Visualizing data structures using R :

Applied Biclustering Methods for Big and High Dimensional Data Using R

Ziv Shkedy

Hasselt University, Belgium

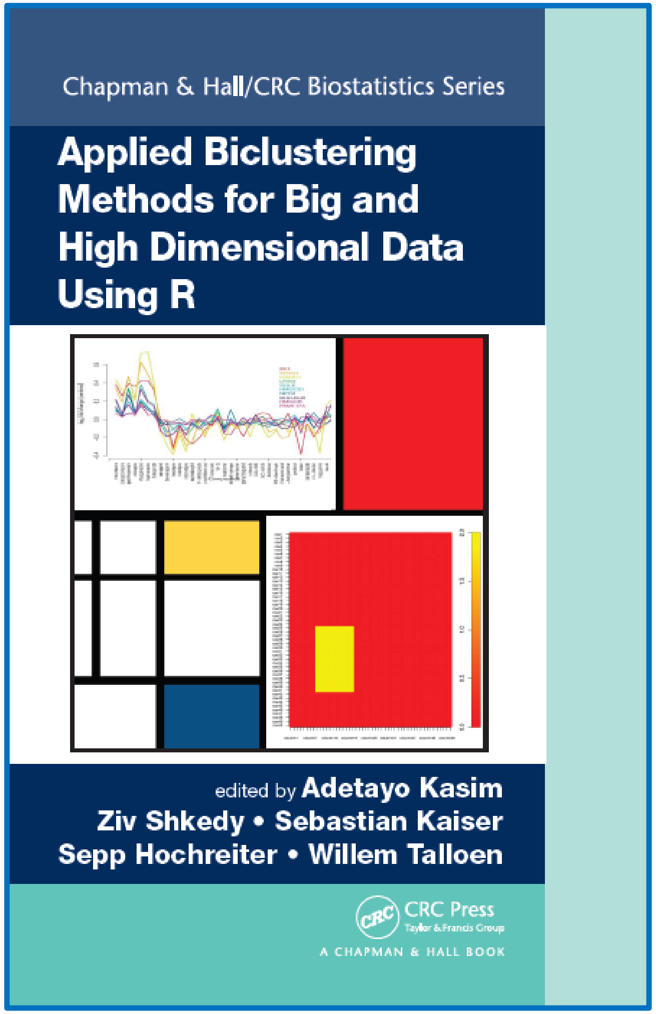
Hasselt University
February-April, 2022

follow us on twitter @ZShkedy

Visit us on Facebook Analysis of DNA Microarray and High-Dimensional Data

Email: ziv.shkedy@uhasselt.be

Research Team



Chapman & Hall/CRC Biostatistics Series

Applied Biclustering Methods for Big and High Dimensional Data Using R

edited by **Adetayo Kasim**
Ziv Shkedy • Sebastian Kaiser
Sepp Hochreiter • Willem Talloen

CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

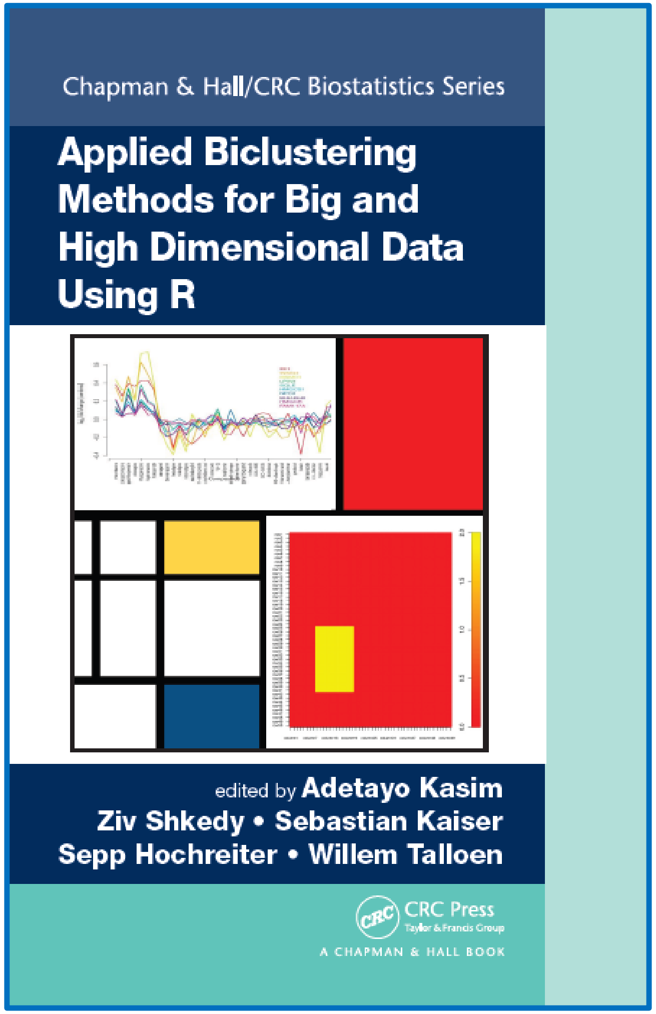
UHasselt:

Ewoud De Troyer
Rudradev Sengupta
Nolen Joy Perualila
Ziv Shkedy
Adetayo Kasim

.

and many others…..

Reference & R packages


Chapman & Hall/CRC Biostatistics Series

Applied Biclustering Methods for Big and High Dimensional Data Using R

edited by Adetayo Kasim
Ziv Shkedy • Sebastian Kaiser
Sepp Hochreiter • Willem Talloen

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

R packages:

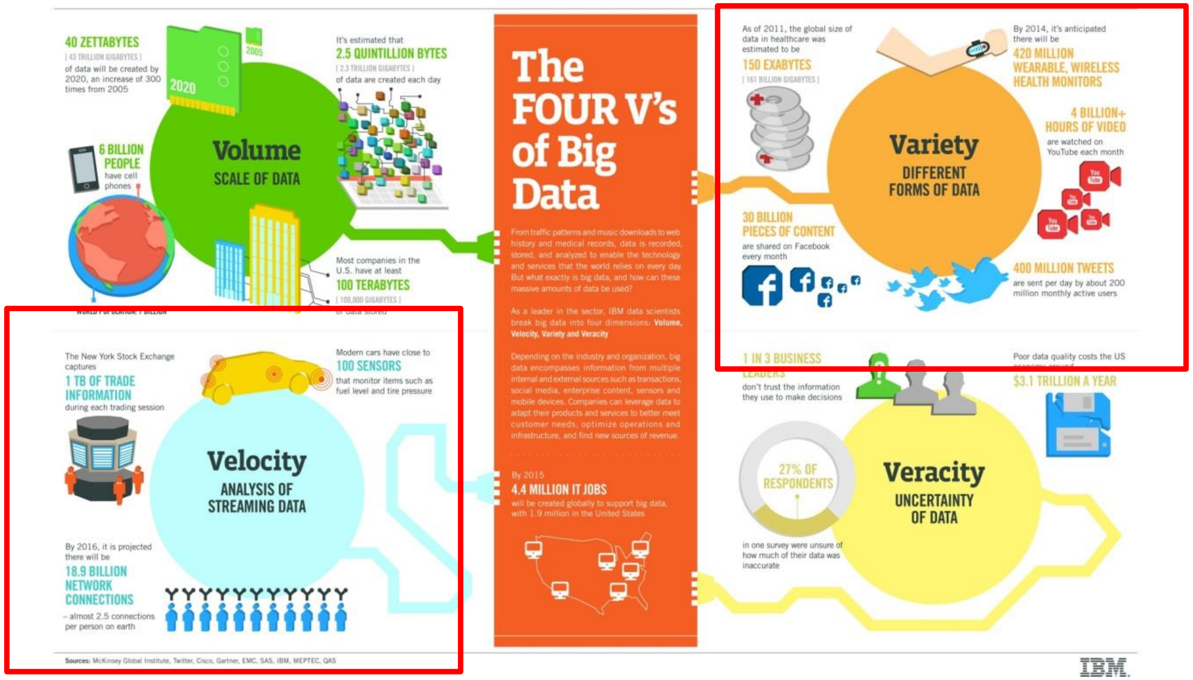
biclust (CRAN)
biclustGUI (CRAN)

PDF file with the first 9 chapters.

Part 1

Introduction

Big data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

Volume SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least 100 TERABYTES [100,000 GIGABYTES] of data stored

Velocity ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session

Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

Variety DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

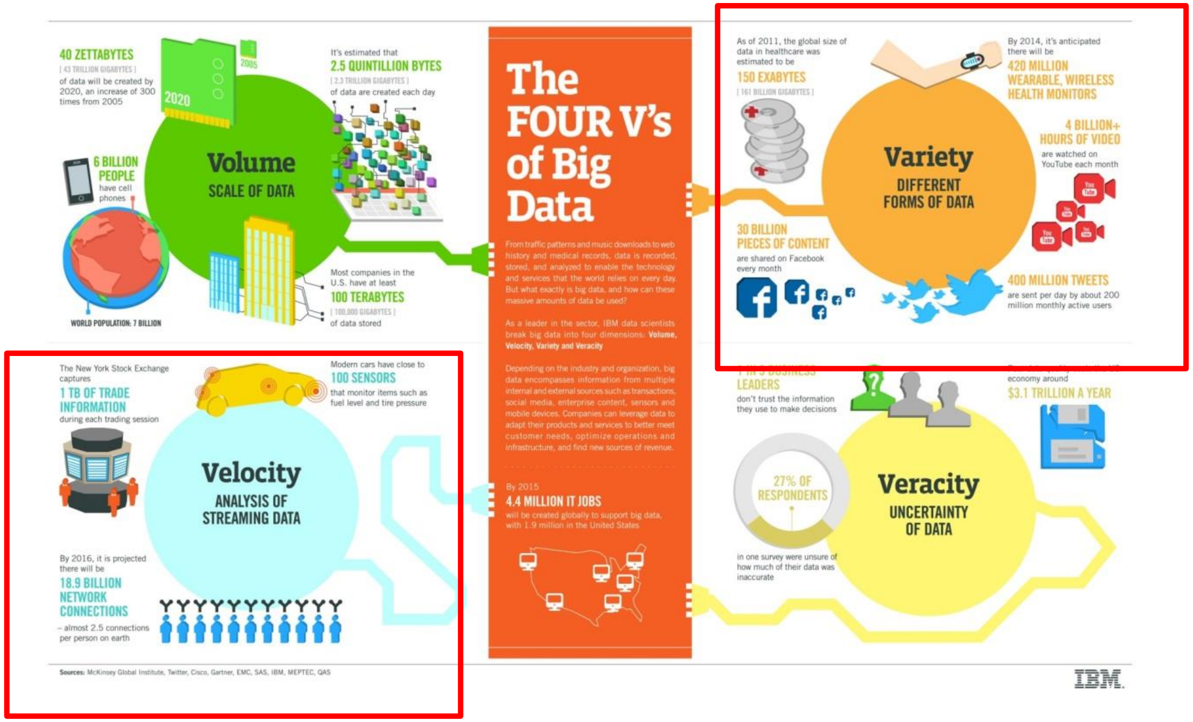
Poor data quality costs the US economy around $3.1 TRILLION A YEAR

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

- Everything is measurable…..
- We can collect a lot of data (and usually very quick)…..
- How can we identify patterns in the data ?

Big data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
LEADERS
don't trust the information they use to make decisions
economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
IBM

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

- Today:
- Data analysis tool to discover local patterns in big data matrices.
- Case studies:
  - Sport.
  - Tourism
  - Drug discovery.

Part 2

Biclustering: local versus global patterns

Data structure

$$\mathbf{X} = \underbrace{\left.\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} . \right\}}_{\text{Observations, samples, conditions ....}} \begin{matrix}\text{Variables,} \\ \text{features...}\end{matrix}$$

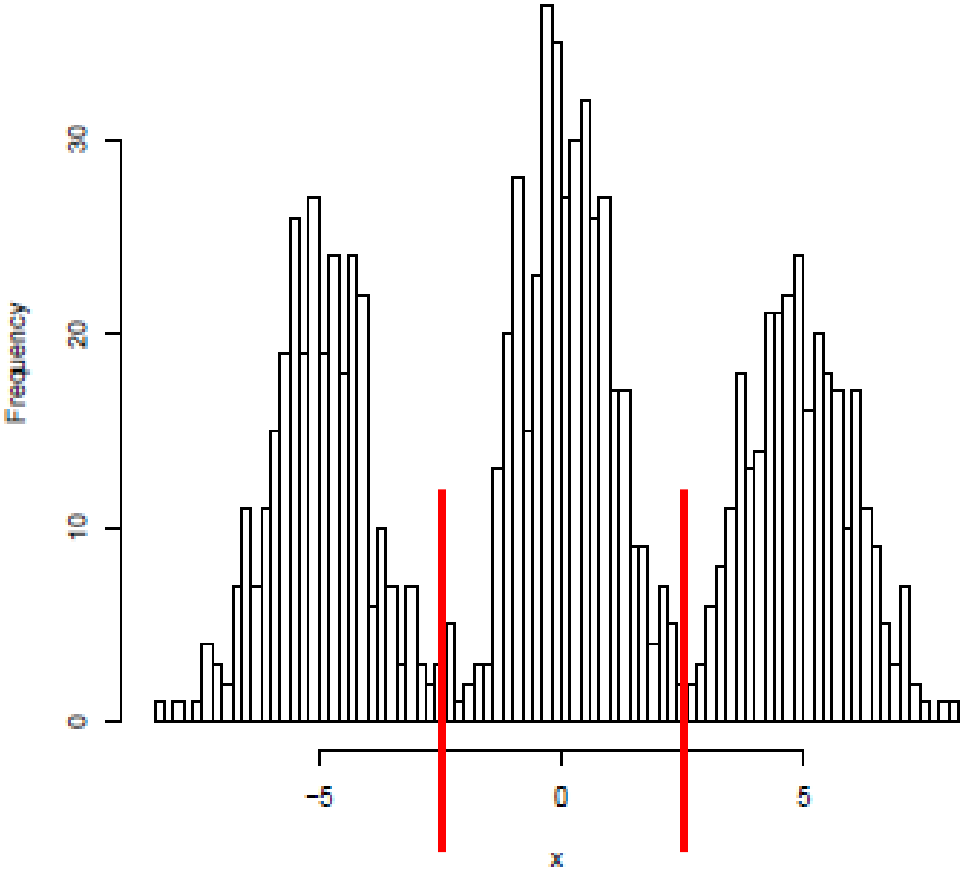
Global patterns

- Find variables (observations) that can be grouped together due to a pattern in the data matrix.

- Examples:
  - All costumers in a supermarkets that have a tendency to buy the same products.
  - Genes with the same expression profiles in an expression matrix.

