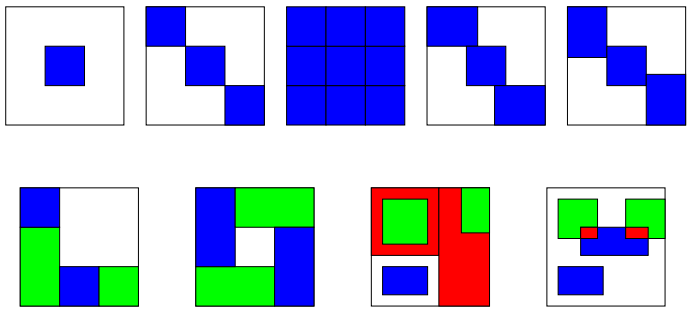
- Matriz $A = (X, Y)$, X filas, Y columnas.
- Bicluster: Submatriz (I, J) de A con $I \subseteq X$ y $J \subseteq Y$ donde a_{ij} representa la relación entre la fila i y la columna j.
- Clasificación:
  - Tipo de biclusters encontrados.

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

  - Estructura del bicluster encontrado.



Biclustering
Ventajas de utilizar el biclustering

- Grupos muy restrictivos y homogéneos.
- Selección automática de variables:
  * Eliminación automática de las variables donde no hay relación.
  * Seguridad de que las variables poco informativas no sesgan la partición.
- Reproducibilidad.
  * Algoritmos deterministas.
  * Seguridad de que los grupos encontrados realmente existen.

Algoritmo de Cheng & Church

Introducción teórica

- $a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{ij}, \quad a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij}, \quad a_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} a_{ij}$
- Bicluster perfecto: $a_{ij} = a_{iJ} + a_{Ij} - a_{IJ}$
- Por la presencia de ruido, los biclusters no son perfectos.
- Concepto de residuo: $r(a_{ij}) = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$
- Residuo cuadrado medio: $H(I,J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(r(a_{ij}))^2$
- Bicluster: Submatriz tal que $H(I,J) < \delta$ siendo $\delta \geq 0$

Algoritmo de Cheng & Church

Propuesta de aplicación sobre los datos antropométricos

Para cada talla defino un objeto nc, un objeto delta y un objeto disac.
nc es el número propuesto de biclusters a encontrar en cada talla.
delta es el valor δ del algoritmo Cheng & Church. Inicialmente es igual a 1 (valor por defecto en BCCC).
disac es el número de mujeres que se van a quedar sin agrupar. Inicialmente es igual al número de mujeres de la talla correspondiente.
La proporción de mujeres no acomodadas que queremos en cada talla la fijamos al 1 %.
mientras disac > ceiling(0.01 * *número de mujeres de la talla correspondiente*)
biclust(datosTalla, method = BCCC(), delta = , alpha = 1.5, number = nc)
disac = número de mujeres que se han quedado sin agrupar.
delta = delta + 1
fin mientras

Algoritmo de Cheng & Church

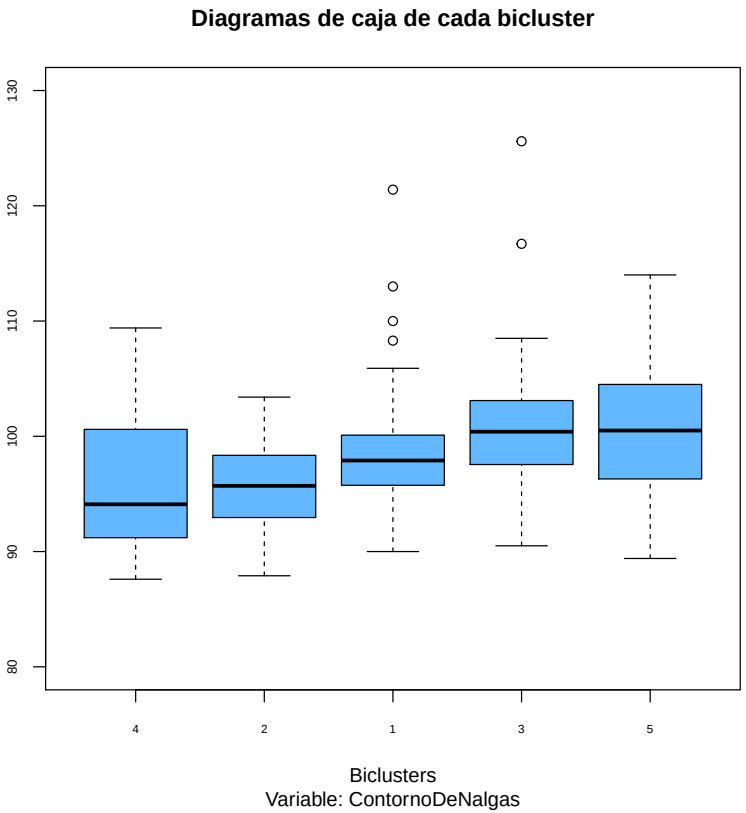
Resultados e interpretación (I). Comentarios.

