

Aprendizaje no Supervisado

Biclustering

Manuel Sánchez-Montañés
Universidad Autónoma de Madrid
manuel.smontanes@gmail.com

Biclustering

$$X_{(N,M)} = \begin{matrix} & \text{ATTRIBUTES} & \\ & \begin{matrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N1} & \dots & x_{Nj} & \dots & x_{NM} \end{matrix} & \\ \text{OBJECTS} & \end{matrix}$$

🌐 “Objetos”: por ejemplo, clientes

🌐 “Atributos”: sus datos

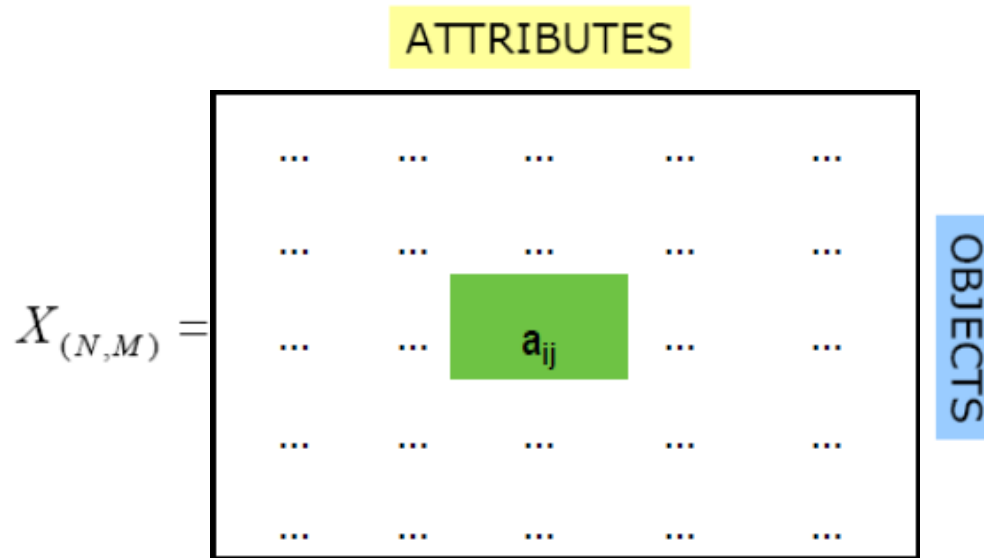
Biclustering

$$X_{(N,M)} = \begin{matrix} & \text{ATTRIBUTES} \\ \begin{matrix} \text{OBJECTS} \\ \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N1} & \dots & x_{Nj} & \dots & x_{NM} \end{bmatrix} \end{matrix} \end{matrix}$$

 Biclustering: Clustering **simultáneo de filas y columnas** de una matriz de datos

- Clustering: Identifica grupos of objetos con valores similares en **todos los atributos**
- Biclustering: Identifica grupos of objetos con valores similares en solo **un subconjunto de atributos** (¿Qué subconjunto es ese?)

Biclustering








 Biclustering: Clustering **simultáneo de filas y columnas** de una matriz de datos

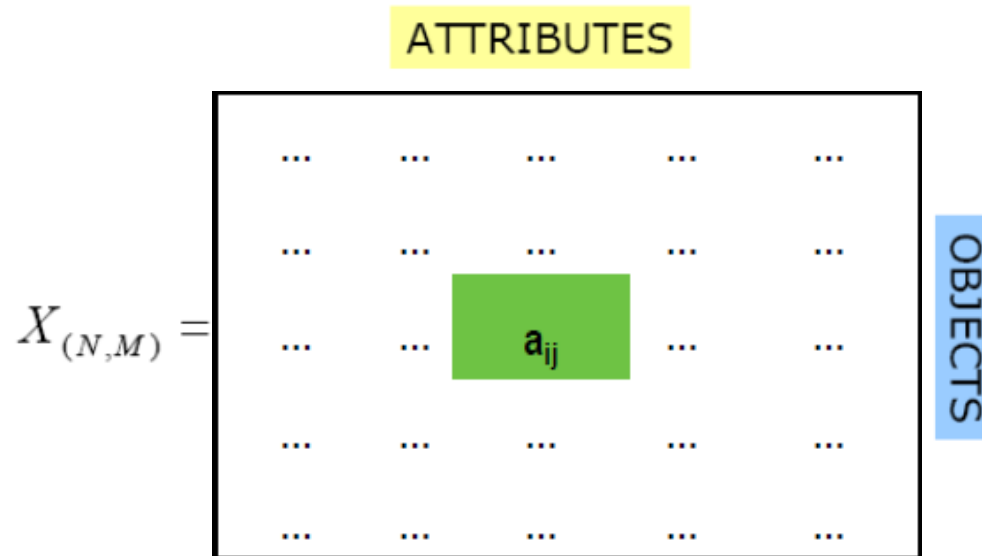
- Clustering: Identifica grupos of objetos con valores similares en **todos los atributos**
- Biclustering: Identifica grupos of objetos con valores similares en solo un **subconjunto de atributos** (¿Qué subconjunto es ese?)