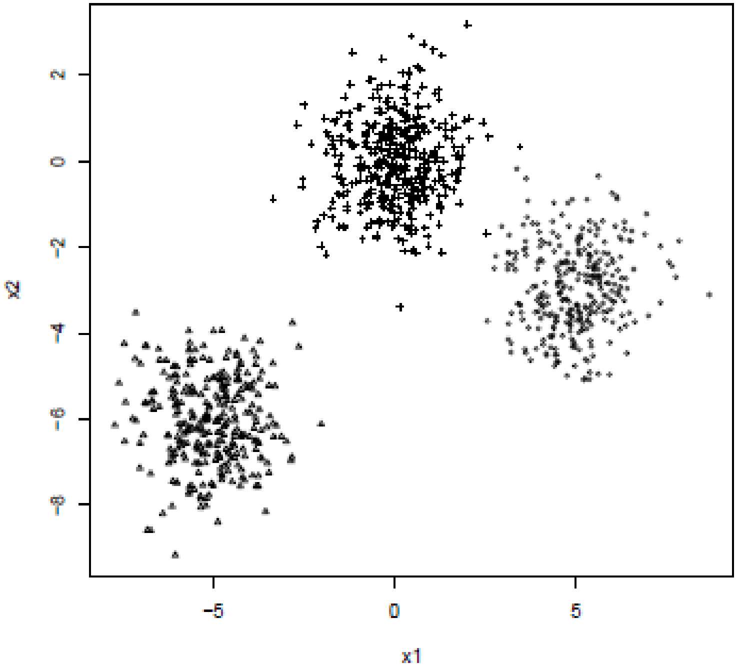
Clusters of observations

Example: three clusters of one variable

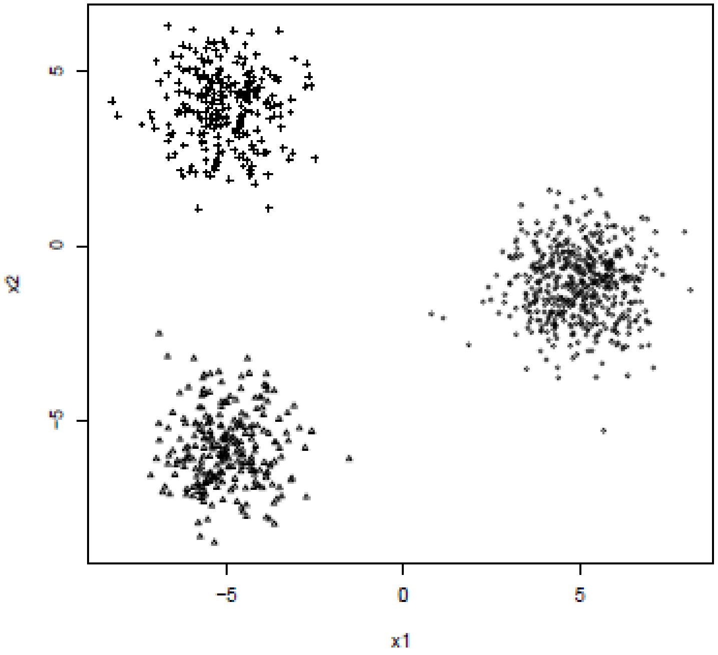


Clusters of observations

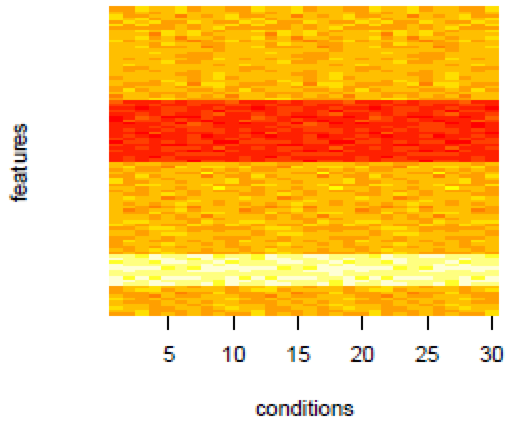
Example: three cluster in two variables

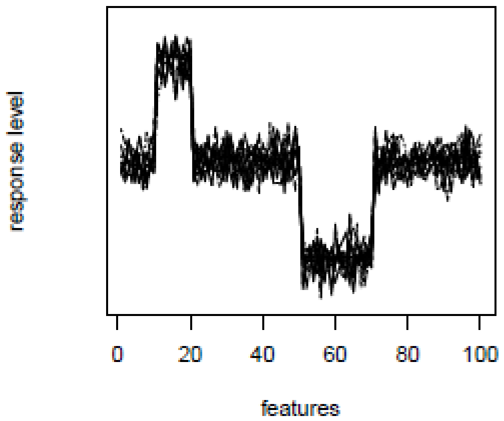


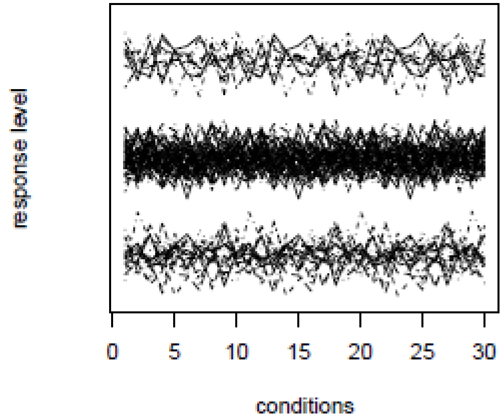
Example: how many clusters ?



Clustering and similarity measures







A data matrix with three clusters (of variables, rows).

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}.$$

Correlation between rows across all columns

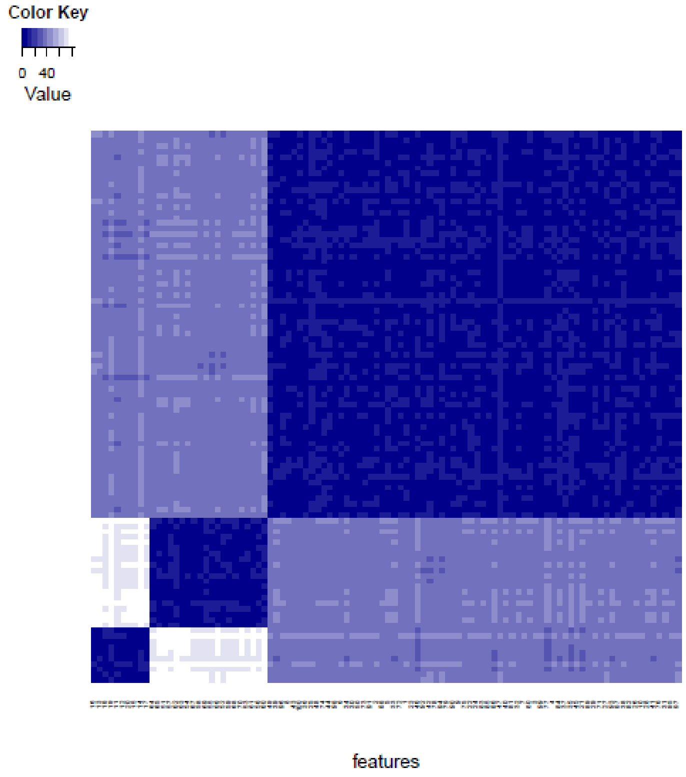
Clustering and similarity measures

The observe data matrix.
Clustering features.

$$\left.\begin{bmatrix} x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{m,1} & x_{m,2} & . & . & x_{m,n} \end{bmatrix}\right\} \text{features}$$

How similar are the features across the samples (conditions).
Eample:correlation across samples of two features:

$$\rho(x_i, x_j)$$



Example of three clusters.

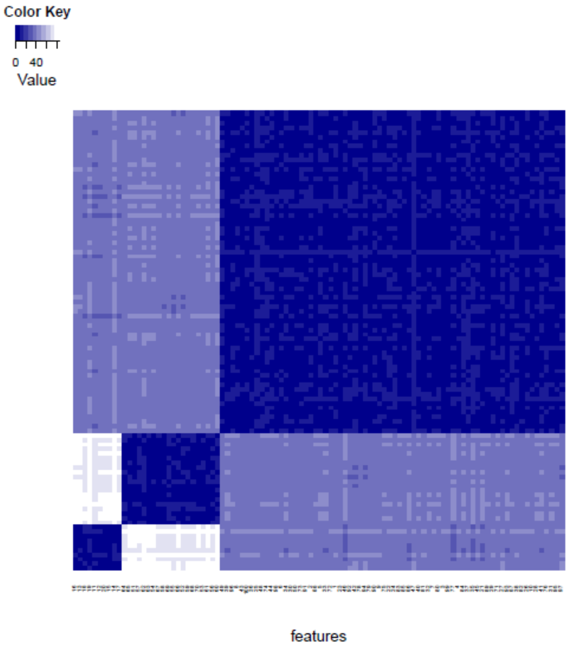
Hierarchical clustering

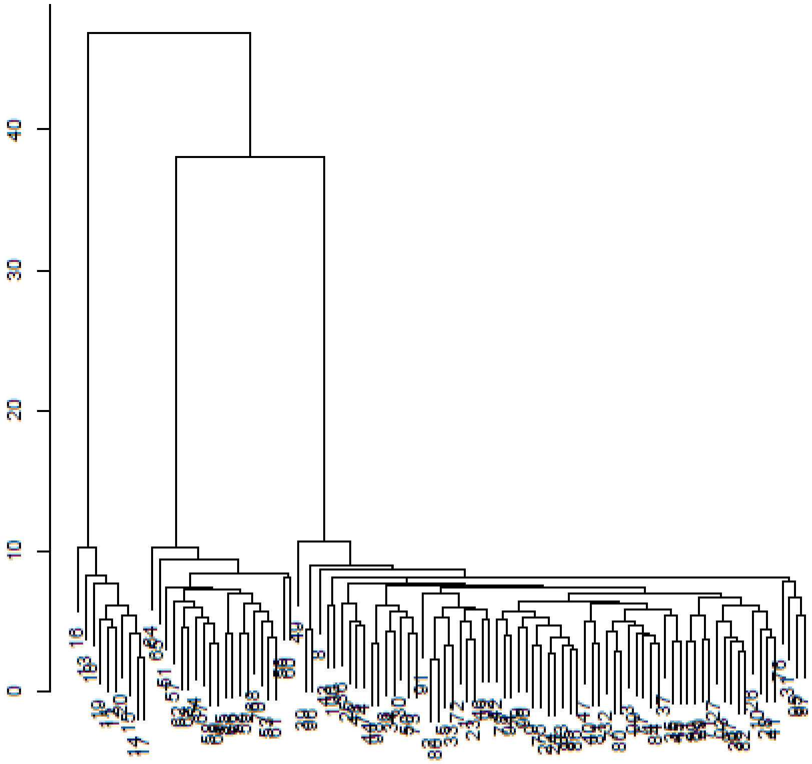
- Group variables according to their correlation (with each other).
- Variables are correlated across all observations.

⇩

Global pattern

Hierarchical clustering





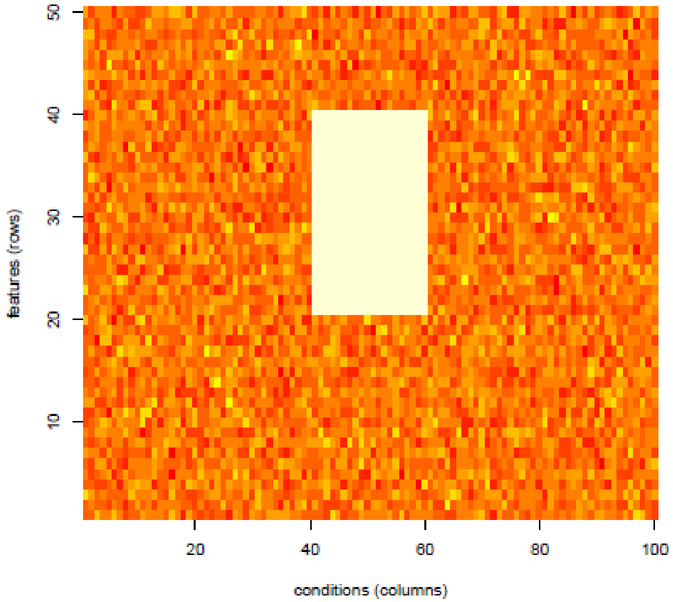
(b) Example 2.

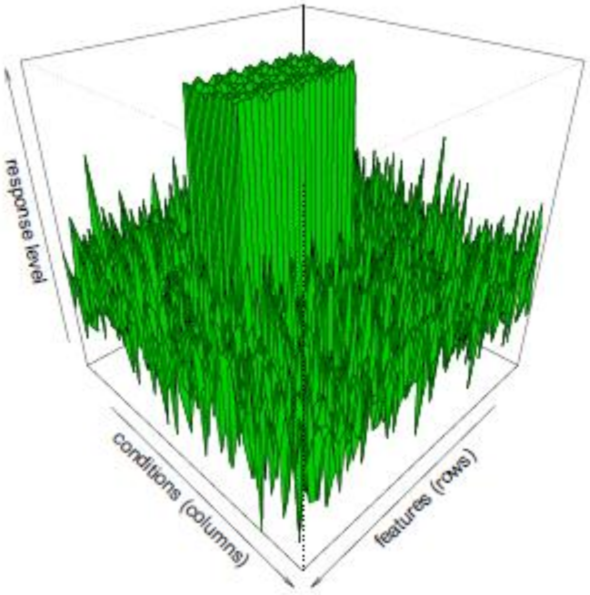
Example of three clusters.

Local patterns

- We are looking for:
  - A subset of features with the same characteristic across a subset of conditions.
  - Example:
    - A group of genes with the same expression patterns across a subset of samples.
    - A group of costumers that buy the same products in a supermarket.
    - A group of students with the same results patters across a group of subjects.
    - …….

Local patterns in a data matrix



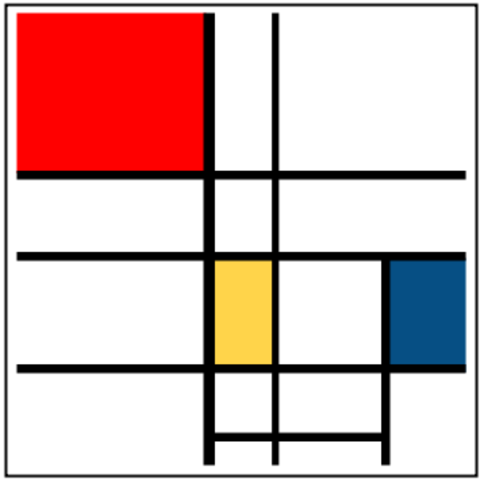


Example of a subset of features with high response level on a subset of conditions.

