# Clasificación Bimodal

Metodos clásicos
Uso de Metaheurísticas
Uso de un índice difuso
Experimento de Monte Carlo

# Clasificación bimodal

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | | | |
| $A_2$ | | | |
| $A_3$ | | | |

- Minimizar

  $$W(P) = \sum_{k=1}^{r} \sum_{l=1}^{m} \sum_{i \in A_k} \sum_{j \in B_l} (x_{ij} - g_{kl})^2$$

  con

$$g_{kl} = \frac{1}{|A_k||B_l|} \sum_{i \in A_k} \sum_{j \in B_l} x_{ij}$$

# Notación

- Data set: $X = (x_{ij})_{n\times m}$ with modes $I$, $J$ ($|I| = n$, $|J| = m$, $I \cap J = \emptyset$), $x_{ij} \geq 0$
- $P = \{A_k | k = 1, \ldots, K\}$: partition of mode $I$
- $Q = \{B_l | l = 1, \ldots, L\}$: partition of mode $J$
- $A_k \times B_l$: two-mode class of $I \times J$
- $(P, Q)$ is a **two-mode partition** of $I \times J$ in $K, L$ classes
- $P, Q$: indicator classes of partitions $P, Q$

CIMPA-UCR

# Notación

| | |
|---|---|
| $\mathbf{X}_{n\times m} = (x_{ij})_{n\times m}$ | Two-mode data matrix of $n$ rows and $m$ columns. |
| $\mathbf{P}_{n\times K} = (p_{ik})_{n\times K}$ | Cluster membership matrix of the rows with $K$ the number of row clusters, $p_{ik} = 1$ if row $i$ belongs to row cluster $k$, and $p_{ik} = 0$ otherwise. |
| $\mathbf{Q}_{m\times L} = (q_{jl})_{m\times L}$ | Cluster membership matrix of the columns with $L$ the number of column clusters, $q_{jl} = 1$ if column $j$ belongs to column cluster $l$, and $q_{jl} = 0$ otherwise. |
| $\mathbf{V}_{K\times L} = (v_{kl})_{K\times L}$ | Matrix with cluster centers for row cluster $k$ and column cluster $l$. |
| $\mathbf{E}_{n\times m} = (e_{ij})_{n\times m}$ | Matrix with errors from cluster centers. |

CIMPA-UCR

# Criterio

Given the two-mode matrix $X$ and fixing $K$ and $L$, the problem is to find partitions $P = (p_{ik})_{n\times K}$ and $Q = (q_{jl})_{m\times L}$, and an association matrix $V = (w_{kl})_{K\times L}$ such that

$$Z = Z(P,Q,\ V) = \sum_{i\in I}\sum_{j\in J}\left(x_{ij} - \sum_{k=1}^{K}\sum_{l=1}^{L} p_{ik}q_{jl}\ v_{kl}\right)^2$$

is **minimum**,
where the $x_{ij}$ are centered $(x_{ij} \leftarrow x_{ij} - \overline{x})$ and $\hat{x}_{ij} = \sum_k \sum_l p_{ik}q_{jl}\ v_{kl}$.

# Simplificación

$Z$ can be written in the form (Gaul & Schader (1996)):

$$Z = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ik} q_{jl} (x_{ij} - v_{kl})^2. \qquad (1)$$

Given $P$ and $Q$, the matrix $V$ that minimizes $Z$ is obtained for (see Gaul & Schader (1996)):

$$w_{kl} = \frac{1}{a_k b_l} \sum_{i \in A_k} \sum_{j \in B_l} x_{ij}, \qquad (2)$$

where $a_k = |A_k|$ and $b_l = |B_l|$ are the cardinalities of the classes.

