Clusters

Walter Sosa Escudero

(wsosa@udesa.edu.ar)

Universidad de San Andres y CONICET

Ideas generales

- X matriz de N filas y p columnas.
- Cada fila es un 'punto' de p dimensiones.
- Cada columna se corresponde con una variable. Ejemplo: 40 alumnos, cuatro preguntas en un examen (cada alumno es un 'punto').
- Cluster: grupo de *puntos*.
- Objetivo: dividir los puntos en clusters de modo que los puntos *dentro* de un cluster sean similares y a su vez distintos a los de cualquier otro cluster.

Dissimilarity

- $y, z \in \Re^p$. Cuan disimiles son y y z, con p coordendas cada uno?
- Variables cuantitativas: distancia euclidea:

$$d(x, z) = \left[\sum_{j=1}^{p} (x_j - y_j)^2 \right]^{1/2}$$

Agrega disimilitudes para cada atributo.

- Distancia de Minkowski:

$$d(x, z) = \left[\sum_{j=1}^{p} |x_j - y_j|^m \right]^{1/m}$$

Variables binarias:

- $\sum_{j=1}^{p}(x_j - y_j)^2 =$ numero de coincidencias.
- En varias ocasiones es relevante ponderar los aciertos en forma distinta (ejemplo: personas que hablan griego).
- Ver Tabla 12.2 Johnson y Wichern para varias alternativas.

Dissimilarity Matrix: D. matriz $N \times N$ donde $D_{ij} = D(x_i, x_j)$

- Este es el 'input' del analisis de clusters.
- Idealmente las D_{ij} son verdaderas distancias, de modo que D es simetrica y con diagonal principal nula, no es siempre el caso.
- El analisis es muy sensible a la eleccion de D.

Cluster analysis

- Cada punto, indizado por $i \in \{1, \ldots, N\}$.
- Supongamos que sabemos de antemano que hay K clusters. Visitadores medicos? Clases sociales? Pobres?
- Cada cluster, indizado por $k \in \{1, \ldots, K\}$.
- Un mecanismo de clusters asigna cada punto a un solo cluster:

$$k = C(i)$$

$C(i) =$ 'enconder', $C(i) : \{1, \ldots, N\} \to \{1, \ldots, K\}$

Analisis de Cluster: encontrar $C^*(i)$ optimo, en base a la matriz de dissimilarities.

Consideremos la siguiente funcion de perdida:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \left[\sum_{i,j/C(i)=C(j)=k} d(x_i, x_j) \right]$$

Intuitivamente: agrega las disimilitudes *dentro* de cada cluster.

Notar que

$$
\begin{aligned}
T &= \frac{1}{2}\sum_{i,j} d_{ij} \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i,j/C(i)=C(j)=k} d(x_i,x_j) + \sum_{i/C(i)=k,j/C(j)\neq k} d(x_i,x_j)\right] \\
&= W(C) + B(C)
\end{aligned}
$$

- T es la disimilitud total, entre todas las observaciones (no depende de la clusterizacion)
- BC es la agregacion de las distancias *entre* clusters.

Entonces, minimizar $W(C)$ es equivalente a maximizar $B(C)$.

Algoritmos combinatorios

- Minimizar $W(C)$ chequeando todas las posibles clusterizaciones.
- Ventaja: conduce a un minimo global.
- Desventaja: computacionalmente impensable en la practica.
- $S(N, K)$: cantidad de clusterizaciones en base a N puntos y K clusters.

$$S(N,K) = \frac{1}{K!} = \sum_{k=1}^{K} (-1)^{K-k} \left(\begin{array}{c} K \\ k \end{array} \right) k^N$$

- $S(10, 4) = 34.105$, $S(19, 4) = 10^{10}$.

Supongamos que como dissimilarity usamos el cuadrado de la distancia euclidea. La funcion de perdida es:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i,j/C(i)=C(j)=k} ||x_i - x_j||^2\right]$$

con $||x_i - x_j|| = \left[\sum_{s=1}^{p}(x_{is} - x_{js})^2\right]^{1/2}$
Es facil mostrar que

$$W(C) = \sum_{k=1}^{K} N_k \left[\sum_{i/C(i)=k} ||x_i - \bar{x}_k||^2\right]$$

con $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$, es el vector de p medias de todas las variables para cada cluster.