Local vs. global

A bicluster

$$\mathbf{X} = \begin{pmatrix} \multicolumn{2}{c}{\boxed{\begin{matrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{matrix}}} & \begin{matrix} \ldots \\ \ldots \end{matrix} & \begin{matrix} X_{1m} \\ X_{2m} \end{matrix} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nm} \end{pmatrix}.$$



A subset of features in a data matrix that have a similar response patterns across a subset of samples.

Example: a group of genes with a similar expression profiles across a group of samples

A bicluster: signal and noise

Within a bicluster: additive or multiplicative structures

$$Y = signal + noise$$

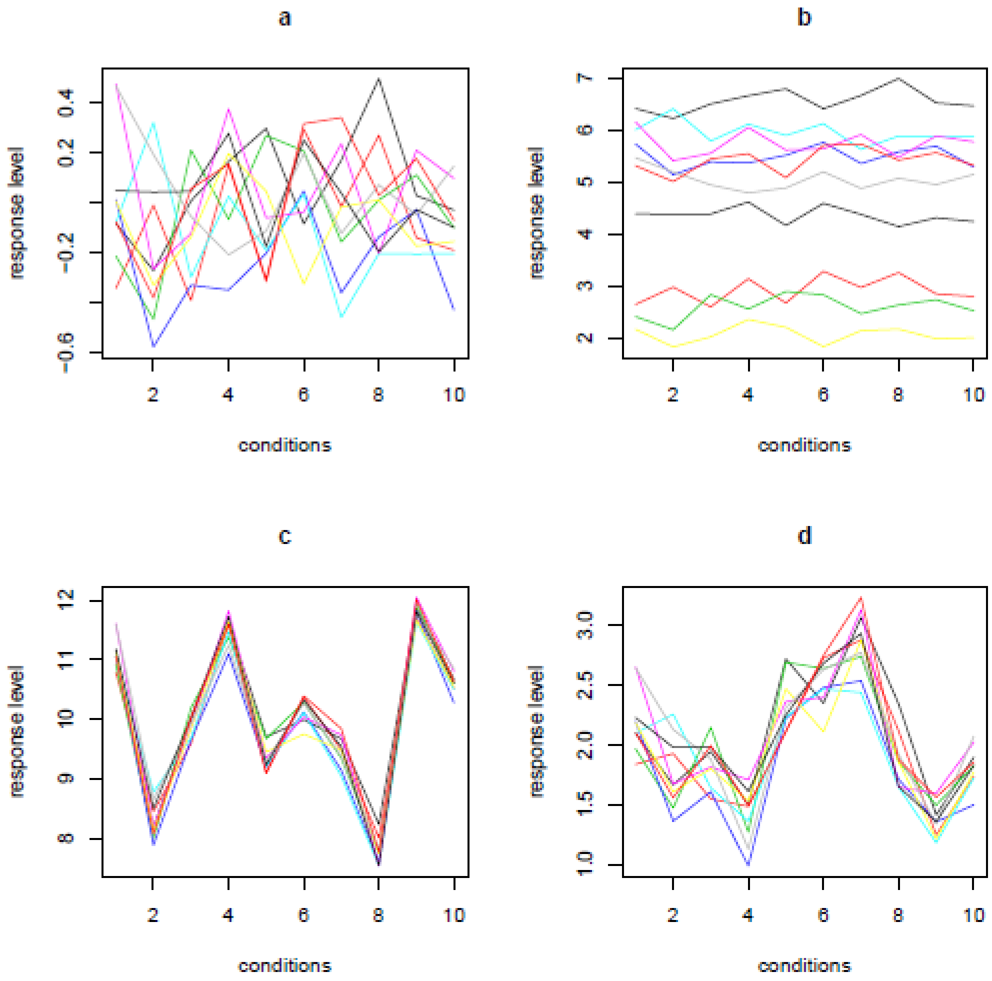
Signal structure: multiplicative or additive.

Outside a bicluster:

$$Y = noise$$

$$\begin{bmatrix} \enclose{circle}{x_{1,1}} & x_{1,2} & . & . & \enclose{circle}{x_{1,n}} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \\ \enclose{circle}{.} & . & . & . & \enclose{circle}{.} \\ . & . & . & . & . \\ \enclose{circle}{x_{m,1}} & x_{m,2} & . & . & \enclose{circle}{x_{m,n}} \end{bmatrix}$$

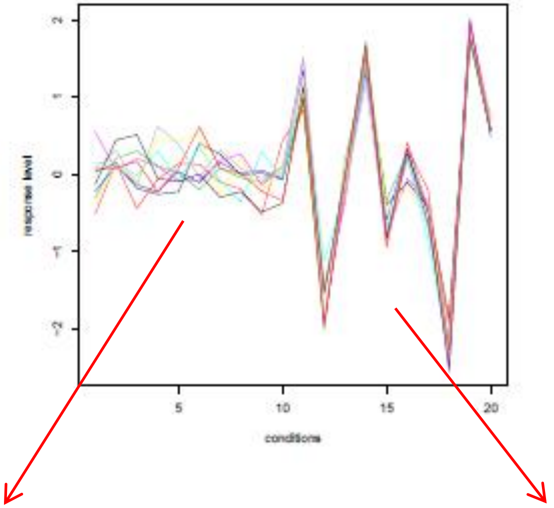
A bicluster: rows and columns effects

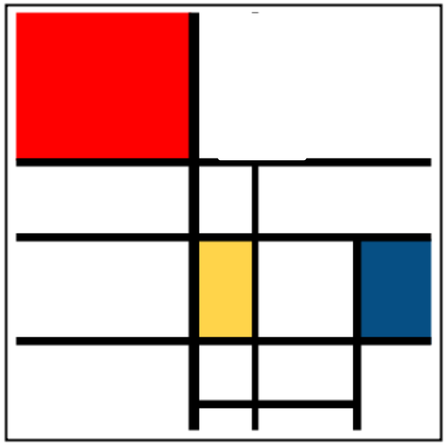


Dominant effects:

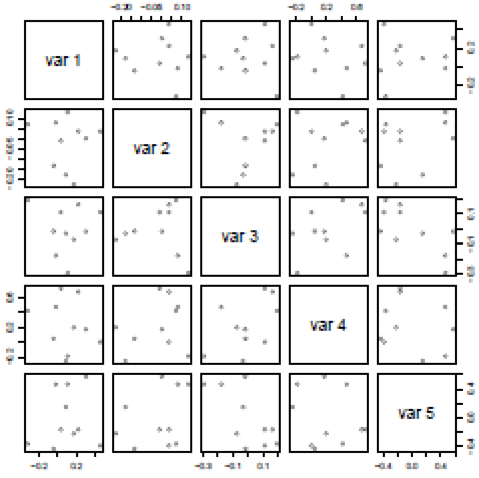
- Rows ?
- Columns ?
- Rows and columns ?

A bicluster: correlation

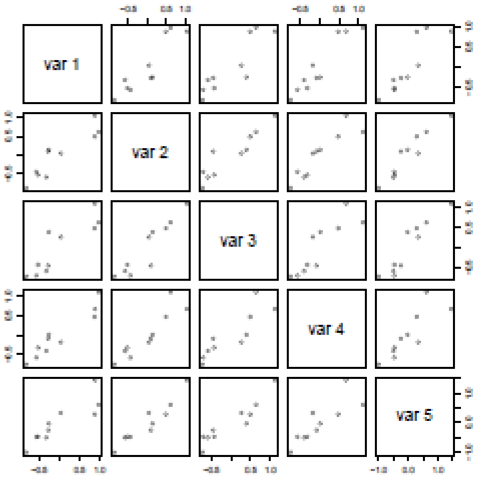




Conditions outside the BC



Conditions within the BC



Structure within and outside the biclsuter:

Not always different levels (i.e. significant different signal)

Signal structure

$$Y = R + Erorr$$

$$Y = C + Erorr$$

$$Y = R + C + Erorr$$

Additive BCs

R: rows.

C: columns.

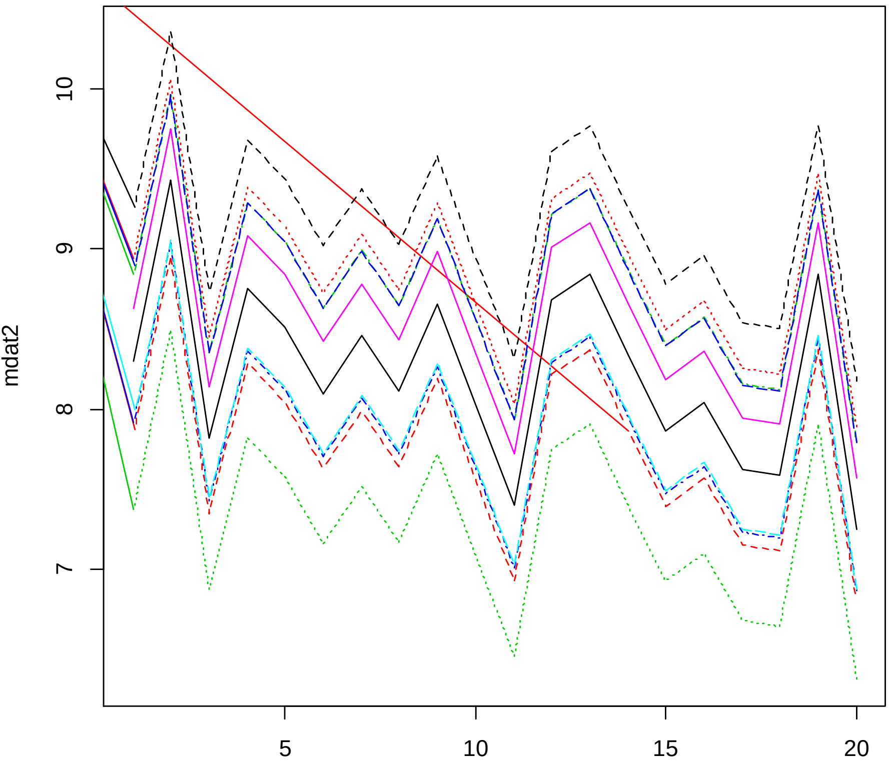
$$\left.\begin{aligned} Y &= R \times C \times Erorr \\ Y &= R \times C^{Error} \end{aligned}\right\} \Rightarrow \begin{aligned} \log(Y) &= \log(R) + \log(C) + Error \\ \log(Y) &= Error \times (\log(R) + \log(C)) \end{aligned}$$