Biclustering: también conocido como “co-clustering”, “block clustering”, etc.

Tipos de Biclustering

-  Biclusters con valores constantes
-  Biclusters con valores constantes en las filas
-  Biclusters con valores constantes en las columnas
-  Biclusters con valores coherentes
-  Biclusters con evoluciones coherentes

Biclusters con valores constantes

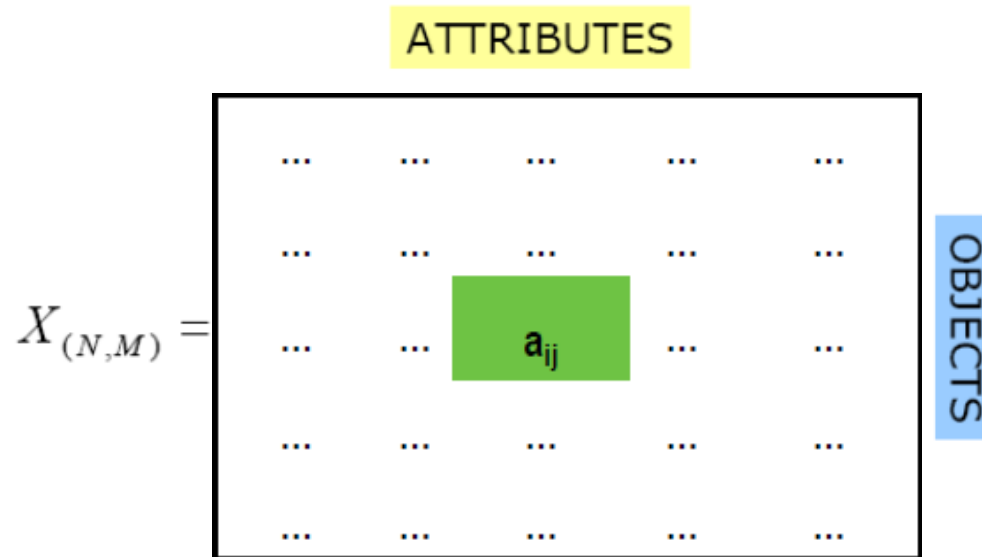


 Bicluster constant perfecto: los valores a_{ij} en cada bicluster son constantes. Por ejemplo:

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

Figuras de “(Clustering and) Biclustering Gene Expression Data” (Sara Madeira 2011)

Biclusters con filas constantes



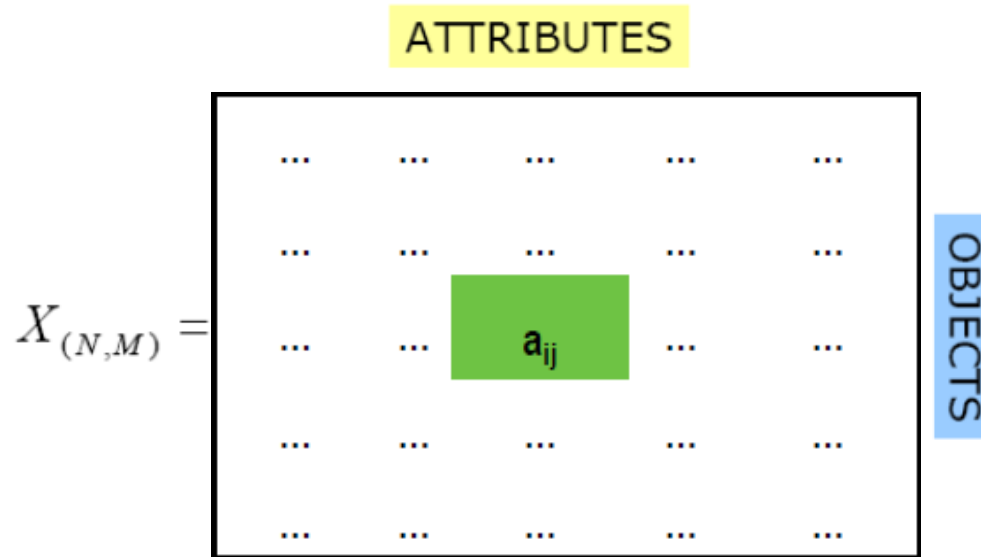
- Bicluster perfecto: los valores dentro del bicluster se pueden obtener usando:

$$a_{ij} = \mu + \alpha_i$$

$$a_{ij} = \mu \times \alpha_i$$

- El ajuste puede ser aditivo o multiplicativo

Biclusters con columnas constantes



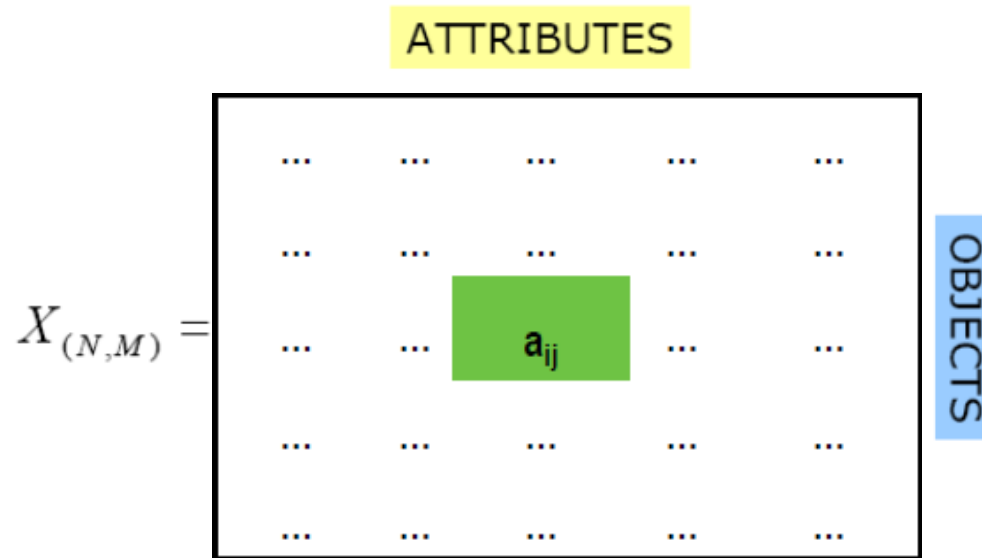
- Bicluster perfecto: los valores dentro del bicluster se pueden obtener usando:

$$a_{ij} = \mu + \beta_j$$

$$a_{ij} = \mu \times \beta_j$$

- El ajuste puede ser aditivo o multiplicativo

Biclusters con filas/columnas constantes



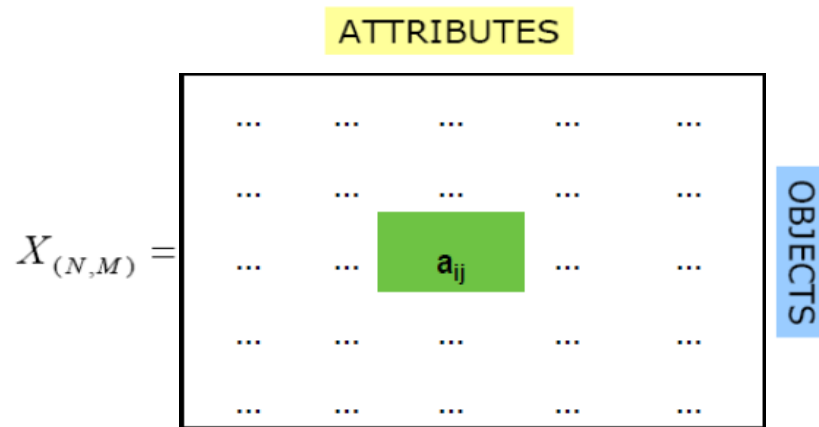
1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

Filas constantes

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

Columnas constantes

Biclusters con valores coherentes



$$a_{ij} = \mu + \alpha_i + \beta_j$$

$$a_{ij} = \mu \times \alpha_i \times \beta_j$$

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

Modelo aditivo

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

Modelo
multiplicativo

Biclusters con valores coherentes

General Additive Model

1.0	1.0	1.0	1.0		
2.0	2.0	2.0	2.0		
3.0	3.0	8.0	8.0	5.0	5.0
4.0	4.0	10	10	6.0	6.0
		7.0	7.0	7.0	7.0
		8.0	8.0	8.0	8.0

Constant Rows

1.0	2.0	3.0	4.0		
1.0	2.0	3.0	4.0		
1.0	2.0	8.0	10	7.0	8.0
1.0	2.0	8.0	10	7.0	8.0
		5.0	6.0	7.0	8.0
		5.0	6.0	7.0	8.0

