# Aprendizaje no supervisado

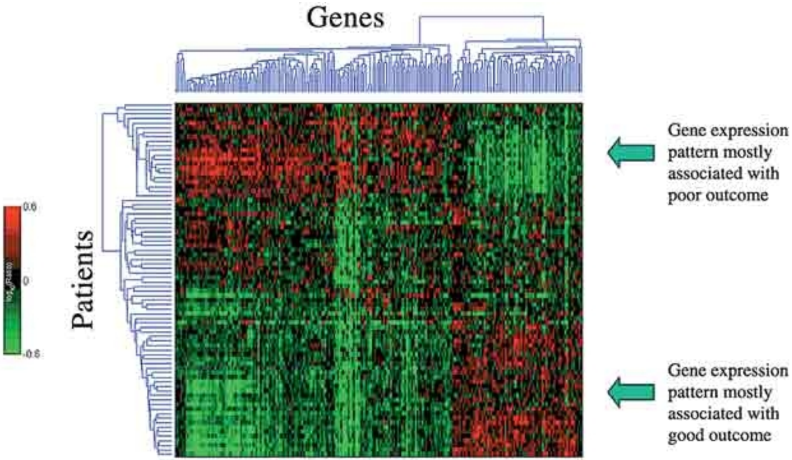
# Coagrupamiento o biclustering

Javier Sevilla

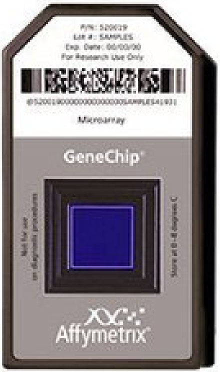
Universidad Internacional de Valencia

17.03.2020

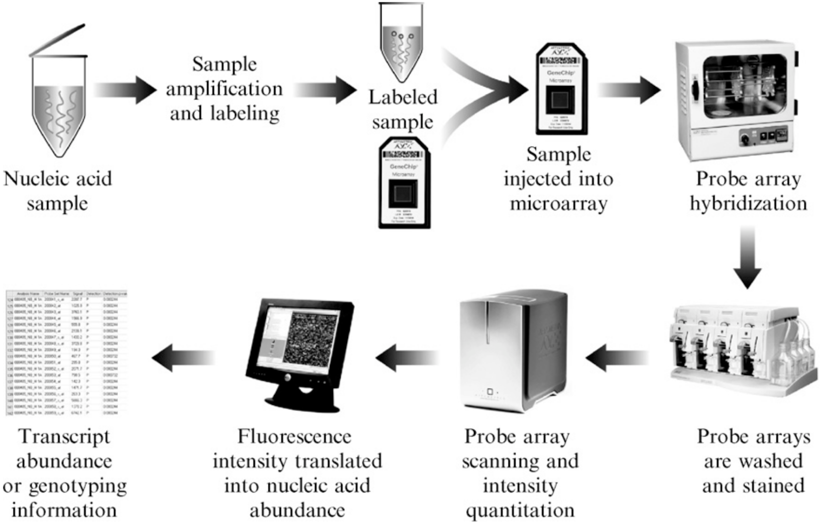
# Microarray data



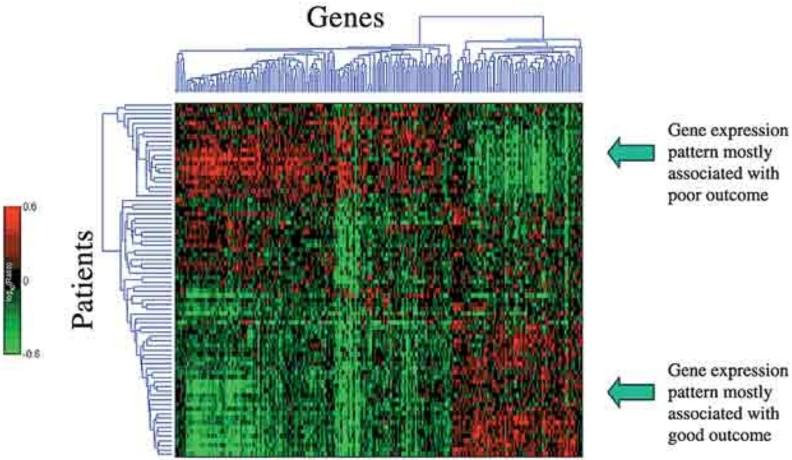
# Microarray data



# Microarray data



# Microarray data



# Microarray data

https://www.youtube.com/watch?v=0ATUjAxNf6U

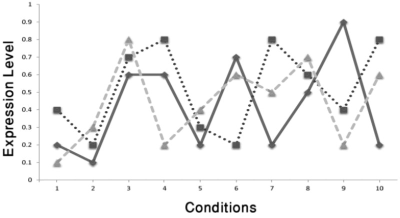
# Agrupamiento vs. coagrupamiento

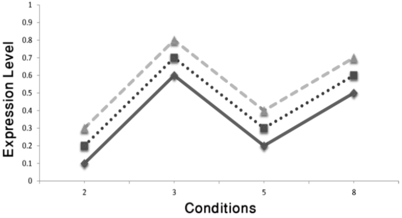
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

# Agrupamiento vs. coagrupamiento



## Coagrupamiento o biclustering

Técnicas que buscan clústeres tal que:

- Un clúster de ejemplos se define sólo mediante un subconjunto de variables

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

# Coagrupamiento o biclustering

## Coagrupamiento o biclustering

Técnicas que buscan clústeres tal que:

- Un clúster de ejemplos se define sólo mediante un subconjunto de variables
- Un clúster de variables se define sólo mediante un subconjunto de ejemplos

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

## Coagrupamiento o biclustering

Técnicas que buscan clústeres tal que:

- Un clúster de ejemplos se define sólo mediante un subconjunto de variables
- Un clúster de variables se define sólo mediante un subconjunto de ejemplos
- Los clústeres no son, respecto a ejemplos y variables, ni exclusivos ni exhaustivos

  Un ejemplo puede pertenecer a uno, varios o ningún clúster

  Una variable puede relacionarse con uno, varios o ningún clúster

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

- $X$ Conjunto de datos de entrenamiento (matriz)
- $n$ Número de ejemplos
- $v$ Número de variables
- $I_{n'}$ Vector de índices de longitud $n'$

Así, definimos una **sub-matriz** $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

$X$ Conjunto de datos de entrenamiento (matriz)

$n$ Número de ejemplos

$v$ Número de variables

$I_{n'}$ Vector de índices de longitud $n'$

Así, definimos una **sub-matriz** $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

Una **sub-matriz** es un **clúster** si los valores que agrupa siguen un cierto patrón

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