$$Y = R \times C + Erorr$$

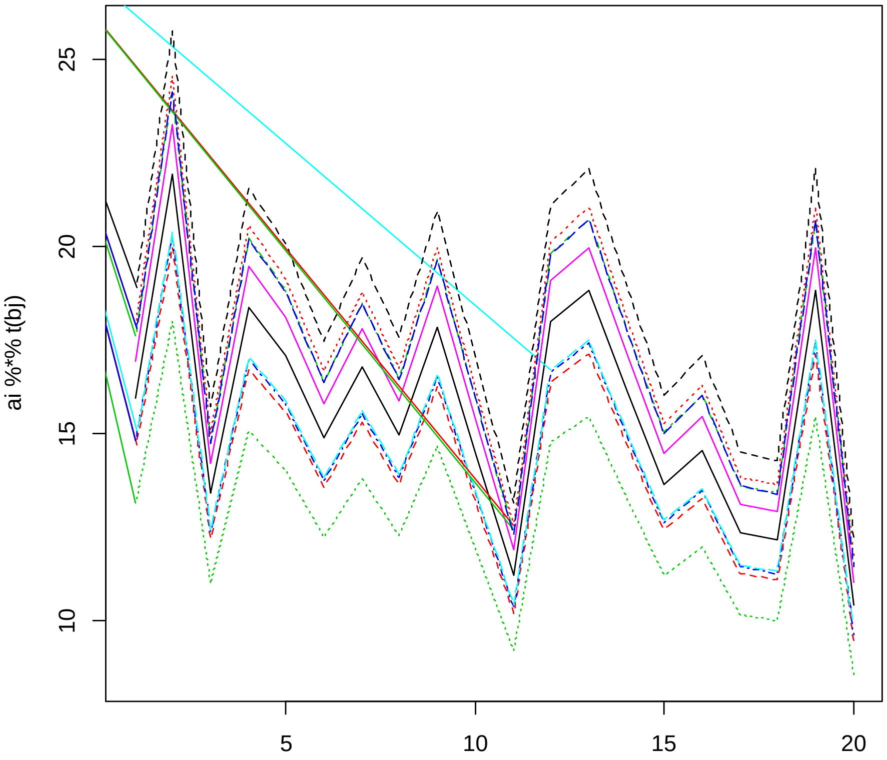
Multiplicative BCs

Signal structure

Additive BC: $Y = R + C$



Multiplicative BC: $Y = R \times C$



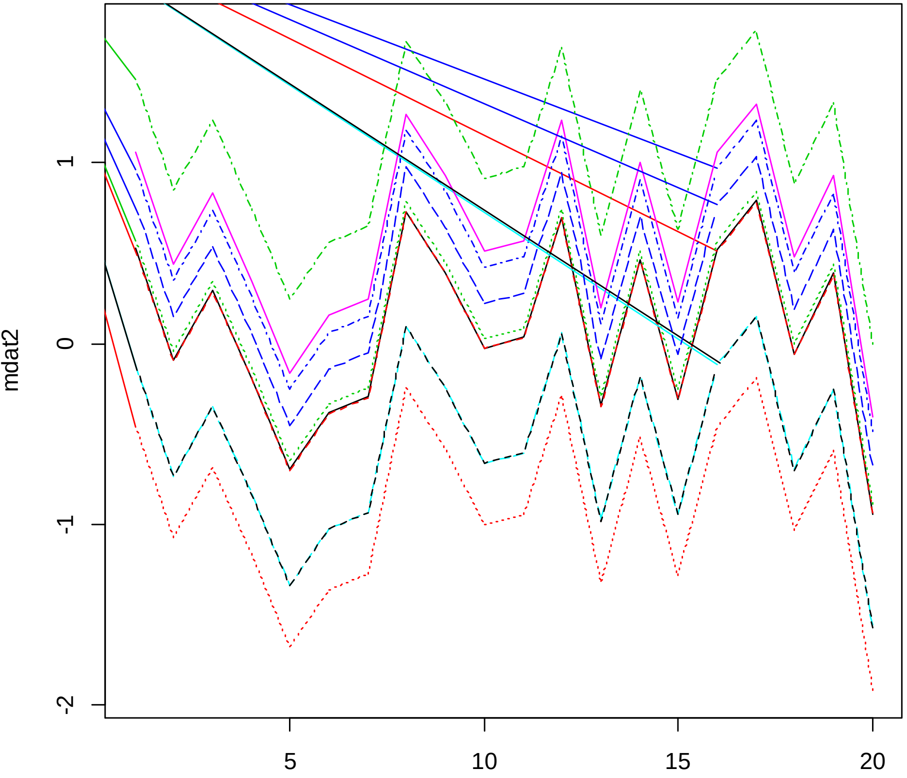
$$R_i \sim N(3,0.25)$$

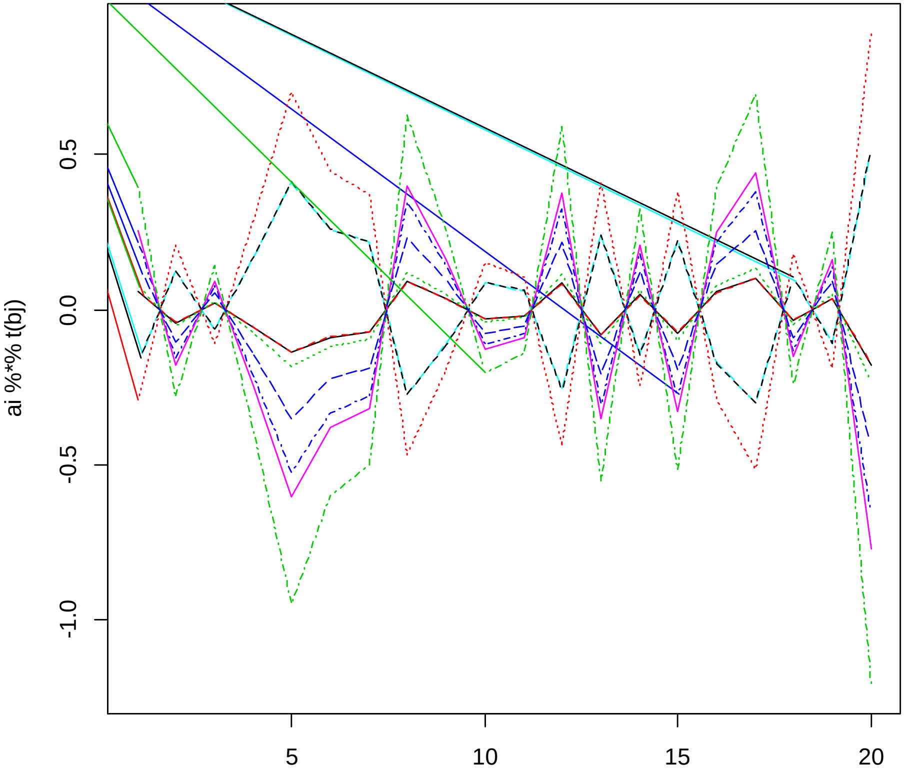
$$C_j \sim N(2,0.25)$$

Signal structure

Additive BC: $Y = R + C$

Multiplicative BC: $Y = R \times C$





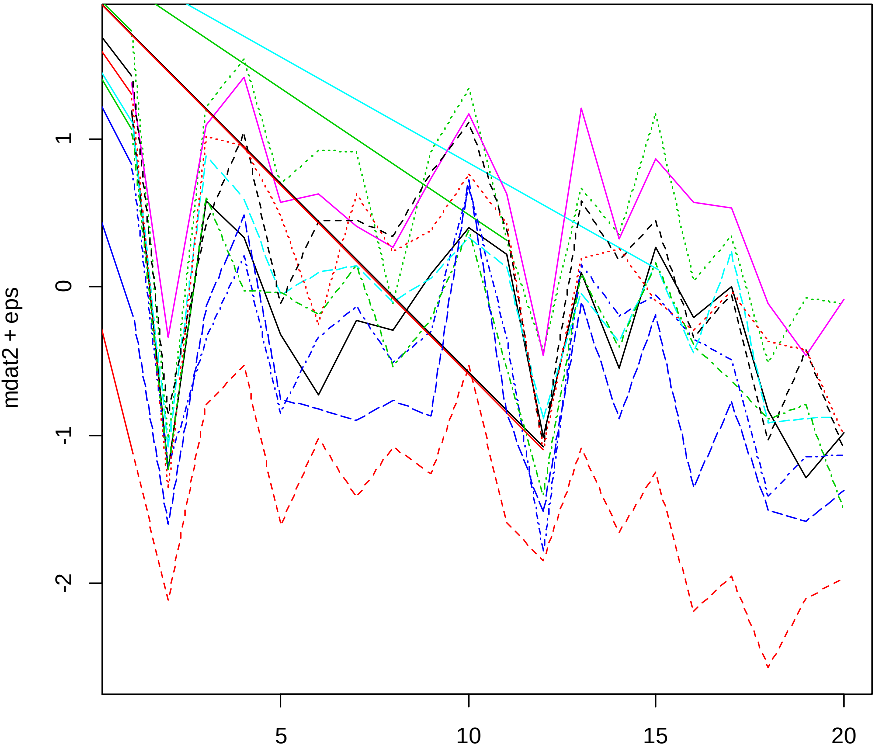
$$R_i \sim N(0,0.25)$$

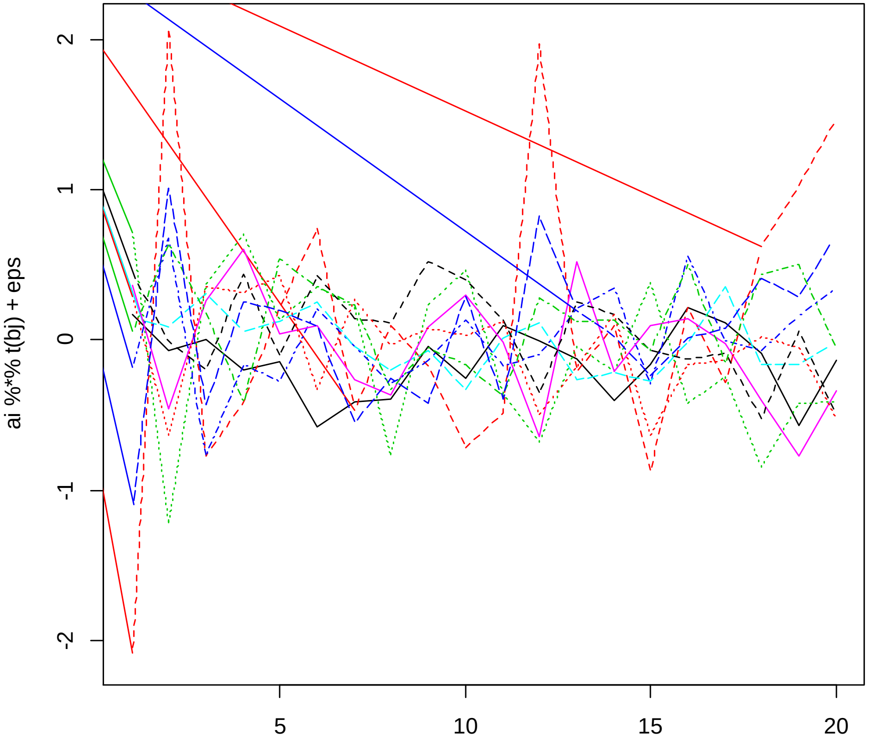
$$C_j \sim N(0,0.25)$$

Signal + noise

Additive BC: $Y = R + C + Erorr$

Multiplicative BC: $Y = R \times C + Erorr$



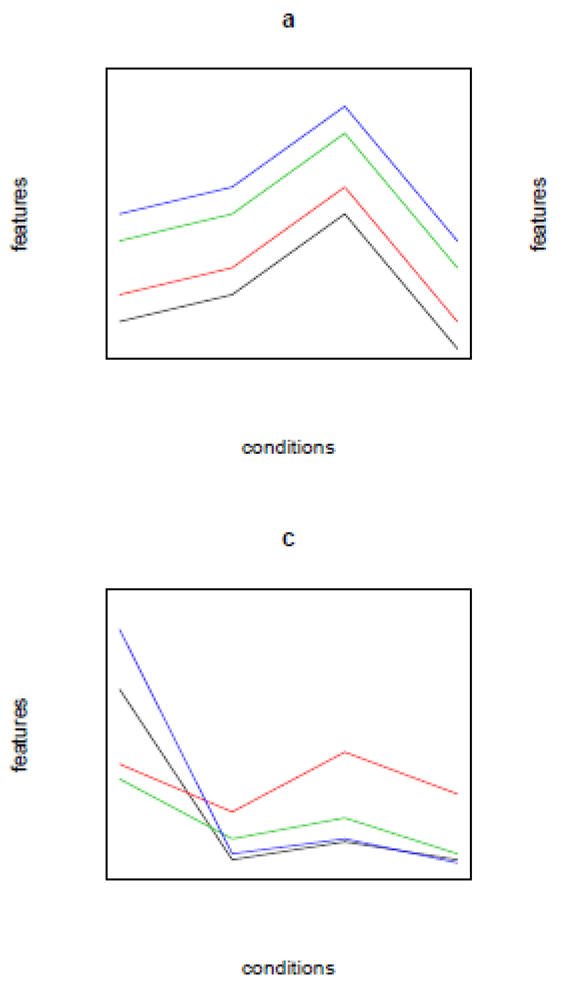


$R_i \sim N(0,0.25)$

$C_j \sim N(0,0.25)$

$E_{ij} \sim N(0,0.0625)$

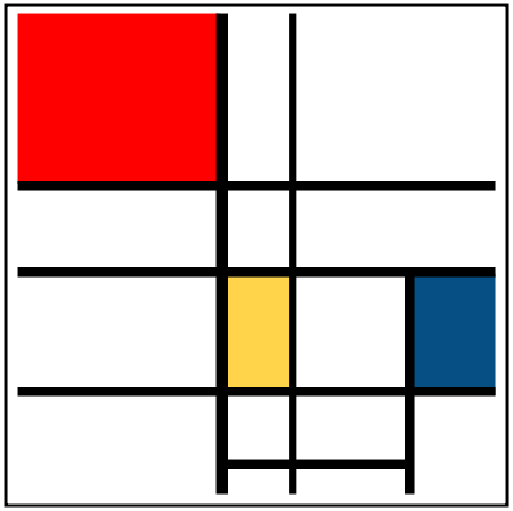
Types of biclusters



- Constant BC.
- Rows effects.
- Columns.
- Rows and columns effects (a).
- Coherent values (b).
- Coherent evolution (c).

Configurations of biclusters in the data matrix

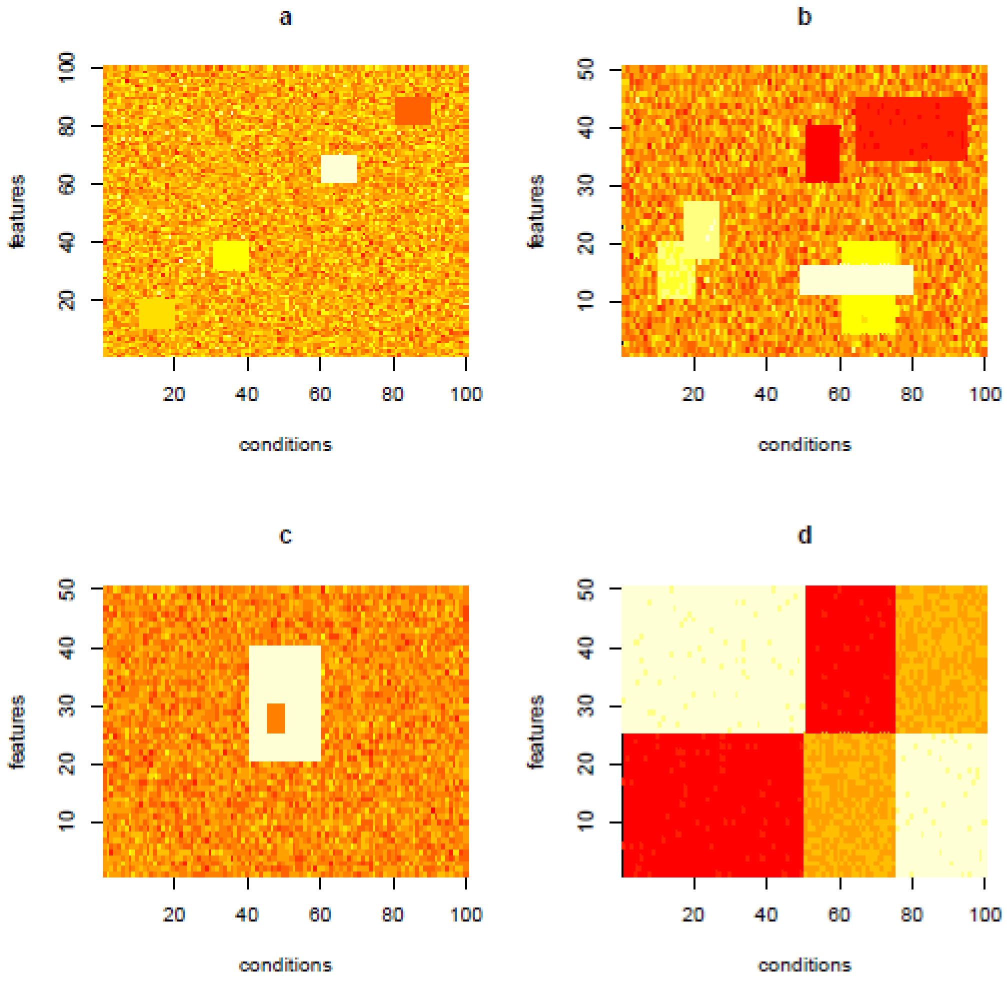
Which structure we observed in the data matrix ?