Constant Columns

Biclusters con valores coherentes

General Additive Model

1.0	2.0	5.0	0.0		
2.0	3.0	6.0	3.0		
4.0	5.0	10	7.0	1.0	3.0
5.0	6.0	11	9.0	2.0	4.0
		5.0	7.0	4.0	6.0
		7.0	9.0	6.0	8.0

Coherent Values

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

2.0	4.0	1.0	3.0
3.0	5.0	2.0	4.0
5.0	7.0	4.0	6.0
7.0	9.0	6.0	8.0

Additive Model

Biclusters con evoluciones coherentes

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

Overall Coherent
Evolution

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

Coherent Evolution
On the Rows

Biclusters con evoluciones coherentes

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

Coherent Evolution
On the Columns

70	13	19	10
49	40	49	35
40	20	27	15
90	15	20	12

Order Preserving
Sub-Matrix (OPSM)

Biclusters con evoluciones coherentes

General Additive Model

1.0	2.0	5.0	0.0		
2.0	3.0	6.0	3.0		
4.0	5.0	10	7.0	1.0	3.0
5.0	6.0	11	9.0	2.0	4.0
		5.0	7.0	4.0	6.0
		7.0	9.0	6.0	8.0

Coherent Values

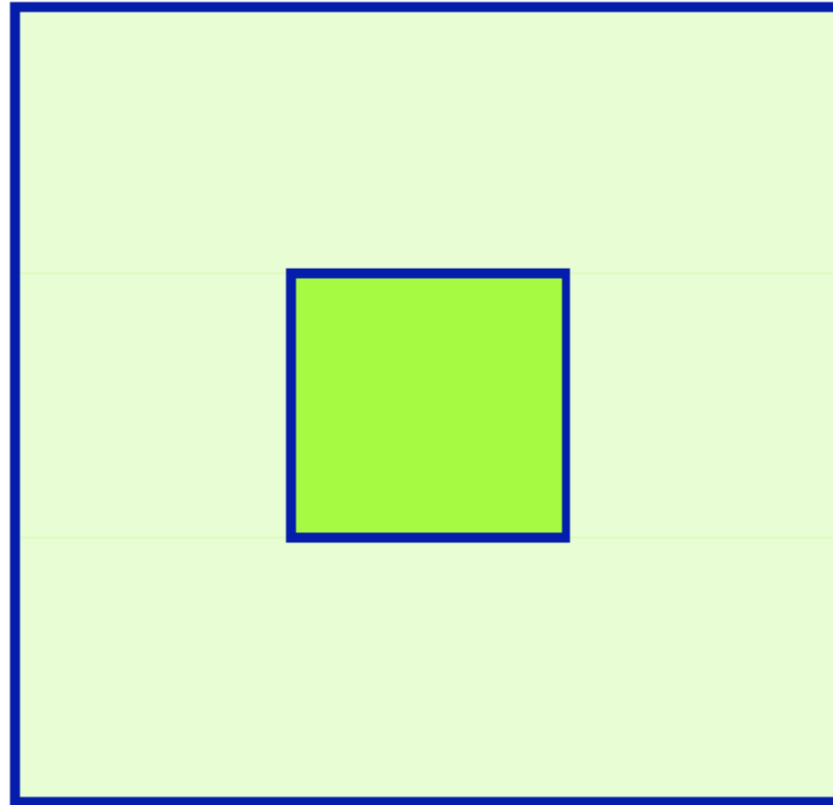
1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