Clúster $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

- **Biclústeres constantes**: todos los valores de la sub-matriz son iguales

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

Clúster $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

- **Biclústeres constantes**: todos los valores de la sub-matriz son iguales
- **Biclústeres con filas/columnas constantes**: todos los valores de una fila/columnas son iguales pero los de diferentes filas/columnas son diferentes

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

Clúster $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

- **Biclústeres constantes**: todos los valores de la sub-matriz son iguales
- **Biclústeres con filas/columnas constantes**: todos los valores de una fila/columnas son iguales pero los de diferentes filas/columnas son diferentes
- **Biclústeres basados en patrones (aditivos o multiplicativos)**: las diferencias o el ratio entre los elementos de dos filas o columnas cualquiera son constantes

# Patrones habitualmente buscados en biclustering

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

| | | |
|---|---|---|
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |
| 5 | 7 | 9 |

| | | |
|---|---|---|
| 2 | 2 | 2 |
| 4 | 4 | 4 |
| 6 | 6 | 6 |
| 8 | 8 | 8 |
| 10 | 10 | 10 |
| 12 | 12 | 12 |

| | | |
|---|---|---|
| 2 | 5 | 4 |
| 5 | 8 | 7 |
| 3 | 6 | 5 |
| 6 | 9 | 8 |
| 4 | 7 | 6 |
| 7 | 10 | 9 |

| | | |
|---|---|---|
| 2 | 6 | 4 |
| 4 | 12 | 8 |
| 5 | 15 | 10 |
| 8 | 24 | 16 |
| 3 | 9 | 6 |
| 6 | 18 | 12 |

| | | |
|---|---|---|
| 2.5 | 7 | 5 |
| 4 | 10 | 8.5 |
| 5 | 13 | 9 |
| 8 | 19 | 13 |
| 3 | 8 | 6.5 |
| 5.5 | 14 | 9 |

Clúster $X'$ dados $I'_n \subset I_n$ e $I'_v \subset I_v$ donde $X'_{ij} = X_{I_{n'}[i], I_{v'}[j]}$

- **Biclústeres constantes**: todos los valores de la sub-matriz son iguales
- **Biclústeres con filas/columnas constantes**: todos los valores de una fila/columnas son iguales pero los de diferentes filas/columnas son diferentes
- **Biclústeres basados en patrones (aditivos o multiplicativos)**: las diferencias o el ratio entre los elementos de dos filas o columnas cualquiera son constantes
- **Biclústeres con evoluciones coherentes**: más allá del valor exacto, busca biclústeres con comportamientos coherentes (aumentan o disminuyen a la vez) por filas, columnas o ambas.