Piet Mondrian

Theo van Doesburg

Configurations of biclusters in the data matrix



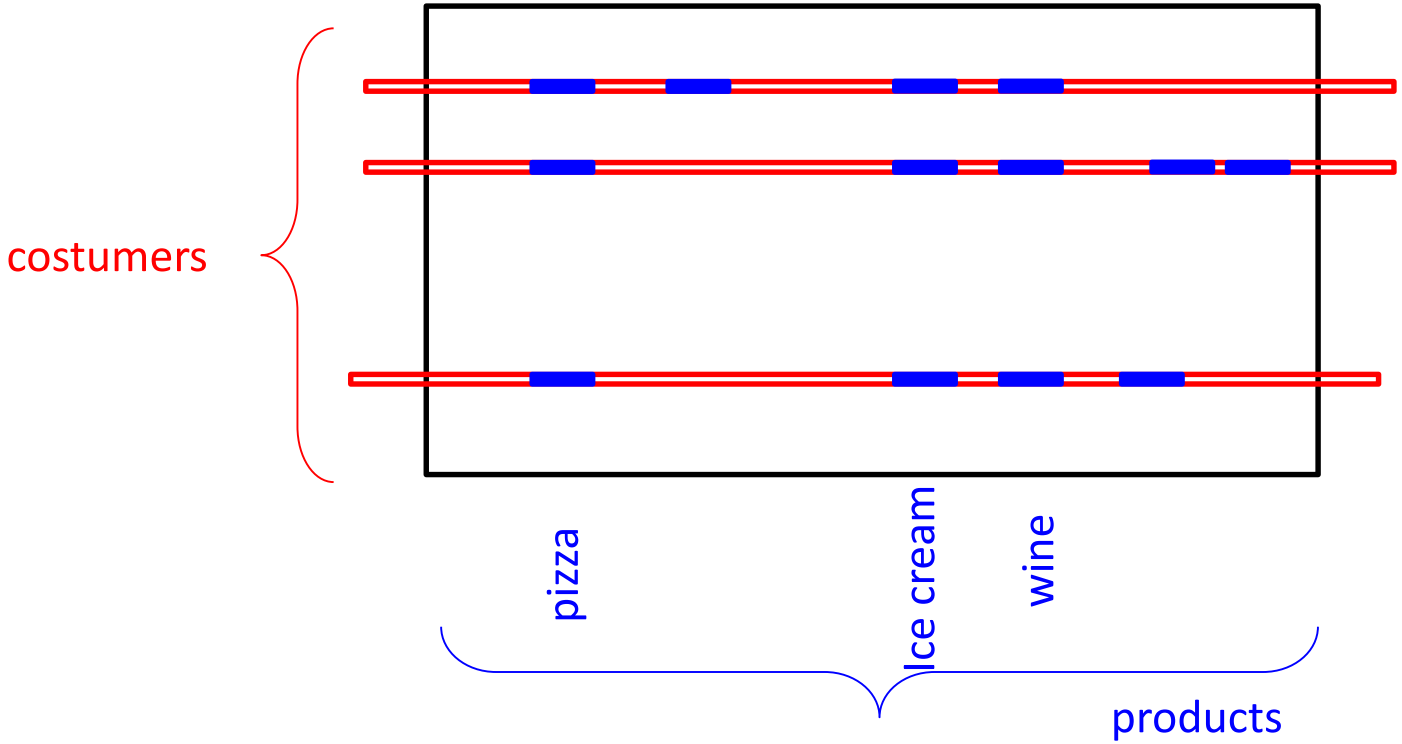
Overlapping
Non overlapping
.
.

Local patterns

- Why local ?
- In a supermarket, if we know that a group of costumers have a tendency to buy : pizza, wine and ice cream we can help them to buy these products.
- In a holiday resort, if we know that costumers like to go to the sea and to have BBQ there....

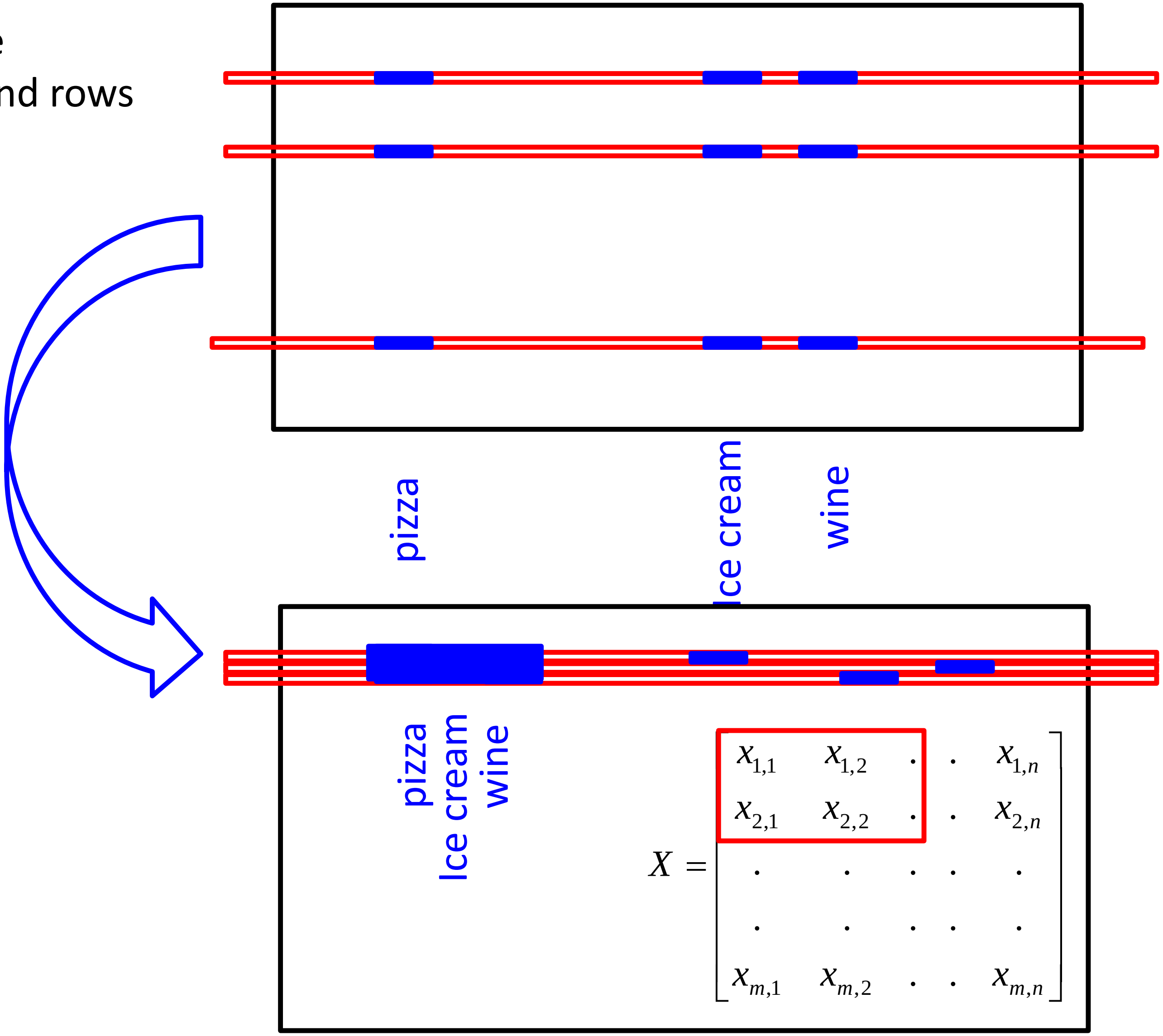
Local patterns

- In a supermarket, if we know that a group of costumers have a tendency to buy : pizza, wine and ice cream we can help them to buy these products.
- Why this is a biclsuter ?



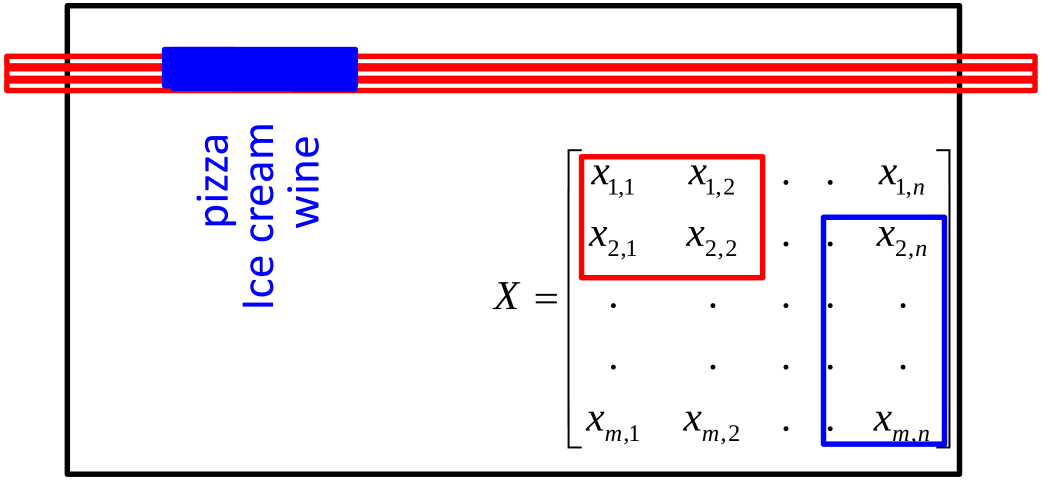
Local patterns

Re arrange
columns and rows



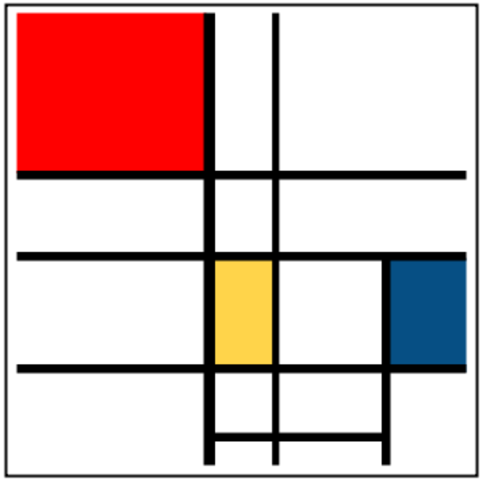
$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{m,1} & x_{m,2} & . & . & x_{m,n} \end{bmatrix}$$

Many local patterns….



$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{m,1} & x_{m,2} & . & . & x_{m,n} \end{bmatrix}$$

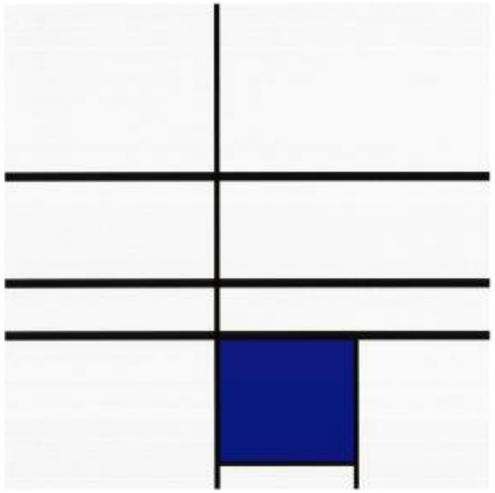
Pizza, ice cream & wine →



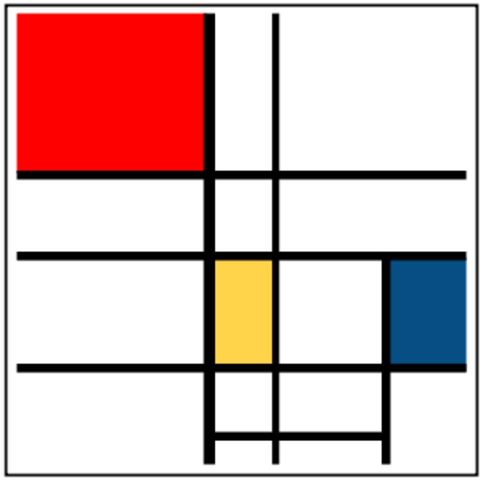
← Organic vegetables, free ranged chicken and cheese…

Exempels of BC

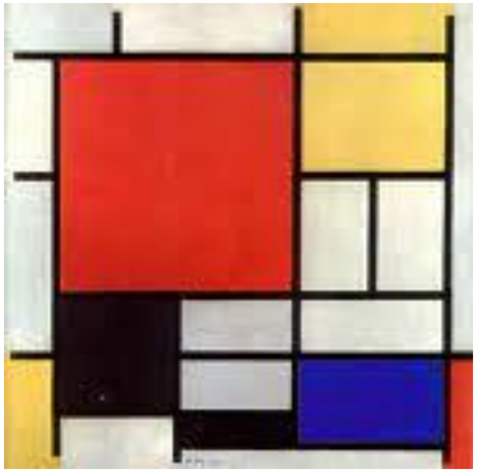
One bicluster



Three biclusters



Overlaping biclusters



Piet Mondrian

Overlapping only in one dimension (rows or columns).

Examples of overlapping biclusters

Many Overlaping biclusters

Theo van Doesburg

Overlapping only in one dimension (rows or columns).

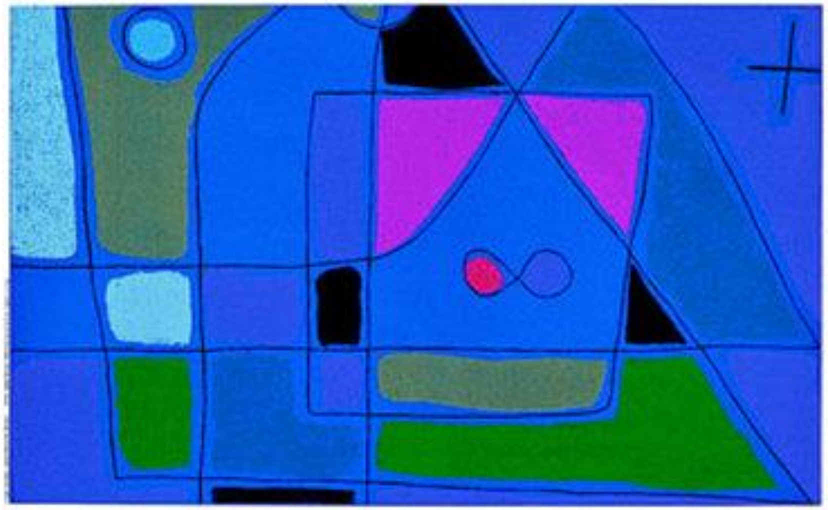
Overlaping biclusters

Theo van Doesburg

Other examples

Jean Arp

Paul Klee



Sonia Delaunay

Not everybody understood the concept of biclustering so good…

Part 3

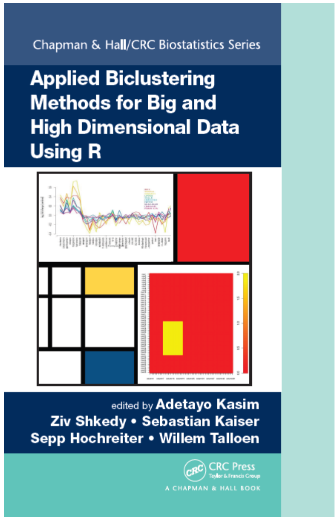
Selection of Biclustering methods

- Computer science methods:
  - Bimax.

- Statistical methods:
  - The plaid model.
  - FABIA.

Part 3.1

Bimax



Chapter 5

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

Prelic el al. 2006

- Bimax (binary inclusion-maximal biclustering algorithm).

- The Bimax is a biclustering algorithm introduced by Prelic 2006 as a reference biclustering method for a comparison with different biclustering methods.