# Simplificación

- We can write:

$$\mathbf{X} = \mathbf{P}\mathbf{V}\mathbf{Q}' + \mathbf{E},$$

- Hence, the criterion is:

$$f(\mathbf{P},\mathbf{Q},\mathbf{V}) = \|\mathbf{X} - \mathbf{P}\mathbf{V}\mathbf{Q}'\|^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ik}q_{jl}(x_{ij} - v_{kl})^2.$$

# Métodos comparados

- Alternating exchanges (Gaul & Schader 1996)
- Two-mode K-means (Govaert 83; Baier, Gaul & Schader 1997)
- Simulated annealing (Trejos & Castillo 2000)
- Tabu search (Castillo & Trejos 2002)
- Fuzzy approach (Groenen, Roosmalen, Trejos & Castillo 2008)

# Métodos existentes

- Intercambios alternantes (Gaul & Schader 96): transferencias en cada modo, iteraciones mientras el criterio mejora
- *K*-medias (Govaert 83; Baier, Gaul, Schader 97): fases de asignación y representación, hasta estabilización

# Notación

- $A_k \xrightarrow{i} A_{k'}$: transfer of the row mode $i$ from class $A_k$ to class $A_{k'}$
- $B_l \xrightarrow{j} B_{l'}$: transfer of column object $j$ from class $B_l$ to class $B_{l'}$

All the following algorithms begin with an initial two-mode partition $(P, Q)$, which can be generated at random or it can be the result of some previous knowledge on the data

CIMPA-UCR

# Intercambios Alternantes

1. Initialize $P, Q$; compute $W$ according to (2).
2. Alternate the following steps until $Z$ does not improve:

    a) Make $A_k \xrightarrow{i} A_{k'}$ for $i$ and $k'$ chosen at random; if $\Delta Z := Z^{\text{new}} - Z^{\text{current}} < 0$ accept the transfer, redefine $P$ and compute $W$ using (2).

    b) Make $B_l \xrightarrow{j} B_{l'}$ for $j$ and $l'$ chosen at random; if $\Delta Z < 0$ accept the transfer, redefine $Q$ and compute $W$ using (2).

CIMPA-UCR

# K-medias Bimodal

1. Initialize $P^0, Q^0$; compute $W^0$ according to (2); let $t := 0$.
2. Repeat the following steps until $Z$ does not improve:

   *a*) Given $W := W^t$ and $Q := Q^t$, define $\tilde{P}$ by

   $$\tilde{A}_k := \{i \in I | \sum_l \sum_{j \in B_l} (x_{ij} - w_{kl})^2 \longrightarrow \min_{k'} \sum_l \sum_{j \in B_l} (x_{ij} - w_{k'l})^2$$

   *b*) Given $W := W^t$ and $P := P^t$, define $\tilde{Q}$ by

   $$\tilde{B}_l := \{j \in J | \sum_k \sum_{i \in A_k} (x_{ij} - w_{kl})^2 \longrightarrow \min_{l'} \sum_k \sum_{i \in A_k} (x_{ij} - w_{kl'})^2$$

   *c*) Calculate $W^t$ according to (2).

   *d*) Let $t := t + 1$, $P^t := \tilde{P}$, $Q^t := \tilde{Q}$.

# Uso de Sob.Sim.

- Vecindario generado por la transferencia en uno de los dos modos (escogido al azar)
- Si se escoge el modo fila: escoger al azar (unif.) *i*, escoger al azar (unif.) *k'*, hacer la transferencia de *i* de su clase actual a la clase *k'*
- Idem para columnas

CIMPA-UCR

# Sobrecalentamiento Simulado

1. Initialize $P, Q$; calculate $W$ according to (2).
2. Estimate the initial temperature $c_0$ using the initial acceptance rate $\chi_0$.
3. For $t := 0$ until $\lambda$ do:

   $a)$ choose at random a mode ($I$ or $J$) with uniform probability 1/2

   $b)$ make $A_k \stackrel{i}{\longrightarrow} A_{k'}$ or $B_l \stackrel{j}{\longrightarrow} B_{l'}$ in the following way:
   - choose at random $i$ or $j$ (according to the chosen mode);
   - choose at random $k'$ or $l'$, different from the current class of $i$ or $j$, noted $k$ or $l$, depending on the corresponding case;
   - calculate $\Delta Z$.