## Biclustering como problema de optimización

- Algoritmo voraz
- Cada posible biclúster, un valor de credibilidad (¿es realmente un biclúster?)
- Busca submatrices **grandes** y **uniformes**
- Se asume (implícito) biclúster constantes, con la posibilidad de cierto comportamiento aditivo por filas y/o columnas

Valor medio sobre una fila de un posible biclúster (sub-matriz):

$$\bar{x}_{iI_{v'}} = \frac{1}{|I_{v'}|} \sum_{j \in I_{v'}} x_{ij}$$

Valor medio sobre una fila de un posible biclúster (sub-matriz):

$$\bar{x}_{iI_{v'}} = \frac{1}{|I_{v'}|} \sum_{j \in I_{v'}} x_{ij}$$

Valor medio de una columna:

$$\bar{x}_{I_{n'}j} = \frac{1}{|I_{n'}|} \sum_{i \in I_{n'}} x_{ij}$$

Valor medio sobre una fila de un posible biclúster (sub-matriz):

$$\bar{x}_{iI_{v'}} = \frac{1}{|I_{v'}|} \sum_{j \in I_{v'}} x_{ij}$$

Valor medio de una columna:

$$\bar{x}_{I_{n'}j} = \frac{1}{|I_{n'}|} \sum_{i \in I_{n'}} x_{ij}$$

Valor medio de la sub-matriz:

$$\bar{x}_{I_{n'},I_{v'}} = \frac{1}{|I_{n'}| \cdot |I_{v'}|} \sum_{i \in I_{n'}} \sum_{j \in I_{v'}} x_{ij}$$

# Coagrupamiento

## Cheng y Church (2000)

### Idea

Si se substrae el valor medio del biclúster, $\bar{x}_{I_{n'},I_{v'}}$, de la fila, $\bar{x}_{iI_{v'}}$, y de la columna, $\bar{x}_{I_{n'}j}$, el **valor residual** restante tendería a cero.

Valor residual de un punto $(i,j) \in (I_{n'}, I_{v'})$:

$$R_{I_{n'},I_{v'}}(i,j) = x_{ij} - \bar{x}_{I_{n'},I_{v'}} - \bar{x}_{iI_{v'}} - \bar{x}_{I_{n'}j}$$

Valor residual cuadrático medio de una sub-matriz:

$$\bar{R}_{I_{n'},I_{v'}} = \frac{1}{|I_{n'}| \cdot |I_{v'}|} \sum_{(i,j)\in(I_{n'},I_{v'})} R_{I_{n'},I_{v'}}(i,j)^2$$

## Idea

Si se substrae el valor medio del biclúster, $\bar{x}_{I_{n'},I_{v'}}$, de la fila, $\bar{x}_{iI_{v'}}$, y de la columna, $\bar{x}_{I_{n'}j}$, el **valor residual** restante tendería a cero.

Valor residual de un punto $(i,j) \in (I_{n'}, I_{v'})$:

$$R_{I_{n'},I_{v'}}(i,j) = x_{ij} - \bar{x}_{I_{n'},I_{v'}} - \bar{x}_{iI_{v'}} - \bar{x}_{I_{n'}j}$$

Valor residual cuadrático medio de una sub-matriz:

$$\bar{R}_{I_{n'},I_{v'}} = \frac{1}{|I_{n'}| \cdot |I_{v'}|} \sum_{(i,j)\in(I_{n'},I_{v'})} R_{I_{n'},I_{v'}}(i,j)^2$$

**Reformulación:** Buscar sub-matrices $X'$ cuyo valor residual medio no supere cierto umbral $\delta$ y que sean máximas

[problema NP-hard]

## Algoritmo

Dadas $X$ y $\delta$:

- Se elimina **iterativamente** la fila/columna que reduce el valor residual medio en mayor medida hasta que $\bar{R}_{I_{n'},I_{v'}} < \delta$
- Se incluye **iterativamente** una fila/columna previamente eliminada (la que menos incremente el valor residual medio) siempre que $\bar{R}_{I_{n'},I_{v'}} < \delta$

Converge a una submatriz $X'$ localmente máxima con $\bar{R}_{I_{n'},I_{v'}} < \delta$

## Algoritmo

Dadas $X$ y $\delta$:

- Se elimina **iterativamente** la fila/columna que reduce el valor residual medio en mayor medida hasta que $\bar{R}_{I_{n'},I_{v'}} < \delta$
- Se incluye **iterativamente** una fila/columna previamente eliminada (la que menos incremente el valor residual medio) siempre que $\bar{R}_{I_{n'},I_{v'}} < \delta$

Converge a una submatriz $X'$ localmente máxima con $\bar{R}_{I_{n'},I_{v'}} < \delta$

## Procedimiento para encontrar **un único** biclúster

Para encontrar uno nuevo, sustituir los valores de $X'$ por valor aleatorios y relanzar el algoritmo.

# Coagrupamiento

## Algoritmo de firma iterativa

### Idea

El valor medio de una fila/columna de un biclúster debería ser inusualmente alto o bajo en comparación con el valor medio del resto de la fila/columna

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $\cdots$ | $x_{1v}$ |
| $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $\cdots$ | $x_{2v}$ |
| $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $\cdots$ | $x_{3v}$ |
| $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ | $\cdots$ | $x_{4v}$ |
| $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ | $\cdots$ | $x_{5v}$ |
| $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ | $\cdots$ | $x_{6v}$ |
| $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ | $\cdots$ | $x_{7v}$ |
| $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ | $\cdots$ | $x_{8v}$ |
| $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ | $\cdots$ | $x_{9v}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $\cdots$ | $x_{nv}$ |

## Algoritmo

Dadas $X$ y $\epsilon$:

- Hace dos copias de $X$: $X^F$ y $X^C$
- Selección inicial de filas, $I_{n'}^{(0)}$
- **Iterativamente**, elegir unas columnas, $I_{v'}^{(t)}$, con respecto a las filas $I_{n'}^{(t-1)}$ y una nueva selección de filas $I_{n'}^{(t)}$ a partir de las columnas $I_{v'}^{(t)}$

El algoritmo para en la $T^{th}$ iteración que satisfaga:

$$\frac{|I_{n'}^{(T)} \backslash I_{n'}^{(T-1)}|}{|I_{n'}^{(T)} \cup I_{n'}^{(T-1)}|} < \epsilon$$

## Paso clave: actualización de $I_{n'}^{(t)}$ y $I_{v'}^{(t)}$

- Se escogen las columnas cuyo valor absoluto medio sobre todas las filas de $I_{n'}^{(t-1)}$ es mayor que un umbral $t_C$ por la desviación estándar $\sigma_C$ de los valores medios de todas las columnas de la matriz original
- Se escogen las filas cuyo valor absoluto medio sobre todas las columnas de $I_{v'}^{(t)}$ es mayor que un umbral $t_F$ por la desviación estándar $\sigma_F$ de los valores medios de todas las columnas de la matriz original

$>>$ Subconjuntos de filas/cols. con valores medios destacados con respecto al resto de filas/cols. (sólo para un subconjunto de cols./filas) $<<$

## Paso clave: actualización de $I_{n'}^{(t)}$ y $I_{v'}^{(t)}$

- Se escogen las columnas cuyo valor absoluto medio sobre todas las filas de $I_{n'}^{(t-1)}$ es mayor que un umbral $t_C$ por la desviación estándar $\sigma_C$ de los valores medios de todas las columnas de la matriz original
- Se escogen las filas cuyo valor absoluto medio sobre todas las columnas de $I_{v'}^{(t)}$ es mayor que un umbral $t_F$ por la desviación estándar $\sigma_F$ de los valores medios de todas las columnas de la matriz original

$>>$ Subconjuntos de filas/cols. con valores medios destacados con respecto al resto de filas/cols. (sólo para un subconjunto de cols./filas) $<<$

## Procedimiento para encontrar **un único** biclúster

Para encontrar uno nuevo, sustituir los valores de $X'$ por valor aleatorios y relanzar el algoritmo.