Data structure: binary data

Data matrix

$$\begin{array}{cc} \text{features} & \begin{bmatrix} Z_{1,1} & 0 & . & . & 1 \\ Z_{2,1} & 1 & . & . & 1 \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_{m,1} & 1 & . & . & 0 \end{bmatrix} \\ & \text{conditions} \end{array}$$

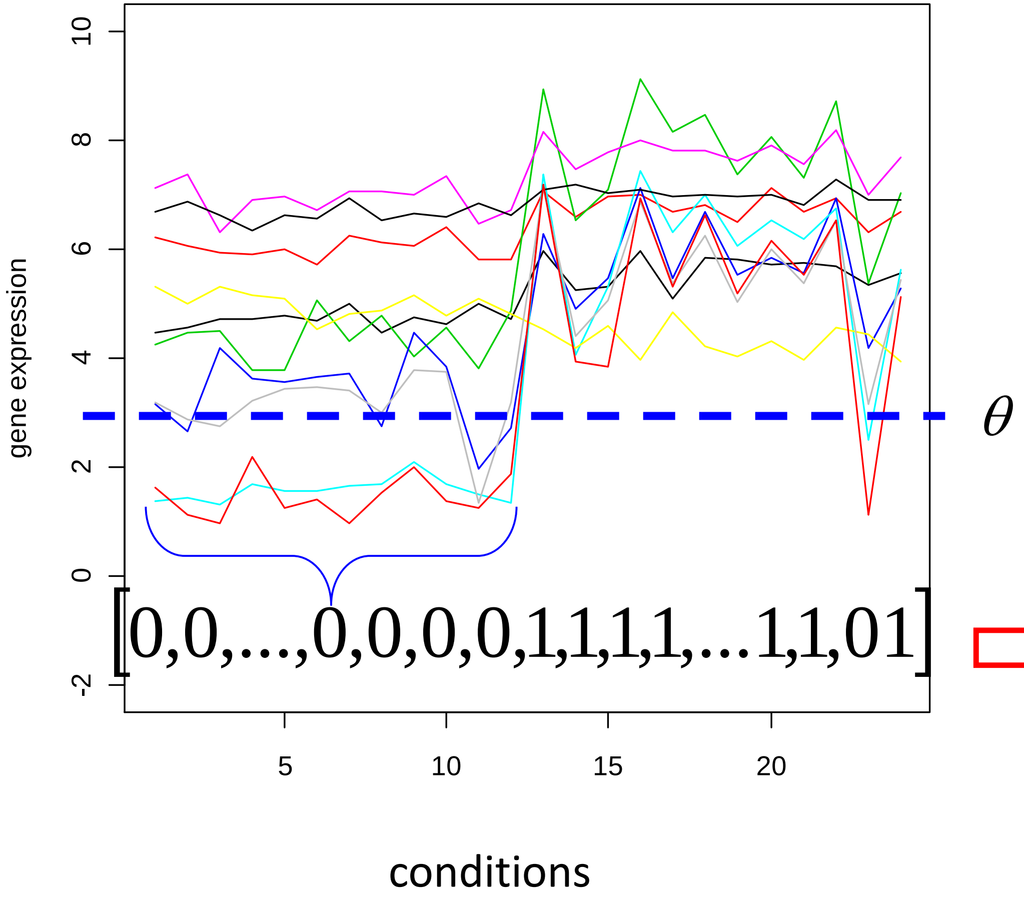
$$Z_{ij} = \begin{cases} 1 & \textit{feature i is active on condition j} \\ 0 & \textit{otherwise} \end{cases}$$

Examples:

- In the supermarket: subject i buy product j
- In football: player i scores a goal in the last 10 minutes of the game.

Data structure: binary data

The original data is continuous.



Examples

- Gene i is expressed under condition j

$$Z_{ij} = \begin{cases} 1 & X_{ij} > \theta \\ 0 & X_{ij} \le \theta \end{cases}$$

Dichotomize the red subject

Data structure: binary data

$$\begin{bmatrix} x_{1,1} & x_{1,2} & . & . & x_{1,n} \\ x_{2,1} & x_{2,2} & . & . & x_{2,n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{m,1} & x_{m,2} & . & . & x_{m,n} \end{bmatrix}$$

We are looking for subset of active features.

$$z_{ij} = \begin{cases} 1 & \text{feature i expressed (active) in condition j} \\ 0 & \text{otherwise,} \end{cases}$$

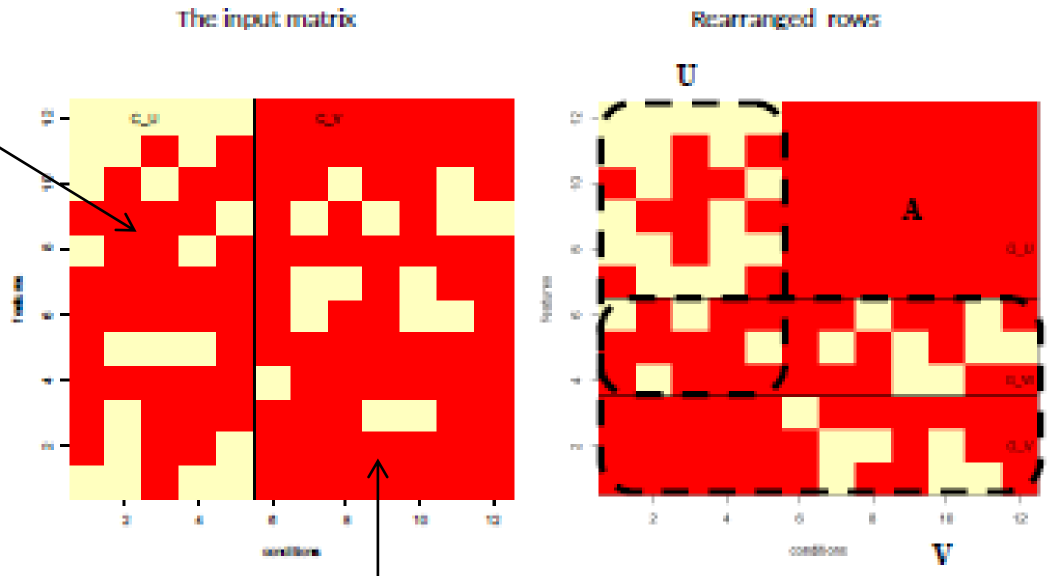
$$\begin{bmatrix} Z_{1,1} & 0 & . & . & 1 \\ Z_{2,1} & 1 & . & . & 1 \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_{m,1} & 1 & . & . & 0 \end{bmatrix}$$

Can we find a sequence of 1s of features across the same conditions ?

The Bimax algorithm

Divide the columns in two sets, C_U and C_V, based on the first row (in the first row, C_U contains only ones, while C_V contains only zeroes.

C_U



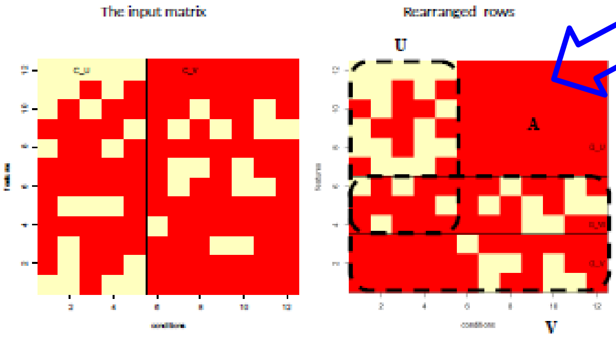
C_V

Re arrange the rows

The Bimax algorithm

Step 1:

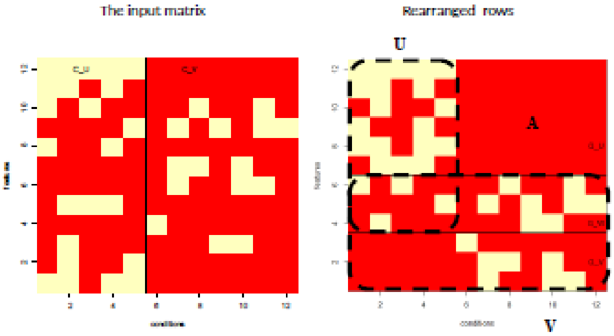
- Re arrange the data matrix.
- Exclude all rows/columns combinations with only zeros.



Step 2:

- Search for rows/columns combinations with ones.

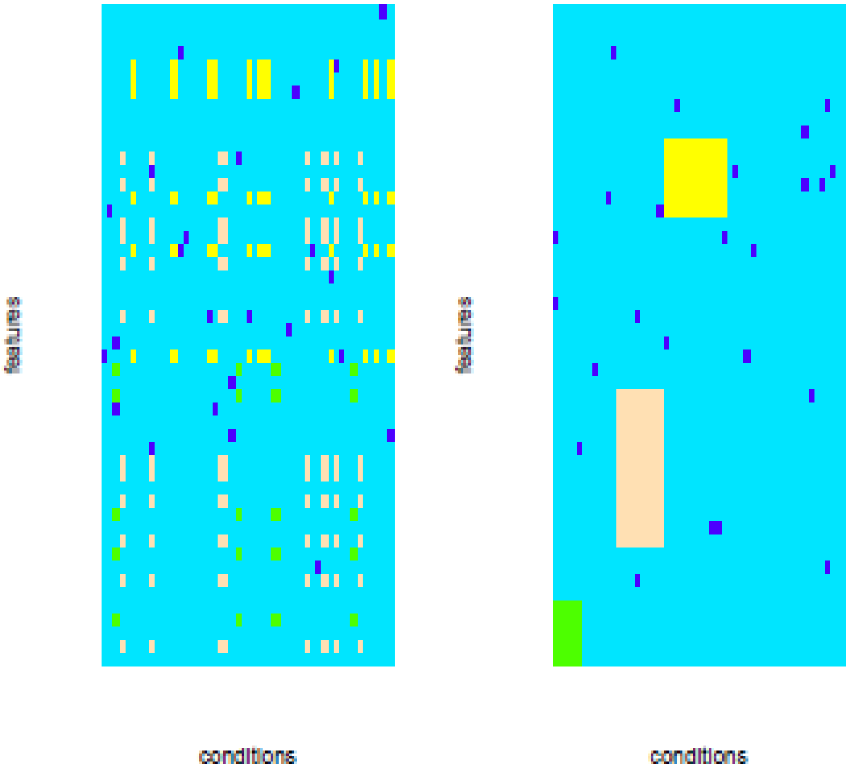
The Bimax algorithm: parameter setting



How many biclsuters we are looking for ?

What is the minimum size of the biclsuter (i.e. number of rows and columns) ?

The Bimax algorithm: an illustration

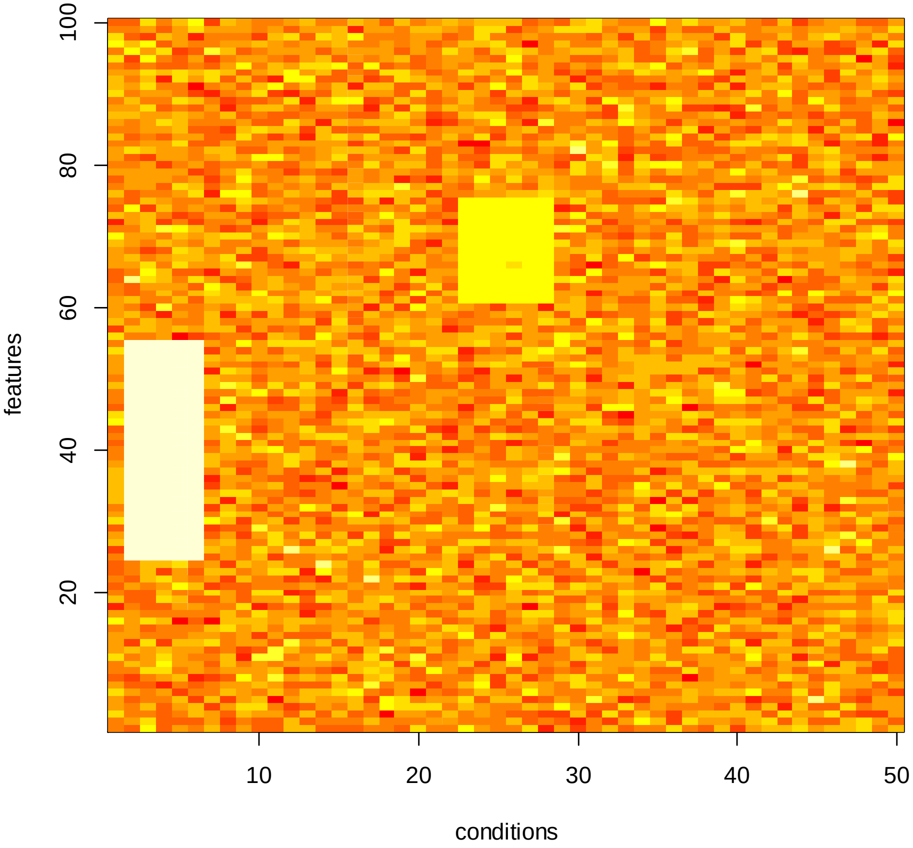


A data matrix
with 3 BCs

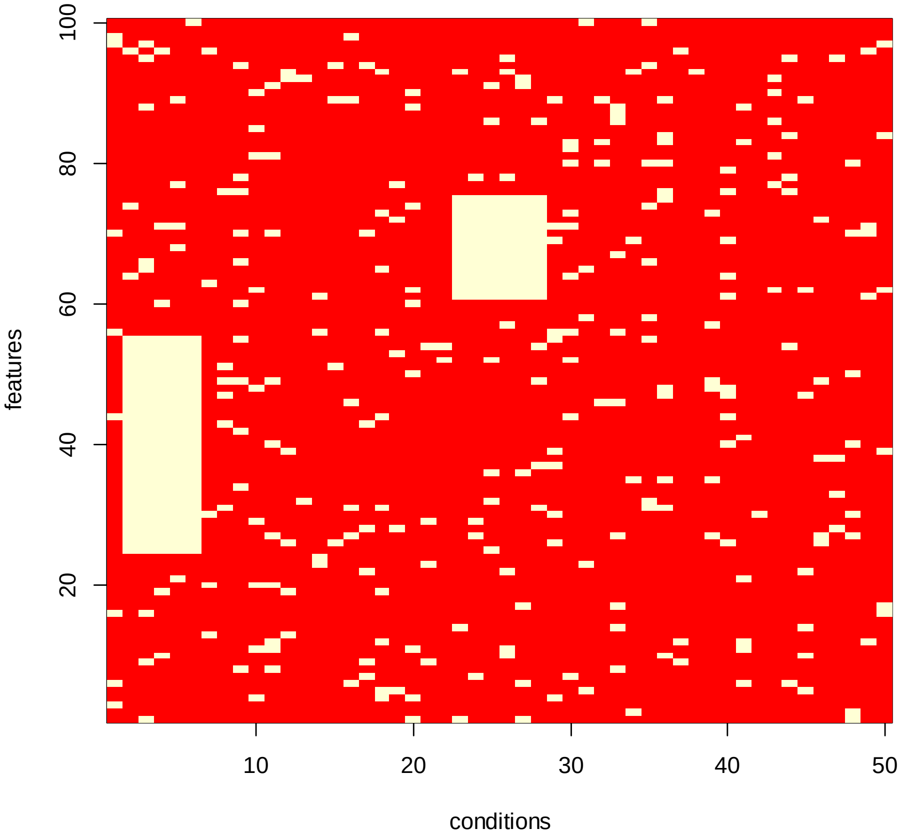
Example

A 100 X 50 matrix with two BCs.