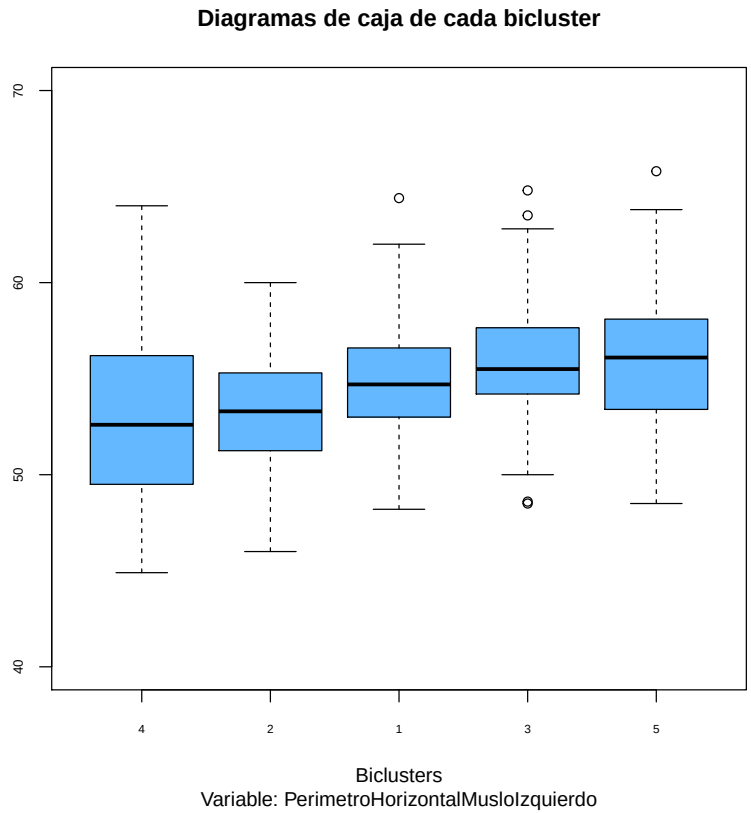
- Biclusters no solapados.
- Pocas mujeres desacomodadas
- δ pequeños.
- Resultados reproducibles.

Resultados para la talla [74,78[					
Filas y columnas de los datos: 809 10					
Mujeres desacomodadas: 4					
Valor de δ: 3					
	BC 1	BC 2	BC 3	BC 4	BC 5
Número de filas:	388	119	136	93	69
Número de columnas:	10	10	10	8	8

Algoritmo de Cheng & Church

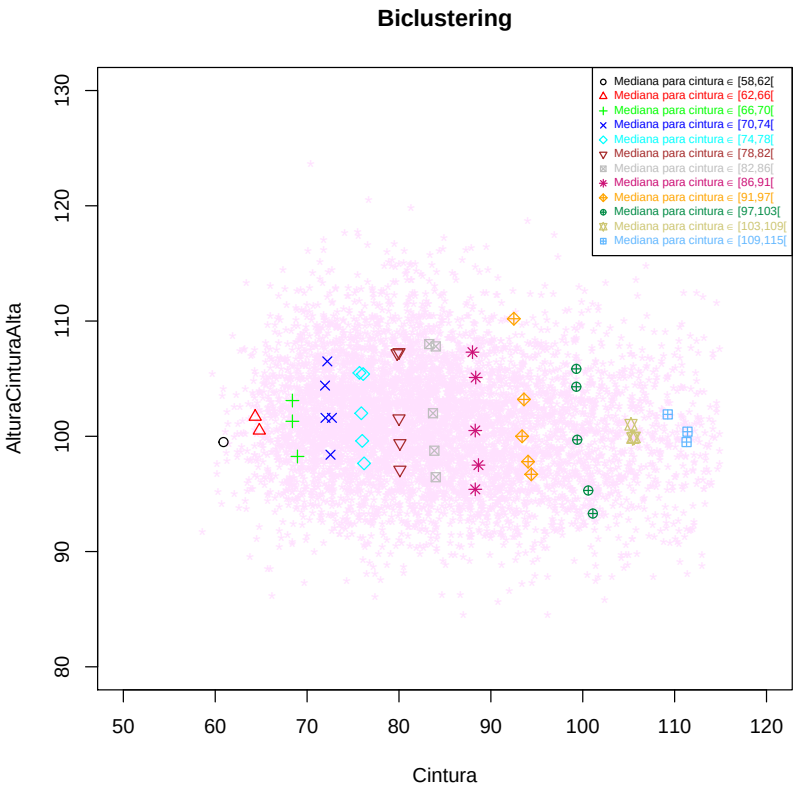
Resultados e interpretación (II). Gráficos para una talla ([74,78[).