   $c)$ Accept the transfer if $\Delta Z < 0$, otherwise accept it with probability $\exp(-\Delta Z/c_t)$.
4. Make $c_{t+1} := \gamma c_t$ and return to step 1, until $c_t \approx 0$.

CIMPA-UCR

# Uso de BT

- Movimiento: transferencia de un objeto fila o un objeto columna a una nueva clase
- Lista tabú contiene los valores de $W(P)$
- Generar una muestra del vecindario
- No se aplica un criterio de aspiración

# Búsqueda Tabú

1. Start with an initial partition $(\mathbf{P}, \mathbf{Q})$ and an empty tabu list. Set $(\mathbf{P}, \mathbf{Q})_{opt} = (\mathbf{P}, \mathbf{Q})$.

2. Choose the number of iterations $S$ and a maximum length of the tabu list.

3. Perform the following steps $S$ times.

   (a) Generate a neighborhood of partitions $N$, consisting of the partitions that can be constructed by transferring one row or column object from $(\mathbf{P},\mathbf{Q})$ to another cluster.

   (b) Choose the partition $(\mathbf{P}, \mathbf{Q})_{cand}$ as the partition in $N$ with the lowest value of $Z(\mathbf{P}, \mathbf{Q})$ that is not on the tabu list.

   (c) Set $(\mathbf{P}, \mathbf{Q}) = (\mathbf{P}, \mathbf{Q})_{cand}$. If $Z((\mathbf{P}, \mathbf{Q})_{cand}) < Z((\mathbf{P}, \mathbf{Q})_{opt})$, then $(\mathbf{P}, \mathbf{Q})_{opt} = (\mathbf{P}, \mathbf{Q})_{cand}$.

   (d) Add $(\mathbf{P},\mathbf{Q})$ to the tabu list. Remove the oldest item from the tabu list, if the list exceeds its maximum length.

# Enfoque Difuso

We handle fuzzy two-mode partitions $(P, Q)$ where $p_{ik}, q_{jl} \in [0,1]$ instead of $p_{ik}, q_{jl} \in \{0,1\}$. Let $s > 1$ be a real number. The criterion to be minimized is

$$f_s(\mathbf{P}, \mathbf{Q}, \mathbf{V}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ik}^s q_{jl}^s (x_{ij} - v_{kl})^2$$

and it is proved that, for fixed $s, P$ and $Q$ the best matrix $V$ is given by

$$v_{kl} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ik}^s q_{jl}^s x_{ij} / \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ik}^s q_{jl}^s \qquad (12)$$

CIMPA-UCR

# Enfoque Difuso (2)

$$L_i(\mathrm{P}, \mathrm{Q}, \mathrm{V}, \lambda) = \sum_{k=1}^{K} p_{ik}^s \sum_{l=1}^{L} \sum_{j=1}^{m} q_{jl}^s (x_{ij} - v_{kl})^2 - \lambda \left( \sum_{k=1}^{K} p_{ik} - 1 \right).$$

Defining $c_{ik} = \sum_{l=1}^{L} \sum_{j=1}^{m} q_{jl}^s (x_{ij} - v_{kl})^2$ and taking partial derivatives of $L_i$ gives

$$\frac{\partial L_i}{\partial p_{ik}} = s p_{ik}^{s-1} c_{ik} - \lambda \quad \text{and} \quad \frac{\partial L_i}{\partial \lambda} = \sum_{k=1}^{K} p_{ik} - 1.$$

Setting these derivatives to zero and solving for $p_{ik}$ yields

$$p_{ik} = \frac{c_{ik}^{1/(1-s)}}{\sum_{k=1}^{K} c_{ik}^{1/(1-s)}}.$$

CIMPA-UCR

# C-Medias Difuso Bimodal

1. Choose initial $\mathbf{P}$ and $\mathbf{Q}$, which can be either crisp or fuzzy, and calculate $\mathbf{V}$ according to (12).