El problema entonces es:

$$C^* = \min_C \sum_{k=1}^{K} N_k \left[\sum_{i/C(i)=k} ||x_i - \bar{x}_k||^2 \right]$$

Notar que para cualquier conjunto S de observaciones:

$$\bar{x}_S = \operatorname*{argmin}_m \sum_{i \in S} ||x_i - m||^2$$

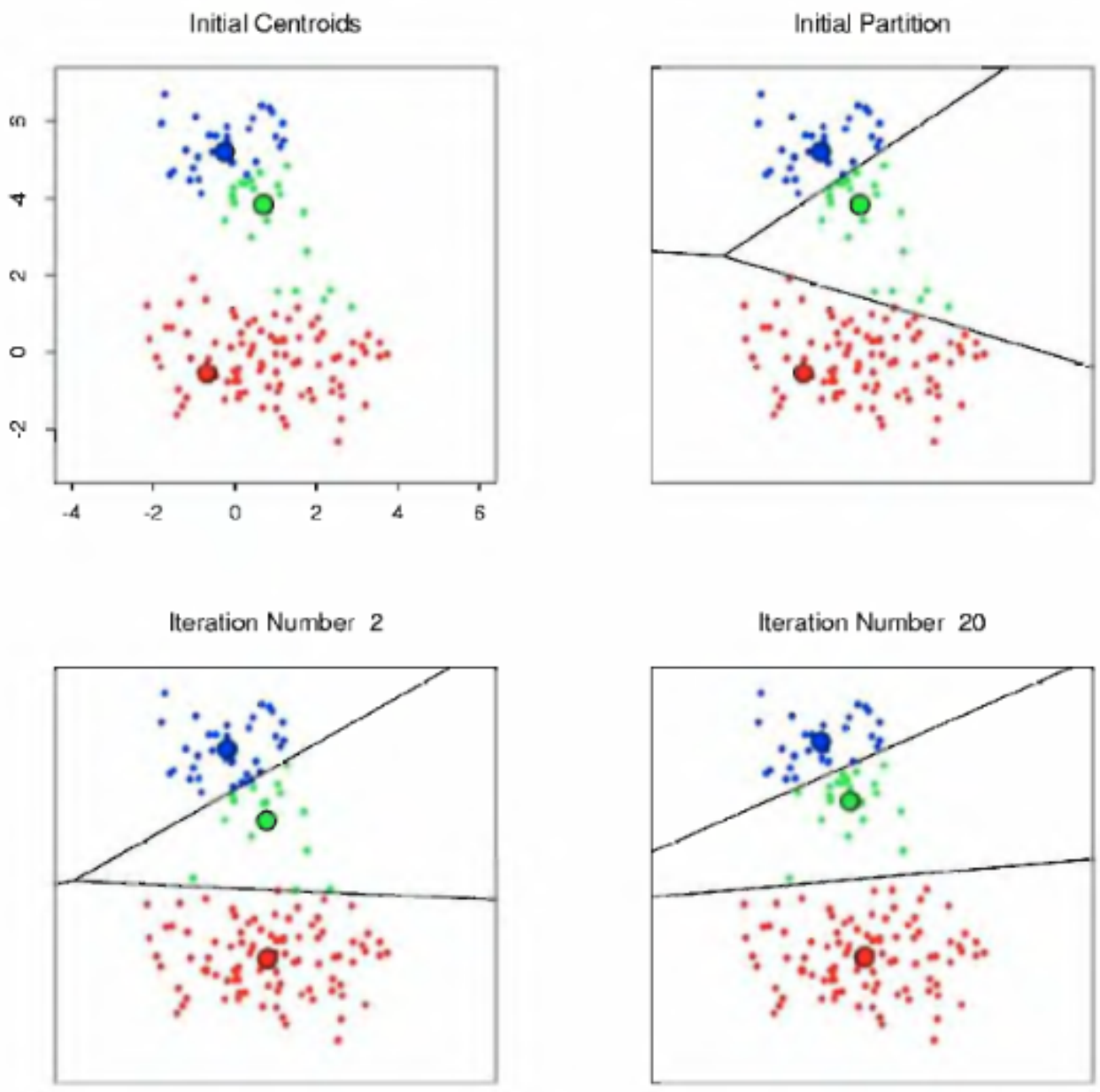
Lo que sugiere el algoritmo de $K-$medias.