2.0	4.0	1.0	3.0
3.0	5.0	2.0	4.0
5.0	7.0	4.0	6.0
7.0	9.0	6.0	8.0

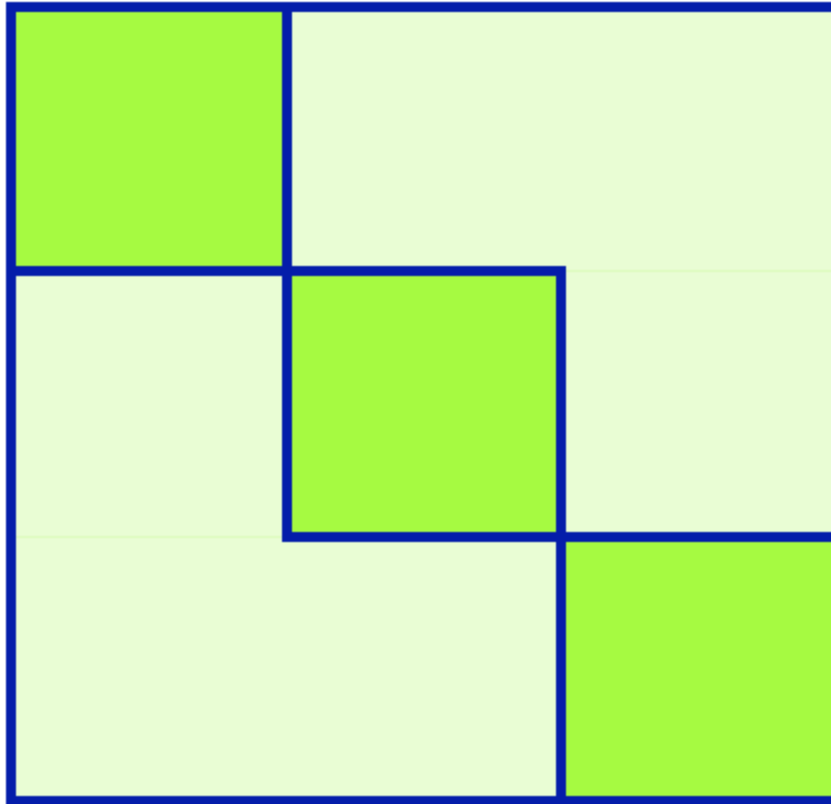
Additive Model

Estructura del Biclustering

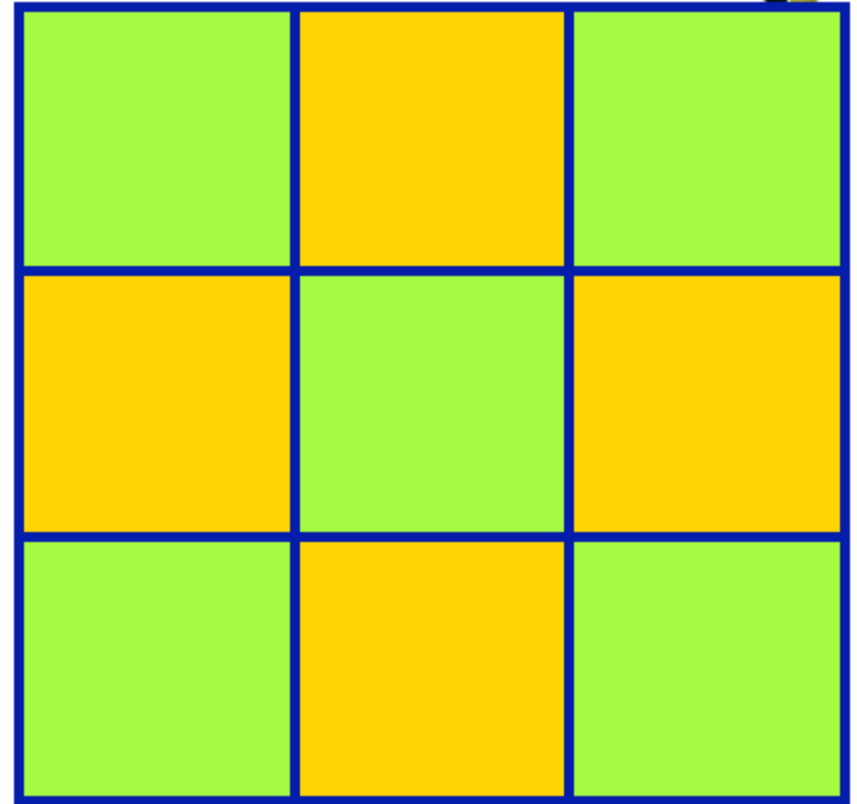
Un bicluster



Varios biclusters (I)