2. Repeat the following, until the decrease in $f_s(\mathbf{P},\mathbf{Q},\mathbf{V})$ is small.

   (a) Calculate $c_{ik} = \sum_{l=1}^{L}\sum_{j=1}^{m} q_{jl}^{s}(x_{ij} - v_{kl})^2$ and update $\mathbf{P}$ as $p_{ik} = c_{ik}^{1/(1-s)} / \sum_{k=1}^{K} c_{ik}^{1/(1-s)}$.

   (b) Calculate $d_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{n} p_{ik}^{s}(x_{ij} - v_{kl})^2$ and update $\mathbf{Q}$ as $q_{ik} = d_{jl}^{1/(1-s)} / \sum_{k=1}^{K} d_{jl}^{1/(1-s)}$.

   (c) Update $\mathbf{V}$ according to (12).

This algorithm lowers the value of $f_s(\mathbf{P},\mathbf{Q},\mathbf{V})$ in each iteration, until convergence has been achieved. Therefore, the algorithm always converges to a saddle point or a local minimum, which may or may not be a global minimum.

# Pasos difusos

1. Choose an initial value of $s$, a fuzzy step size $\gamma < 1$, and a threshold value $s_{min}$.

2. Choose initial $\mathbf{P}_0$ and $\mathbf{Q}_0$ and calculate $\mathbf{V}_0$ according to (12). The initial $\mathbf{P}_0$ and $\mathbf{Q}_0$ can be either crisp or fuzzy.

3. Repeat the following while $s > s_{min}$.

   (a) Perform the two-mode fuzzy $c$-means algorithm starting with $\mathbf{P}_0$, $\mathbf{Q}_0$, and $\mathbf{V}_0$. The results are in $\mathbf{P}_1$, $\mathbf{Q}_1$, and $\mathbf{V}_1$.

   (b) Set $s = 1 + \gamma(s - 1)$ and set $\mathbf{P}_0 = \mathbf{P}_1$, $\mathbf{Q}_0 = \mathbf{Q}_1$, and $\mathbf{V}_0 = \mathbf{V}_1$.

4. Apply the two-mode $k$-means algorithm starting from $\mathbf{P}_0$, $\mathbf{Q}_0$, and $\mathbf{V}_0$.

CIMPA-UCR

# Algunos resultados

| | *K*-medias | | Inter Alter. | | Sob. Sim | |
|---|---|---|---|---|---|---|
| | # | $\tau$ | # | $\tau$ | # | $\tau$ |
| Cognac | 1000 | 5% | 500 | 10% | 150 | 85% |
| Cigarros 1 | 1000 | 12% | 500 | 35% | 150 | 80% |
| Cigarros 2 | 1000 | 8% | 500 | 40% | 150 | 100% |

$$r = m = 3$$

# Datos de sanitarios (82x25)

| *r=m* | Int.Alt. | k-means | SS | Fuzzy | BT |
|---|---|---|---|---|---|
| **2** | 0.2289 | 0.2289 | 0.2289 | 0.2289 | 0.2289 |
| **3** | 0.3234 | 0.3234 | 0.3234 | 0.3234 | 0.3234 |
| **4** | 0.3729 | 0.3729 | 0.3804 | 0.3703 | 0.3804 |
| **5** | 0.4293 | 0.4293 | 0.4293 | 0.4293 | 0.4293 |
| **6** | 0.4635 | 0.4579 | 0.4651 | 0.4641 | 0.4651 |

CIMPA-UCR

# Estudio de Simulación

- Simulate **P**, **Q**, **V**, **E**
- Generate $\mathbf{X} = \mathbf{P}\mathbf{V}\mathbf{Q}^t + \mathbf{E}$
- 4 factors with 3 levels each: 81 combinations
- We simulate 50 data sets for each combination
- All 5 methods are performed for each data set
- Computation time is balanced