## Consideraciones

- Dependencia de la inicialización

  Ejecutar el algoritmo con diferentes conjuntos iniciales de filas, $I_{n'}^{(0)}$, puede ser una manera de obtener diferentes biclústeres

- Al menos 3 parámetros (umbrales) a fijar:

  $\epsilon$ Parámetro de parada-convergencia
  $t_F$ Selección de filas
  $f_C$ Selección de columnas

# Coagrupamiento

## Algoritmo de co-agrupamiento espectral

### Idea

Álgebra lineal para encontrar biclústeres en una matriz $X$ que tiene una estructura de bloques:

- Biclústeres de patrón multiplicativo por fila (realista)
- Detecta estructuras incluso con filas/columnas desordenadas

| | | | | | |
|---|---|---|---|---|---|
| 1 | 1 | 11 | 11 | 3 | 3 |
| 1 | 1 | 11 | 11 | 3 | 3 |
| 4 | 4 | 7 | 7 | 12 | 12 |
| 4 | 4 | 7 | 7 | 12 | 12 |

- La estructura de bloques de $X$ se refleja en los vectores propios de $XX^T$ y $X^TX$
- Los valores propios de $XX^T$ y de $X^TX$ son los mismos:
  $f = Xe$, donde $e$ es un vector propio de $X^TX$, $f$ lo es de $XX^T$ y ambos tienen el mismo valor propio

| | | | | | |
|---|---|---|---|---|---|
| 1 | 1 | 11 | 11 | 3 | 3 |
| 1 | 1 | 11 | 11 | 3 | 3 |
| 4 | 4 | 7 | 7 | 12 | 12 |
| 4 | 4 | 7 | 7 | 12 | 12 |

$$e = (p, p, q, q, o, o)^T \text{ y } f = (r, r, s, s)^T$$

## Algoritmo

- Normalizar $X$: $\check{X} = F^{-1/2} X C^{-1/2}$
  Con $F$: $(n \times n)$-matriz diagonal con $F_{ii} = \sum_{j \in \{1,\dots,v\}} x_{ij}$
  y $C$: $(v \times v)$-matriz diagonal con $C_{jj} = \sum_{i \in \{1,\dots,n\}} x_{ij}$
- Obtener la descomposición en valores singulares: $\check{X} = A \Sigma B^T$
- Clustering por **filas**:
  - Seleccionar los $p$ mejores vectores propios de $B$ para construir la $(v \times p)$-matriz $B'$
  - Proyectar $X$ en ese espacio $p$-dimensional: $\check{X} B'$
  - Aplicar $K$-means a esa matriz resultante
  - Las filas asignadas por $K$-means al mismo clúster pertenecen al mismo bloque de $X$
- Clustering por **columnas**:
  - Mismo procedimiento usando $A$ y $\check{X}^T A'$

## Consideraciones

- Gran número de parámetros: $p$
- Se basa en $K$-means
- El número de $K$ en filas y columnas puede ser diferente

### Idea

Aplicar clustering sobre las filas (matriz $X$) y las columnas (matriz $X^T$) de manera iterativa y jerárquica

Cada clúster de filas se obtendría a partir de un clúster previo de columnas y viceversa.

Estrategia general: permite usar cualquier algoritmo de clustering estándar

## Algoritmo

Dada la matriz $X$ y un algoritmo de clustering estándar

- Se guarda un registro de particiones de filas $F$ y columnas $C$

  Se parte de una partición única de filas en $F$ y de columnas en $C$

- **Iterativamente**:
  - Definir una sub-matriz $X'$ con una partición de $F$ y otra de $C$
  - Aplicar el algoritmo de clustering estándar a $X'$ y a $(X')^T$
  - Añadir a $F$ y $C$ los clústeres *estables* encontrados

    Se mantiene la traza de la jerarquía a través de la cual han surgido los respectivos (sub)clústeres

- Parar cuando no se detectan más clústeres estables

** Una correcta implementación implica asegurarse de que un par de particiones de $F$ y $C$ se considera sólo una vez.

## Consideraciones

- Rendimiento fuertemente dependiente del rendimiento del algoritmo estándar
- Importancia de la medida de estabilidad considerada para detectar clústeres relevantes
- La parametrización de los algoritmos básicos puede dificultar su integración con esta estrategia
- La naturaleza jerárquica de los resultados de esta estrategia suele aportar **información relevante**

# Gracias

viu Universidad Internacional de Valencia