Algoritmo de $K-$medias:

1. Empezar con una clusterizacion C, computar los K vectores de medias para cada variable.
2. Reasignar las observaciones al cluster mas cercano en base a las medias computadas anteriormente
3. Iterar hasta que no haya reasignaciones.

Idea: el mecanismo optimiza primero dentro del cluster (elige las medias) y luego optimiza reasignando las observaciones, dejando quietas las medias.

Problema: es convergente, pero puede hacerlo a un minimo local.


Initial Centroids
Initial Partition
Iteration Number 2
Iteration Number 20
6
4
2
0
-2
-4
-2
0
2
4
6

K-mediods

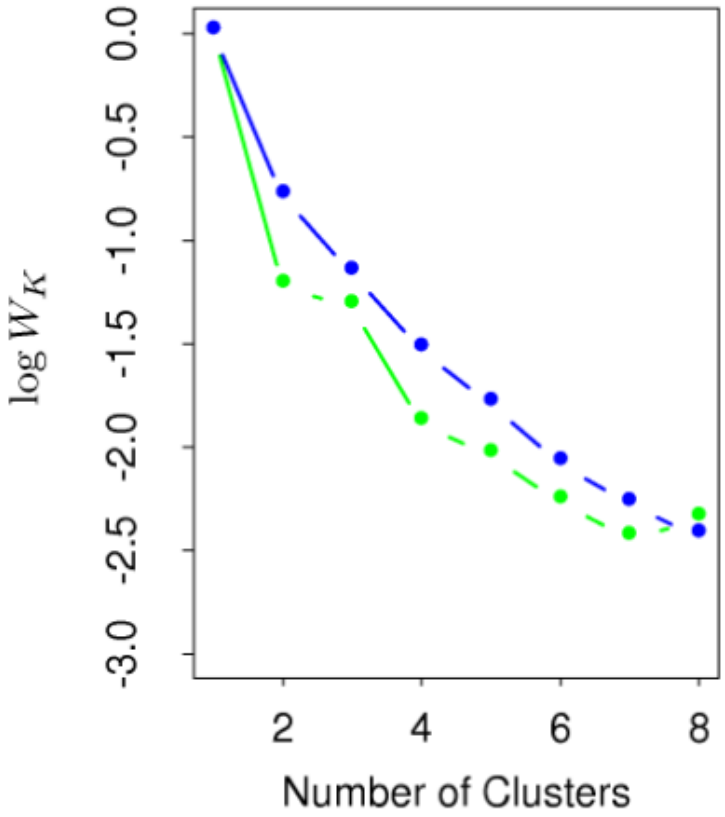
1. Empezar con una clusterizacion C encontrar la observacion dentro del cluster que minimiza la distancia con respecto a los puntos restantes. (la 'observacion central').
2. Reasignar los puntos con respecto a la observacion central.
3. Iterar hasta que no haya reasignaciones.

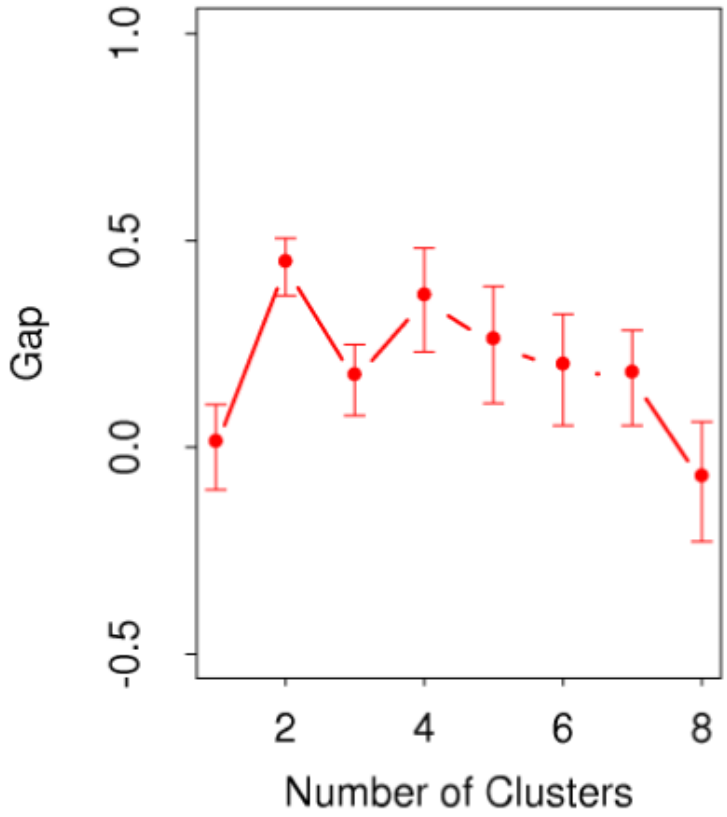
La media como 'centroide' es poco robusta (sensible a outliers).
Es computacionalmente mas costoso. Alternativa: $K-$medianas.