# Factores

1. Size of data: ($n=m=60$; $n=150, m=30$; $n=m=120$)
2. Number of clusters ($K=L=3$; $K=L=5$, $K=L=7$)
3. Size of error perturbations: elements of **E** are independent normally distributed with mean 0 and sta dev 0.5, 1 or 2
4. Distribution of the objects over the clusters:
   - Objects divided over the clusters with = probab
   - One cluster contains 10% of the objects
   - One cluster contains 60% of the clusters

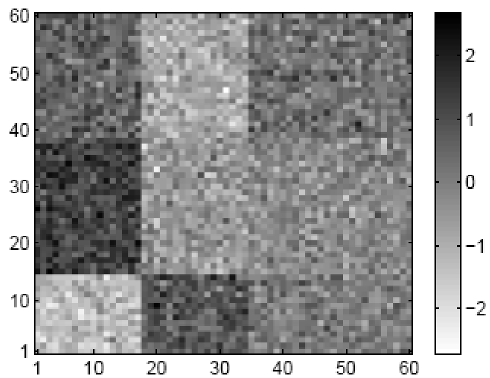
# Centros

- Assign randomly the numbers

$$\Phi\left(\frac{i}{K \times L+1}\right), i = 1, \ldots, K \times L$$

  to the elements of **V**, where $\Phi(.)$ is the inverse standard normal cumulative distribution function
- As a result, the elements of **V** appear standard normally distributed, and a fixed minimum distance between the cluster centers is ensured

# Representación Visual

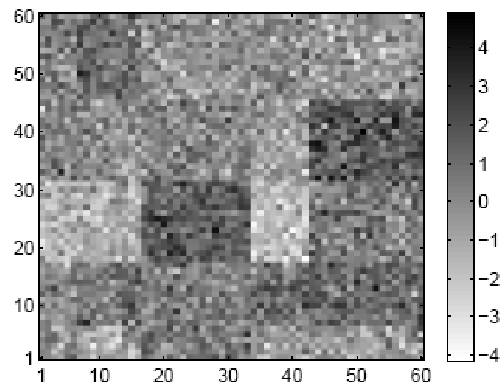



$$n = m = 60,$$
$$K = L = 3,$$

error s.d.=0.5

unif dist over clusters

*Easy to recognize*

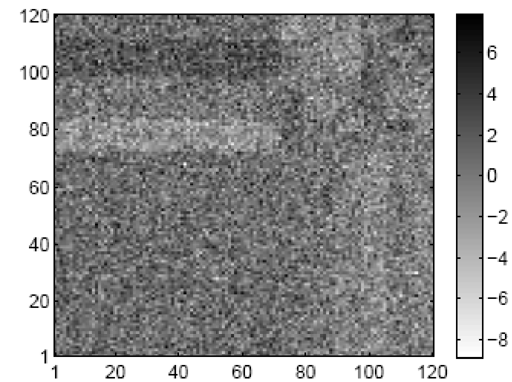



$$n = m = 60$$
$$K = L = 5$$

error s.d.=1

10% of obj in 1 cluster

*Difficult to recognize*




$$n = m = 120,$$
$$K = L = 7$$

error s.d.=2

60% of obj in 1 cluster

*Impossible to recognize*

# Parámetros

- Alt.Exch.: 30 times for every simulated data set
- Two-mode K-Means: 500 times
- Sim.Annealing: $T = 1$, $\gamma = 0.95$, $t_{max} = 10$, $R=(n(K-1)+m(L-1))/4$
- Tabu S.: number of iterations $S=3(n(K-1)+m(L-1))^{1/2}$, length of tabu list = $S/3$
- Fuzzy steps: $s_0 = 1.05$, $\gamma=0.85$, $s_{min} = 1.001$, run 20 times

# Criterios de comparación

- Variance accounted for *(VAF)*
- Adjusted Rand index *(ARI)*
- Average CPU time

CIMPA-UCR

# Variance accounted for

$$\mathrm{VAF} = 1 - \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ik}q_{jl}(x_{ij} - v_{kl})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x})^2}$$

where $\bar{x} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}$

$$VAF \in [0,1]$$

$$VAF \rightarrow \max$$

CIMPA-UCR

# Resultados de Simulación

Table 1: Average VAF for data sets with different error variances and numbers of rows and columns.