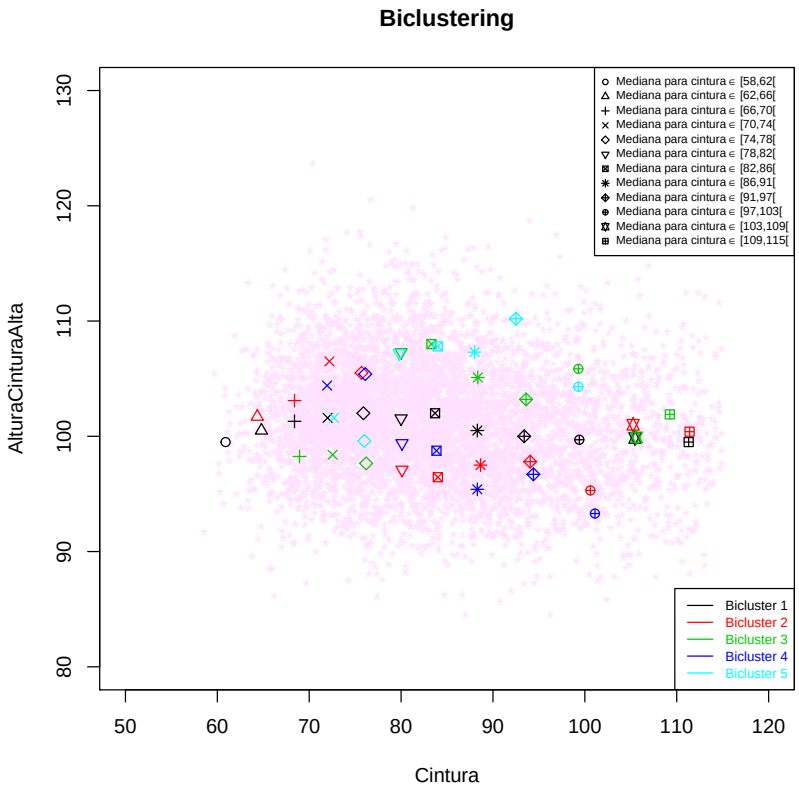




Algoritmo de Cheng & Church

Resultados e interpretación (III). Gráficos para todas las tallas.





Otros algoritmos (I)

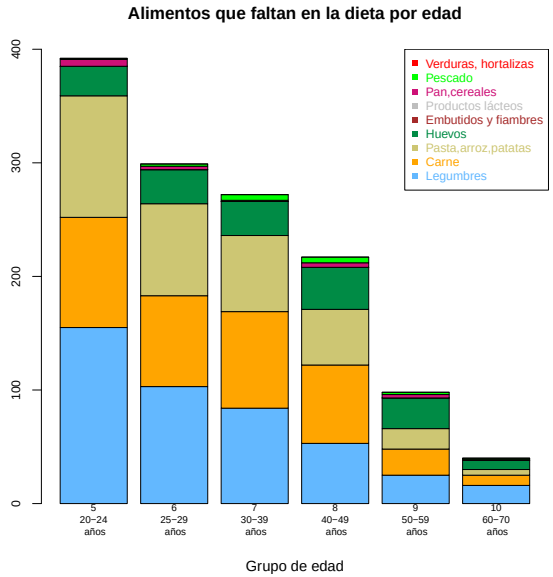
Modelo plaid, Xmotifs

- Algoritmo del *modelo plaid*:
  - Se trabaja con el total de variables (sin estandarizar).
  - Biclusters solapados.
  - Muchas mujeres desacomodadas.
  - Resultados no reproducibles.
- Algoritmo *Xmotifs*:
  - Primera opción: Función *discretize* de `biclust` $\longrightarrow$ No se encuentran biclusters.
  - Alternativa: Trasponer la matriz de datos y quitar los decimales $\longrightarrow$ No se encuentran biclusters.