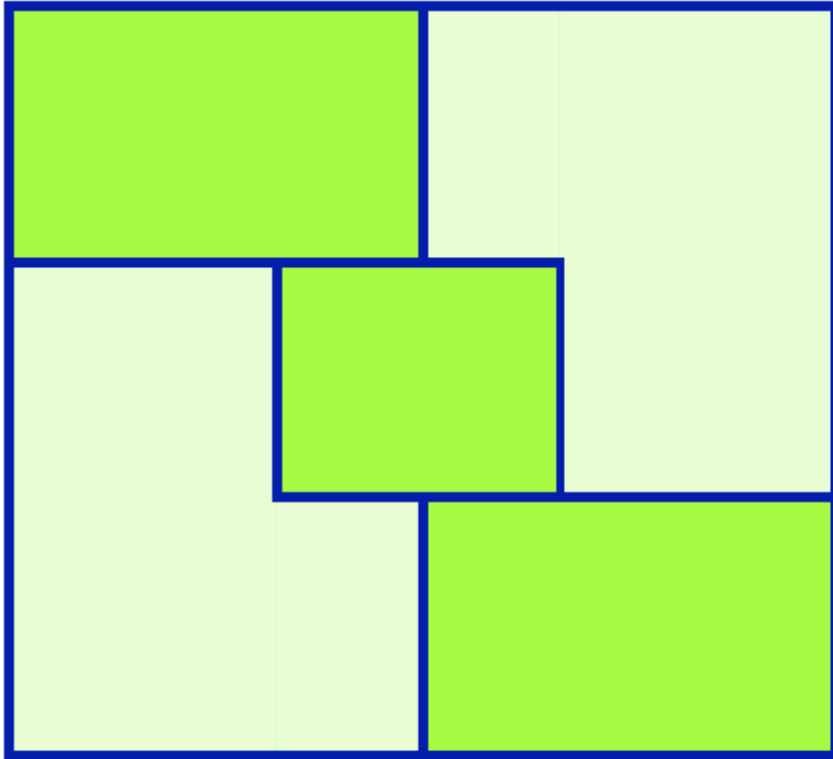


Exclusive Row and Column
Biclusters

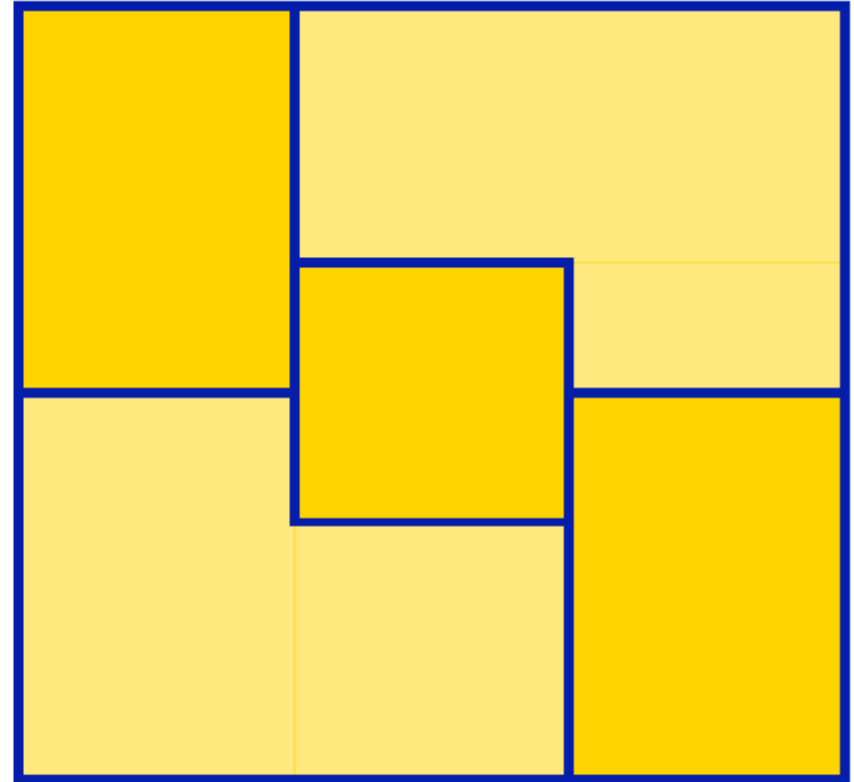


Checkerboard Structure

Varios biclusters (II)