Cuestiones Practicas

- Inicializacion: puede ser en base a clusters o medias. Stata elige por default K observaciones al azar como centro (similar a K-mediods).
- Numero de clusters: No hay un mecanismo comunmente aceptado. En algunos casos es exogeno (visitadores medicos, pobreza?).
- Cuestion: La within dissimilarity $W(C)$ cae con el numero de clusters (en el extremo cuanto da?).
- K optimo se correponde con un quiebre en el dibujo de $W(C)$ incrementando la cantidad de clusters.

Gap statistic (Tibshirani (2001): Comparar la curva $\log W(C)$ para los datos actuales, con la curva correspondiente de un conjunto de 'pseudo-datos' uniformemente distribuidos el (hiper) rectangulo que contiene a los datos actuales. El K optimo corresponde a la maxima distancia


log W_K
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
2
4
6
8
Number of Clusters


Gap
1.0
0.5
0.0
-0.5
2
4
6
8
Number of Clusters

Cluters de pobres

the review of income and wealth

Review of Income and Wealth 2014
DOI: 10.1111/roiw.12127

DEPRIVATION AND THE DIMENSIONALITY OF WELFARE: A VARIABLE-SELECTION CLUSTER-ANALYSIS APPROACH

BY GERMÁN CARUSO

University of Illinois at Urbana-Champaign

WALTER SOSA-ESCUDERO* AND MARCELA SVARC

Universidad de San Andrés and CONICET

Data

- Gallup World Poll. Large socio-economic data set. Gasparini et al. (2010). Includes many aspects that define welfare.
- Our data set: same 15 variables used by Gasparini et al. (2010).
- Monetary (income), Non-monetary (access to water, electricity, etc.), subjective (self rated welfare).

The poor as a cluster

- All tests (Calinsky/Harabasz, and Tibshirani's Gap) suggest two-groups.
- Multivariate Kolmogorov/Smirnov test (Cuesta Albertos et al. (2006) suggest that groups are statistically different.
- One group presents substantially and systematically lower values for all variables and factor (economically different) and is labeled as the 'multidimensional poor'
- Results robust to initial conditions and clustering strategies.

Dimensionality

The reduced set contains:

1. Monthly household income,
2. Not having enough money to buy food over the last year in at least three opportunities.
3. Having a computer at home or at the place you live.

- Computationally expensive.
- Results not essentially different to previous results.
- Multidimensionality: income is not enough. Three variables.
- If only income is kept, 40% of the multidimensional poor are reallocated wrongly.
- 54% of individuals in the lowest decile are classified as multidimensionally poor. This proportion decreases with income.
- The significant but weak relation with income speaks about multidimensionality.
- Rankings: Honduras (high multidimensionally poor, low income poor). Argentina (very similar rates (22%).

Concluding remarks

- A procedure that finds the group as a coherent (statistically and economically) group that is systematically deprived.
- The reduced set is readily interpretable and can be useful for policy purposes.
- Results point towards the multidimensionality of welfare. Inability if income to capture deprivation.
- Use of multidimensional methods in economics.
- Further work: ex-ante rule (poverty definition), more sophisticated clustering strategies, other data sets and cases.