Before dichotomization (X).

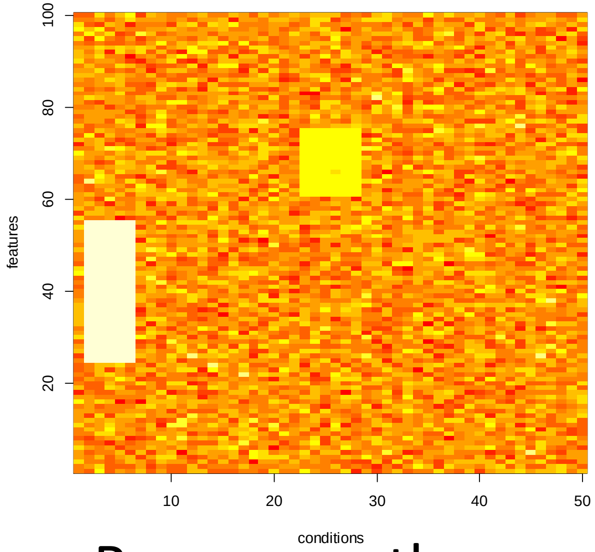


After dichotomization (Z).

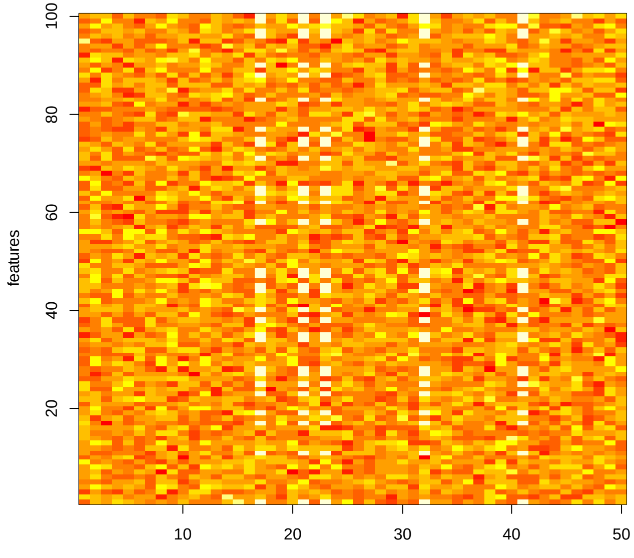


The same example: the input data

This is an unobserved matrix

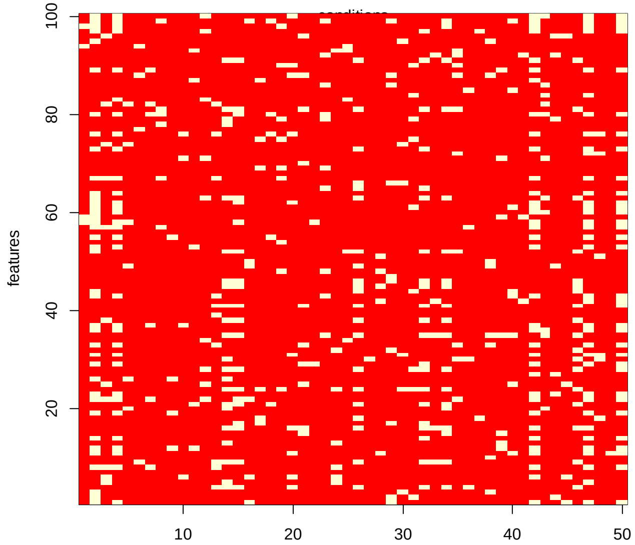


Observed matrix



Re arrange the rows and columns

The input data

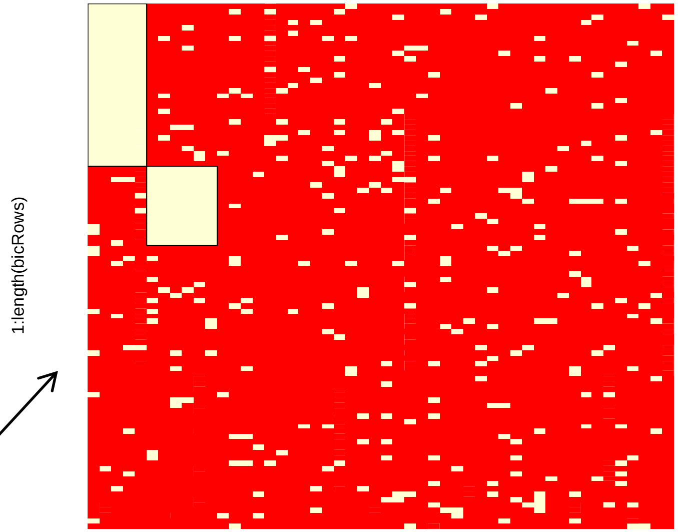


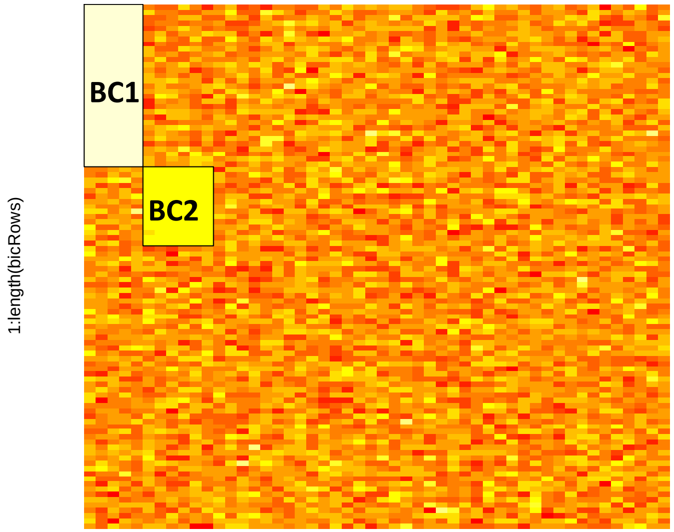
$$Z_{ij} = \begin{cases} 1 & X_{ij} > \theta \\ 0 & X_{ij} > \theta \end{cases}$$

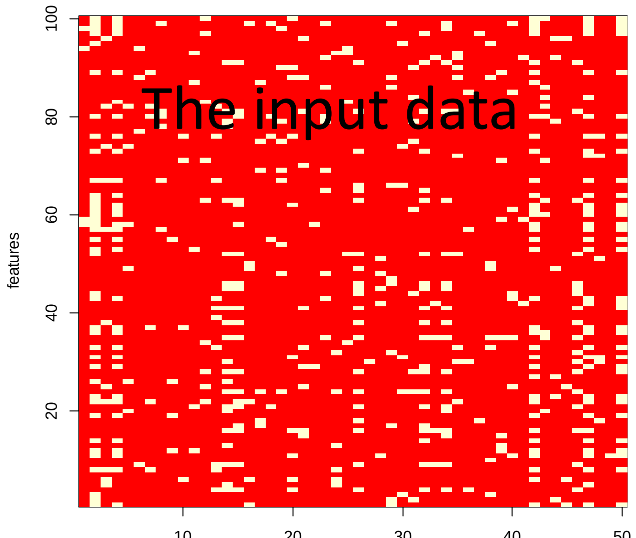
Results

```
> test.b1<-binarize(test2, threshold=1.5)
> image(c(1:dim(test)[2]),c(1:dim(test)[1]),t(test.b1),ylab="features",xlab="conditions")
>
> bimaxbic<-biclust(test.b1,method=BCBimax(),minr=10,minc=5,number=2)
> summary(bimaxbic)

An object of class Biclust

call:
	biclust(x = test.b1, method = BCBimax(), minr = 10, minc = 5,
	   number = 2)

Number of Clusters found:  2

Cluster sizes:
                      BC 1 BC 2
Number of Rows:       31   15
Number of Columns:     5    6
```

Solutions

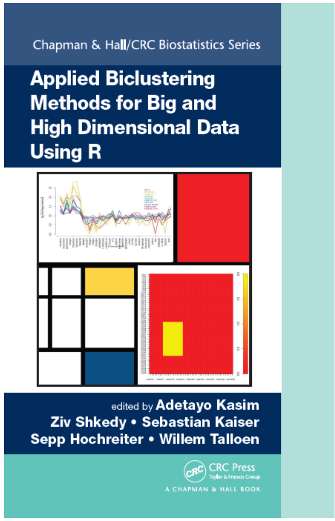






Part 3.2

The plaid model



Chapman & Hall/CRC Biostatistics Series

Applied Biclustering Methods for Big and High Dimensional Data Using R

edited by Adetayo Kasim
Ziv Shkedy • Sebastian Kaiser
Sepp Hochreiter • Willem Talloen

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapter 6

Paper:
Turner, H., Bailey, T. and Krzanowski, W. (2005) Improved biclustering of microarray data demonstrated through systsystem performance tests. *Com-putational Statistics and Data Analysis*, 48, 235–254.

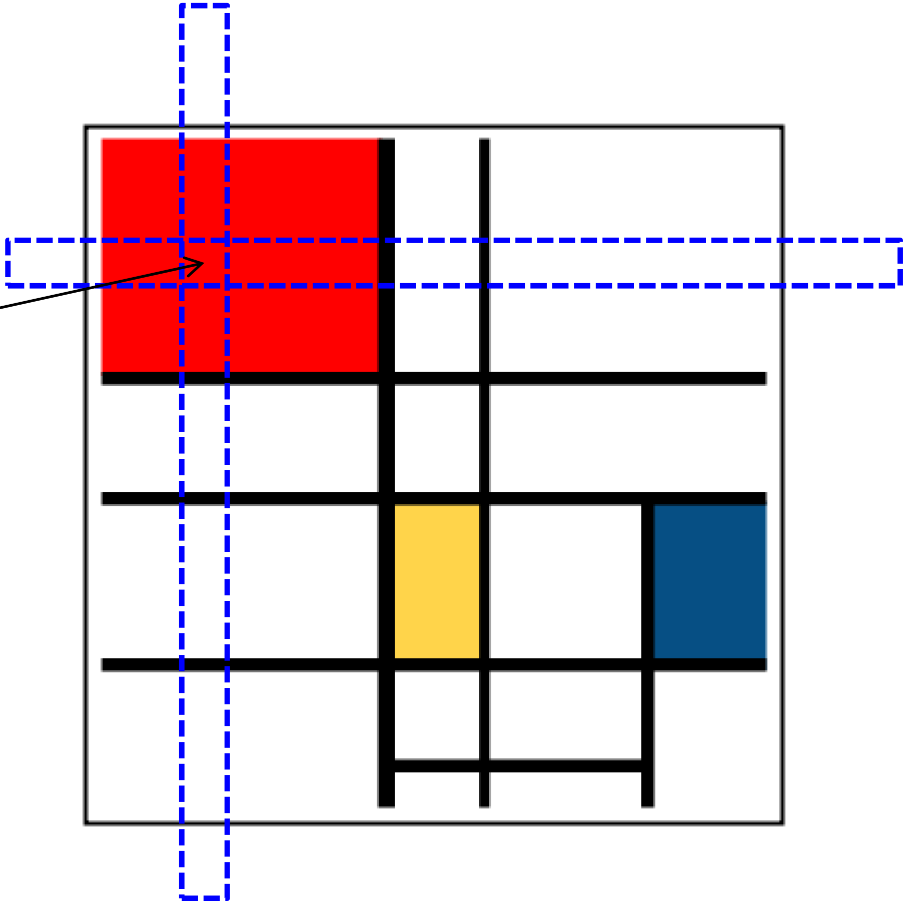
Additive biclusters: the signal structure

Continuous data.