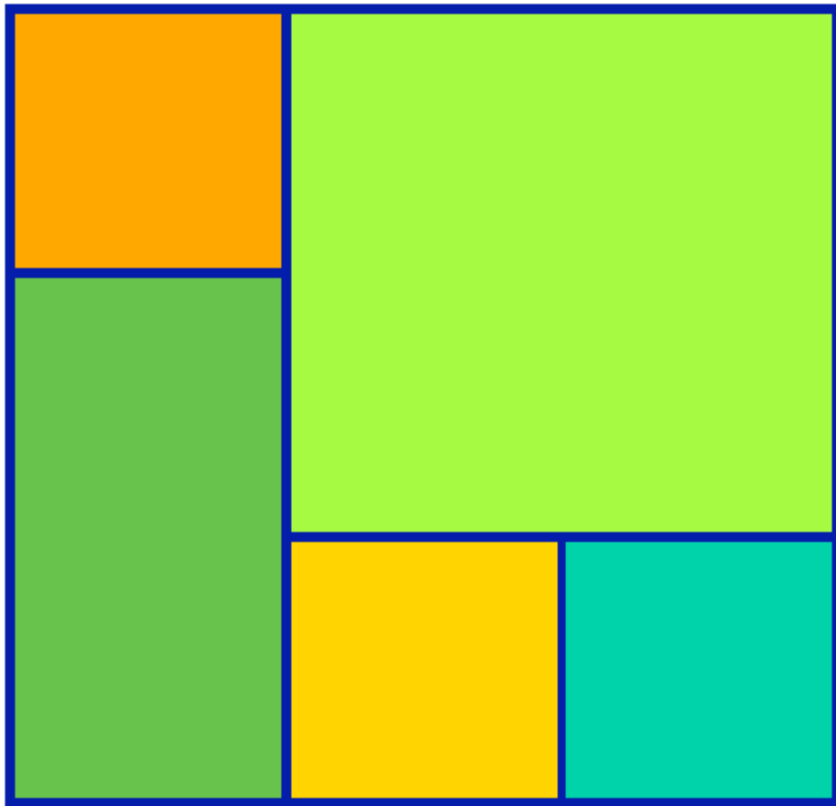


Exclusive-Rows Biclusters

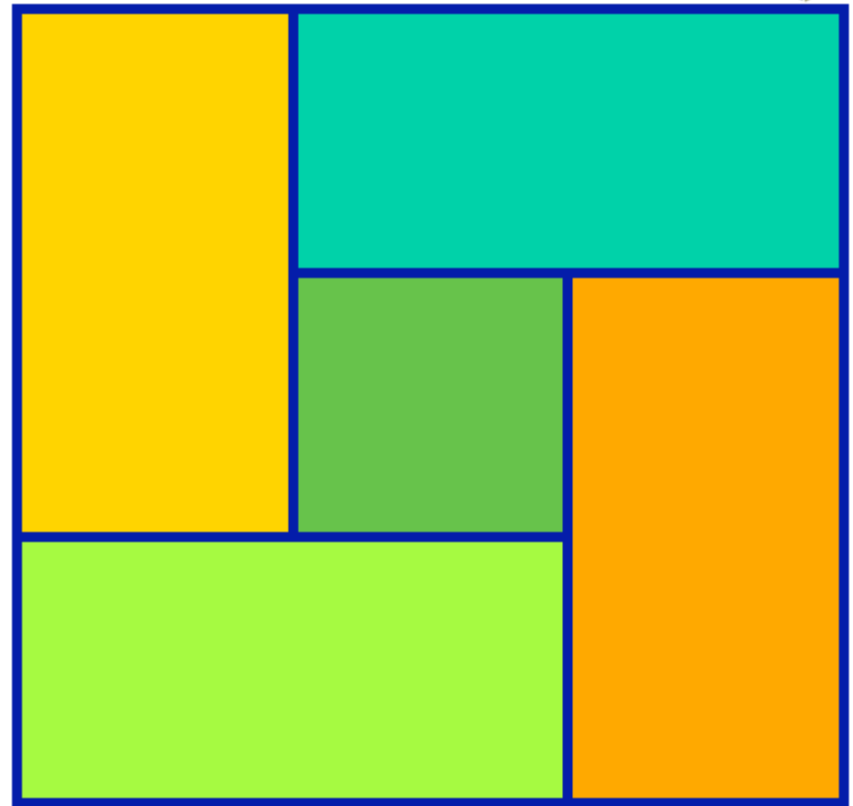


Exclusive-Columns Biclusters

Varios biclusters (III)