Otros algoritmos (II)

Spectral, Bimax

- Algoritmo *Spectral*:
  - Tres métodos de preprocesamiento → No se encuentran biclusters.
  - La fase del preprocesamiento forma parte del propio algoritmo.
  - No se puede adaptar como se intentó con el algoritmo *Xmotifs*.
- Algoritmo *Bimax*: Ref. [Dolnicar et al, (2011)] → Variables binarias de respuesta múltiple → Variable *Listado de alimentos de los que se carece*.



- Biclusters no solapados.
- Resultados reproducibles.

Número de biclusters encontrados: 7							
	BC 1	BC 2	BC 3	BC 4	BC 5	BC 6	BC 7
Número de filas:	6	22	11	17	8	42	12
Número de columnas:	7	5	5	4	4	3	3

Conclusiones

1. Objetivos planteados en el trabajo:
   - Revisión teórica de los cinco métodos biclustering del paquete `biclust` de `R`.
   - Aplicación sobre la base de datos antropométrica.
   - Interpretación en términos de tallaje.
2. Conclusiones más importantes:
   - El preprocesamiento de los datos impide encontrar grupos (*Modelo plaid*, *Xmotifs*, *Spectral*).
   - El método que ofrece mejores resultados es el de Cheng & Church.
   - El método *Bimax* es una alternativa interesante en estudios de mercado.

Referencias básicas I


Alemany, S., González, J. C., Nácher, B., Soriano, C., Arnáiz, C., Heras, H., 2010. Anthropometric survey of the spanish female population aimed at the apparel industry. In: Proceedings of the 2010 Intl. Conference on 3D Body scanning Technologies. Lugano, Switzerland.


Bagherzadeh, R., Latifi, M., Faramarzi, A., 2010. Employing a three-stage data mining procedure to develop sizing system. World Applied Sciences Journal 8 (8), 923–929.


Cheng, Y., Church, G. M., 2000. Biclustering of expression data. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (1), 93–103.


Chunga, M., Lina, H., , Wang, M.-J. J., 2007. The development of sizing systems for taiwanese elementary- and high-school students. International Journal of Industrial Ergonomics 37, 707–716.

Referencias básicas II


Dolnicar, S., Kaiser, S., Lazarevski, K., Leisch, F., 2011. Biclustering: Overcoming Data Dimensionality Problems in Market Segmentation. Journal of Travel Research.


European Committee for Standardization, 2002. European Standard EN 13402-2: Size system of clothing. Primary and secondary dimensions.


Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch., F., 2011. biclust: BiCluster Algorithms. R package version 1.0.1.
URL `http://CRAN.R-project.org/package=biclust`


Kaufman, L. and Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, New York.

Referencias básicas III


Kluger, Y., Basri, R., Chang, J. T., Gerstein, M., 2003. Spectral Biclustering of Microarray Data:Coclustering Genes and Conditions. Genome Research (13), 703–716.


Lazzeroni, L., Owen, A., 2002. Plaid models for gene expression data. Statistica Sinica (12), 61–86.


Madeira, S. C., Oliveira, A. L., 2004. Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE Transactions on Computational Biology and Bioinformatics 1, 24–45.


Murali, T., Kasif, S., 2003. Extracting conserved gene expression motifs from gene expression. Pacific Symposium on Biocomputing (8), 77–88.

Referencias básicas IV


Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E., 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22 (9), 1122–1129.


R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org/`


Tanay, A., Sharan, R., Shamir, R., 2004. Biclustering Algorithms: A Survey. Handbook of bioinformatics.


Turner, H., Bailey, T., Krzanowski, W., 2005. Improved biclustering of microarray data demonstrated through systematic performance tests. Computational Statistics and Data Analysis (48), 235–254.