| Size of data set | $n = 60, m = 60$ | | | $n = 150, m = 30$ | | | $n = 120, m = 120$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Error st. dev. | .5 | 1 | 2 | .5 | 1 | 2 | .5 | 1 | 2 |
| Alternating exchanges | .7260 | .4128 | .1685 | .7288 | .4140 | .1750 | .7251 | .4058 | .1532 |
| Two-mode $k$-means | *.7283* | *.4134* | .1667 | *.7322* | *.4146* | .1733 | *.7293* | *.4082* | *.1535* |
| Simulated annealing | .7092 | .4068 | .1668 | .7166 | .4078 | .1739 | .7108 | .3987 | .1518 |
| Tabu search | .6823 | .4003 | .1642 | .6904 | .4015 | .1660 | .6703 | .3853 | .1496 |
| Fuzzy steps | .7248 | .4129 | *.1688* | .7299 | .4143 | *.1755* | .7199 | .4064 | .1531 |
| Original partition | .7292 | .4126 | .1540 | .7326 | .4122 | .1531 | .7308 | .4088 | .1515 |

CIMPA-UCR

# Resultados de Simulación

Table 2: Average VAF for data sets with different error variances and numbers of clusters.

| Numbers of clusters | $K=3,\ L=3$ | | | $K=5,\ L=5$ | | | $K=7,\ L=7$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Error st. dev. | .5 | 1 | 2 | .5 | 1 | 2 | .5 | 1 | 2 |
| Alternating exchanges | .6843 | .3558 | .1294 | .7399 | .4260 | *.1725* | .7557 | .4508 | .1948 |
| Two-mode $k$-means | *.6843* | *.3558* | *.1296* | *.7440* | *.4276* | .1724 | *.7614* | *.4529* | .1915 |
| Simulated annealing | .6804 | .3525 | .1284 | .7221 | .4186 | .1706 | .7340 | .4422 | .1935 |
| Tabu search | .6584 | .3476 | .1259 | .6868 | .4066 | .1663 | .6980 | .4329 | .1876 |
| Fuzzy steps | .6838 | .3558 | .1288 | .7394 | .4264 | .1725 | .7515 | .4515 | *.1960* |
| Original partition | .6843 | .3549 | .1214 | .7442 | .4265 | .1601 | .7640 | .4522 | .1771 |

CIMPA-UCR

# Resultados de Simulación

Table 3: Average VAF for data sets with different error variances and distributions of the objects over the clusters.

| Object distribution | uniform distribution | | | 10% in one cluster | | | 60% in one cluster | | |
|---|---|---|---|---|---|---|---|---|---|
| Error st. dev. | .5 | 1 | 2 | .5 | 1 | 2 | .5 | 1 | 2 |
| Alternating exchanges | .7460 | .4293 | .1695 | .7330 | .4171 | .1666 | .7009 | .3862 | .1606 |
| Two-mode $k$-means | *.7467* | *.4294* | .1683 | *.7336* | *.4174* | .1655 | *.7094* | *.3894* | .1596 |
| Simulated annealing | .7352 | .4255 | .1684 | .7199 | .4105 | .1650 | .6814 | .3774 | .1593 |
| Tabu search | .6858 | .4127 | .1641 | .6818 | .3993 | .1606 | .6755 | .3751 | .1552 |
| Fuzzy steps | .7419 | .4291 | *.1698* | .7299 | .4170 | *.1668* | .7029 | .3875 | *.1608* |
| Original partition | .7467 | .4288 | .1602 | .7337 | .4167 | .1566 | .7121 | .3881 | .1418 |