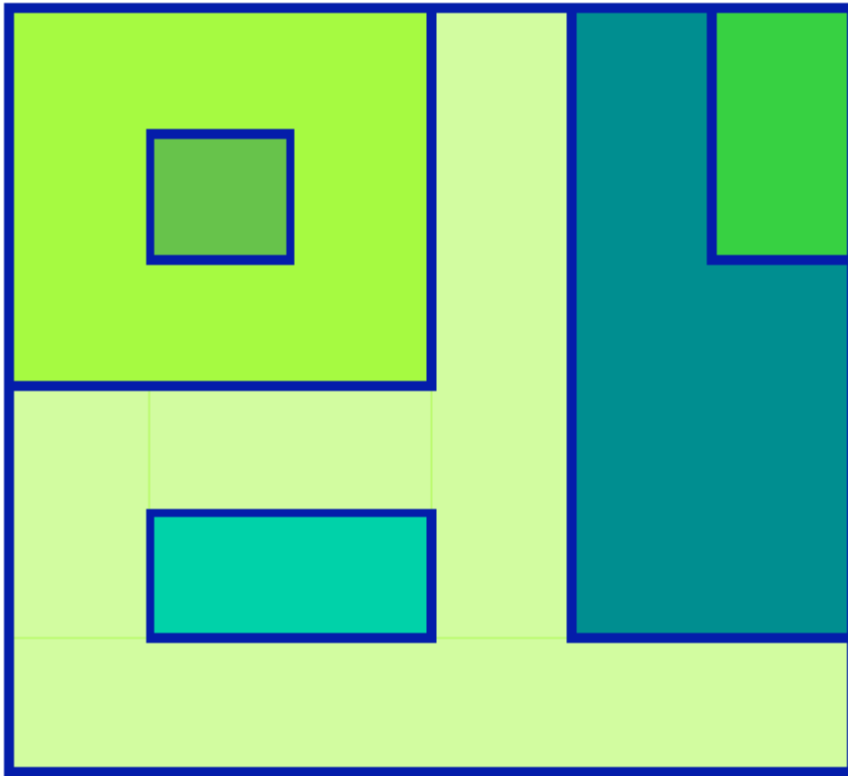


Non-Overlapping Biclusters
with Tree Structure

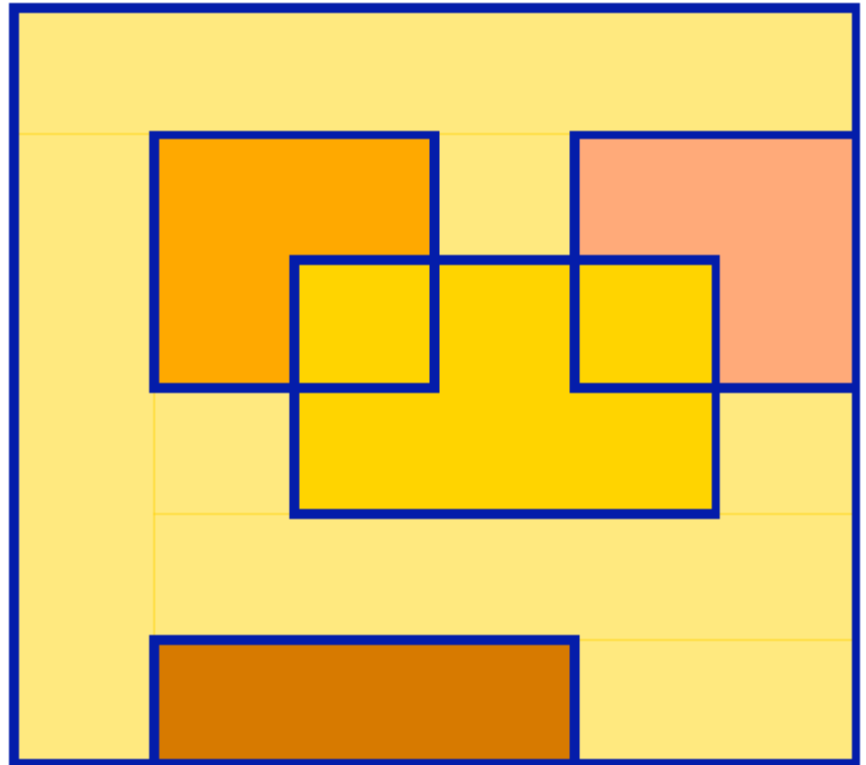


Non-Overlapping Non-Exclusive
Biclusters

Varios biclusters (IV)



Overlapping Biclusters with
Hierarchical Structure



Arbitrarily Positioned
Overlapping Biclusters

Co-clustering espectral

- Particiona filas y columnas bajo la suposición de que los datos tienen una estructura subyacente de tablero de ajedrez
- El algoritmo encuentra biclusters con valores superiores a los de las otras filas y columnas correspondientes
- Cada fila y cada columna pertenecen exactamente a un bicluster, por lo que reorganizar las filas y columnas para hacer que las particiones contiguas revelen estos valores altos a lo largo de la diagonal (cantidad de clusters)
- En sklearn (Python):

```
class sklearn.cluster.bicluster. SpectralCoclustering (n_clusters=3, svd_method='randomized',  
n_svd_vecs=None, mini_batch=False, init='k-means++', n_init=10, n_jobs=1, random_state=None)
```

Biclustering: Preprocesado

- Todos los datos deben ser numéricos
- Normalizar / estandarizar
- Funciona mejor si solo unos pocos atributos son diferentes de cero en cada fila / columna son diferentes de cero (“sparse coding”)

Aprendizaje no Supervisado

Biclustering

Manuel Sánchez-Montañés
Universidad Autónoma de Madrid
manuel.smontanes@gmail.com