CIMPA-UCR

# Índice Ajustado de Rand

$$\text{ARI} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R}\binom{a_{ij}}{2} - \sum_{i=1}^{R}\binom{a_{i.}}{2}\sum_{j=1}^{R}\binom{a_{.j}}{2}/\binom{a}{2}}{[\sum_{i=1}^{R}\binom{a_{i.}}{2} + \sum_{j=1}^{R}\binom{a_{.j}}{2}]/2 - \sum_{i=1}^{R}\binom{a_{i.}}{2}\sum_{j=1}^{R}\binom{a_{.j}}{2}/\binom{a}{2}}$$

where $a_{ij}$ is the number of elements of $\mathbf{X}$ that simultaneously belong to cluster $i$ of the original partition and cluster $j$ of the retrieved partition, $R$ is the total number of clusters, $a_{i.} = \sum_{j=1}^{R} a_{ij}$, $a_{.j} = \sum_{i=1}^{R} a_{ij}$, and $a = \sum_{i=1}^{R}\sum_{j=1}^{R} a_{ij} = nm$. Here we consider a pair of elements to be in the same cluster only if they belong to same row cluster and to the same column cluster, so that $R = K \times L$.

Table 4: Average adjusted Rand indices for each level of each factor.

| Size of data set | $n = 60,\ m = 60$ | $n = 150,\ m = 30$ | $n = 120,\ m = 120$ |
|---|---|---|---|
| Alternating exchanges | .739 | .721 | .846 |
| Two-mode $k$-means | *.759* | *.748* | *.893* |
| Simulated annealing | .678 | .673 | .793 |
| Tabu search | .636 | .613 | .729 |
| Fuzzy steps | .745 | .737 | .851 |
| Numbers of clusters | $K = 3,\ L = 3$ | $K = 5,\ L = 5$ | $K = 7, L = 7$ |
| Alternating exchanges | .865 | .772 | .669 |
| Two-mode $k$-means | *.868* | *.815* | *.718* |
| Simulated annealing | .830 | .702 | .612 |
| Tabu search | .775 | .637 | .567 |
| Fuzzy steps | .858 | .783 | .692 |
| Error standard deviation | .5 | 1 | 2 |
| Alternating exchanges | .915 | .853 | *.538* |
| Two-mode $k$-means | *.974* | *.889* | .537 |
| Simulated annealing | .852 | .788 | .503 |
| Tabu search | .782 | .742 | .454 |
| Fuzzy steps | .934 | .863 | .537 |
| Distribution of the objects | uniform distribution | 10% in one cluster | 60% in one cluster |
| Alternating exchanges | *.874* | *.861* | .571 |
| Two-mode $k$-means | .873 | .860 | *.668* |
| Simulated annealing | .837 | .805 | .501 |
| Tabu search | .741 | .727 | .510 |
| Fuzzy steps | .868 | .856 | .610 |

# Tiempo de CPU

Table 5: Average CPU times of all optimization methods in seconds.

| Size of data sets | $n = 60$, $m = 60$ | | | $n = 150$, $m = 30$ | | | $n = 120$, $m = 120$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Numbers of clusters $(K, L)$ | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 |
| Alternating exchanges | 1.21 | 2.54 | 3.96 | 1.92 | 4.22 | 6.43 | 2.39 | 5.59 | 8.72 |
| Two-mode $k$-means | 1.54 | 2.36 | 3.09 | 1.96 | 3.58 | 4.57 | 2.50 | 5.31 | 7.66 |
| Simulated annealing | 1.08 | 2.30 | 3.54 | 1.75 | 3.82 | 5.78 | 2.42 | 5.47 | 8.49 |
| Tabu search | 0.82 | 2.16 | 3.74 | 1.47 | 3.82 | 6.77 | 2.34 | 6.04 | 10.75 |
| Fuzzy steps | 0.99 | 1.91 | 2.66 | 1.05 | 2.48 | 3.47 | 2.22 | 7.92 | 11.28 |

CIMPA-UCR

# Conclusiones

- Error variance and relative sizes of the clusters strongly affect how well a clusering structure can be retrieved
- Two-mode *k*-means most often had the best performance, especially in favorable conditions
- When the optimal partition was hard to find, the AE and FS methods often performed well
- Sim. Annealing was also fairly well
- Tabu search had an inferior performance