The signal structure inside a bicluster:

$$Y = signal + noise$$

$$Y = con. + row + column + noise$$

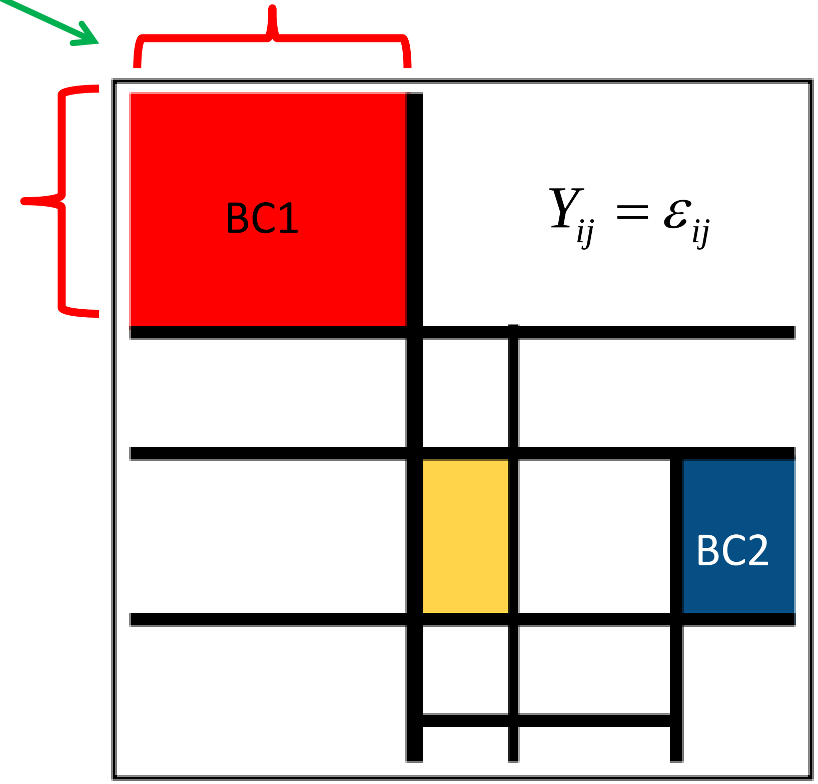


The plaid model

$$Y_{ij1} = \mu_1 + \alpha_{i1} + \beta_{j1} + \varepsilon_{ij1}$$

α_{i1} & β_{j1} :

Rows and columns effect in BC1



$$Y_{ij} = \varepsilon_{ij}$$

Rows and columns effect in BC2

$$Y_{ij2} = \mu_2 + \alpha_{i2} + \beta_{j2} + \varepsilon_{ij2}$$

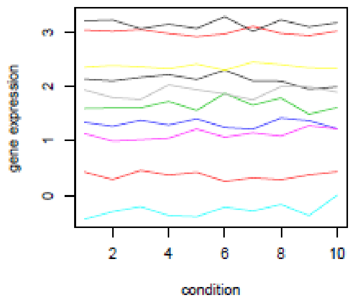
Constant effect specific for BC2

The plaid model: mean structure

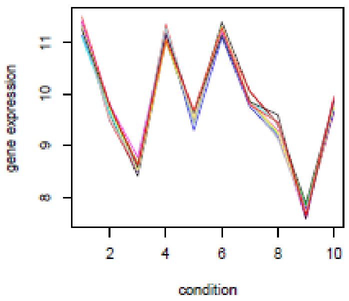
$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

$$\theta_{ijk} = \begin{cases} \mu_k & \text{Constant biclsuter} \\ \mu_k + \alpha_{ik} & \text{Constant rows} \\ \mu_k + \beta_{jk} & \text{Constant cols.} \\ \mu_k + \alpha_{ik} + \beta_{jk} & \text{Rows and cols. efects} \end{cases}$$

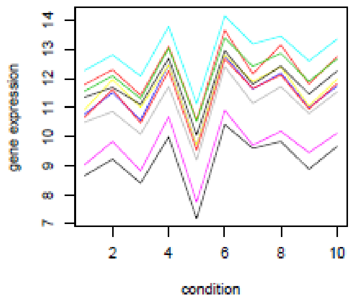
expression levels: constant rows



expression levels: constant columns



expression levels: coherent values



The plaid model: membership

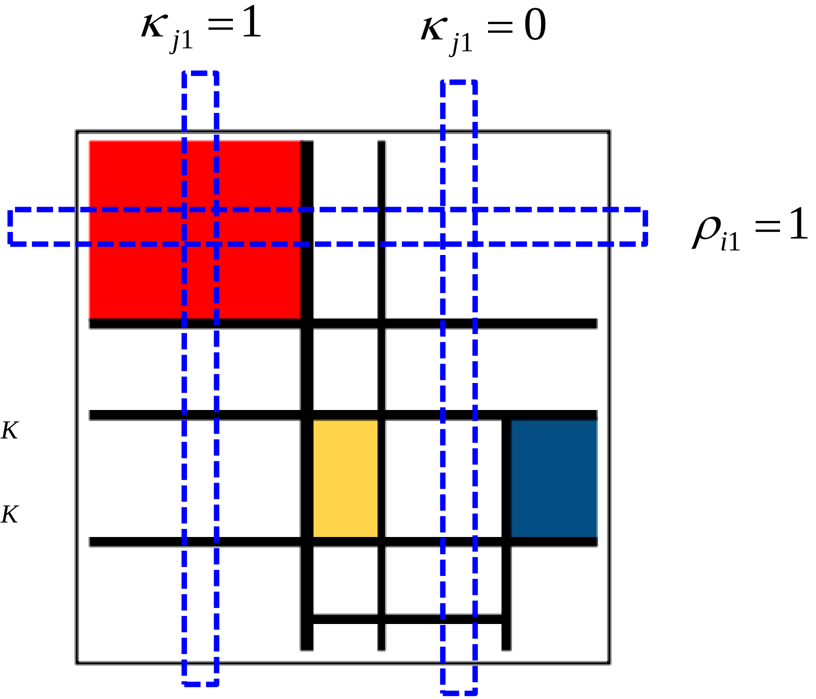
$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

$$\kappa_{jk} = \begin{cases} 1 & \text{condition } j \text{ belongs to bicluster } k, \\ 0 & \text{otherwise.} \end{cases}$$

$$\rho_{ik} = \begin{cases} 1 & \text{gene } i \text{ belongs to bicluster } k, \\ 0 & \text{otherwise,} \end{cases}$$

Membership in the first BC

$$Y_{ijk} = \begin{cases} \mu_0 + \sum_{k=1}^{K} \theta_{ijk} + \varepsilon_{ijk} & Y_{ijk} \in BC_K \\ \mu_0 + \varepsilon_{ijk} & Y_{ijk} \notin BC_K \end{cases}$$



Estimation the BC parameters

Given the membership in the k'th BC

$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

$$Y_{ijk} = \left(\underset{\Uparrow}{\mu_k} + \alpha_{ik} + \beta_{jk} + \varepsilon_{ijk}\right) \underbrace{\rho_{ik} \times \kappa_{jk}}_{\rho_{ik} = \kappa_{jk} = 1} + \varepsilon_{ijk}$$

$$Y_{ijk} = \underbrace{\mu_k + \alpha_{ik} + \beta_{jk}}_{\theta_{ijk}} + \varepsilon_{ijk}$$

A two way ANOVA with one observation per cell.

Estimation the BC parameters

Minimize the sum of squares for the k’th BC

$$Q_k = \sum_{ij} \left(Y_{ijk} - \mu_k + \alpha_{ik} + \beta_{jk} \right)^2$$

For all BCs

$$Q = \sum_{k=1}^{K} \sum_{ij} \left(Y_{ijk} - \mu_k + \alpha_{ik} + \beta_{jk} \right)^2$$

In practice, per BC, two-way ANOVA with one observation per cell.

For detailed information, see Chapter 6 in the biclustering book !!!

Estimation the membership parameters (rows)

Given the rows and columns effects and the membership for the columns.

$$Y_{ijk} = (\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik} \times \underbrace{\kappa_{jk}}_{\kappa_{jk} = 1} + \varepsilon_{ijk}$$

$$Y_{ijk} = \rho_{ik}\left[\underbrace{(\mu_k + \alpha_{ik} + \beta_{jk}) \times \kappa_{jk}}_{\text{“Known”}}\right] + \varepsilon_{ijk}$$

The only unknown is the row membership.

Condition on the parameter Estimates for BC effects and membership (columns).

$$Y_{ijk} = \rho_{ik}\left[(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}) \times \hat{\kappa}_{jk}\right] + \varepsilon_{ijk}$$

Minimize the residuals sum of squares:

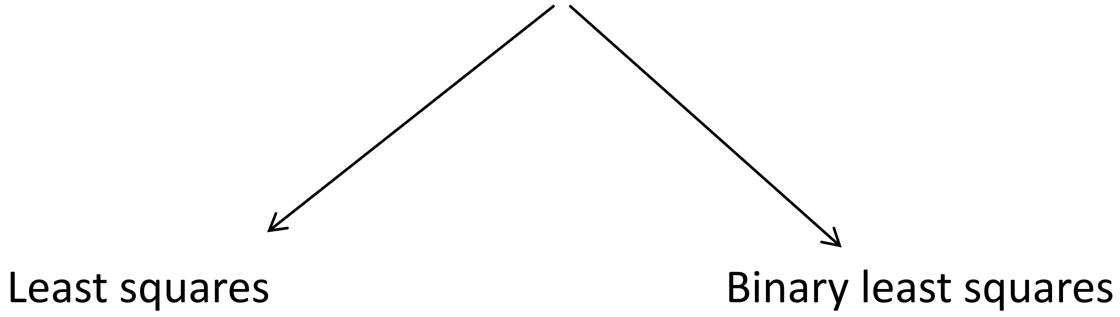
$$Q = \sum\left(Y_{ijk} - \rho_{ik}\left[(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}) \times \hat{\kappa}_{jk}\right]\right)^2$$

Estimation the membership parameters (rows)

Condition on the parameter estimates, linear regression model with one parameter

$$Y_{ijk} = \rho_{ik}\left[\left(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}\right)\times \hat{\kappa}_{jk}\right] + \varepsilon_{ijk}$$

$$Q = \sum \left(Y_{ijk} - \rho_{ik}\left[\left(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}\right)\times \hat{\kappa}_{jk}\right]\right)^2$$



Least squares

Binary least squares

See later

Estimation the membership parameters (columns)

Given the rows and columns effects and the membership for the rows

$$Y_{ijk} = (\mu_k + \alpha_{ik} + \beta_{jk} + \varepsilon_{ijk})\underbrace{\rho_{ik} \times \kappa_{jk}}_{\rho_{ik} = 1} + \varepsilon_{ijk}$$

$$Y_{ijk} = \kappa_{jk}\left\lfloor \underbrace{(\mu_k + \alpha_{ik} + \beta_{jk}) \times \rho_{ik}}_{\text{“Known”}} \right\rfloor + \varepsilon_{ijk}$$

The only unknown is the columns membership

Condition on the parameter estimates and membership (rows):

$$Y_{ijk} = \kappa_{ik}\left\lfloor(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}) \times \hat{\rho}_{jk}\right\rfloor + \varepsilon_{ijk}$$

Estimation the membership parameters (columns)

Condition on the parameter estimates, linear regression model with one parameter

$$Y_{ijk} = \kappa_{ik} \left[\left(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk} \right) \times \hat{\rho}_{jk} \right] + \varepsilon_{ijk}$$

$$Q = \sum \left(Y_{ijk} - \kappa_{ik} \left[\left(\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk} \right) \times \hat{\rho}_{jk} \right] \right)^2$$

Search algorithm

Data structure for K BCs

$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

Residuals

$$Z_{ijk} = Y_{ijk} - \left(\mu_0 + \sum_{k=1}^{K} \hat{\theta}_{ijk} \hat{\rho}_{ik} \hat{\kappa}_{jk} \right)$$

Observed data

Estimated BC and
membership parameters

Search algorithm

Let us assume that L-1 BCs were found and we are looking for the L'th BC

$$Y_{ijk} = \mu_0 + \sum_{k=1}^{L-1} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

Residuals:

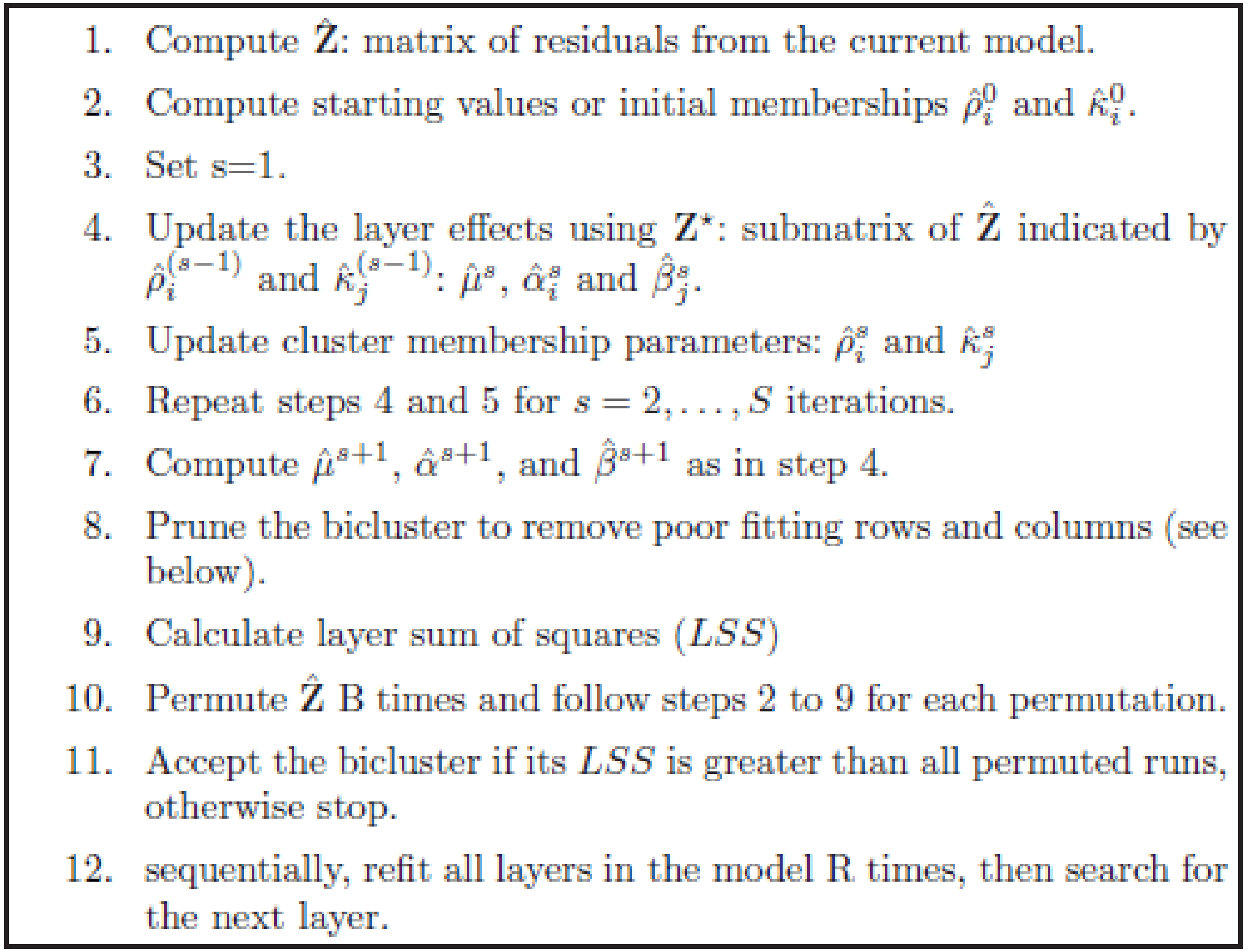
$$\hat{Z}_{ijk} = Y_{ijk} - \left(\mu_0 + \sum_{k=1}^{L-1} \theta_{ijk} \rho_{ik} \kappa_{jk} \right)$$

Residuals matrix:

$\hat{Z}$ The input matrix for the next BC (the L'th BC)

Search algorithm

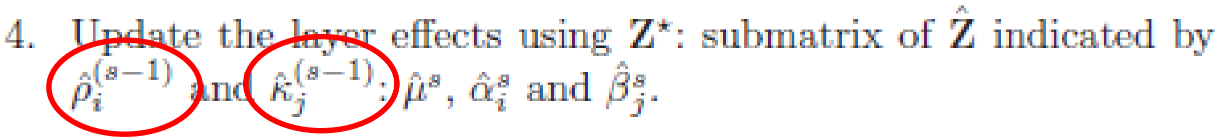
Input for the analysis of the L’th BC $\hat{Z}$



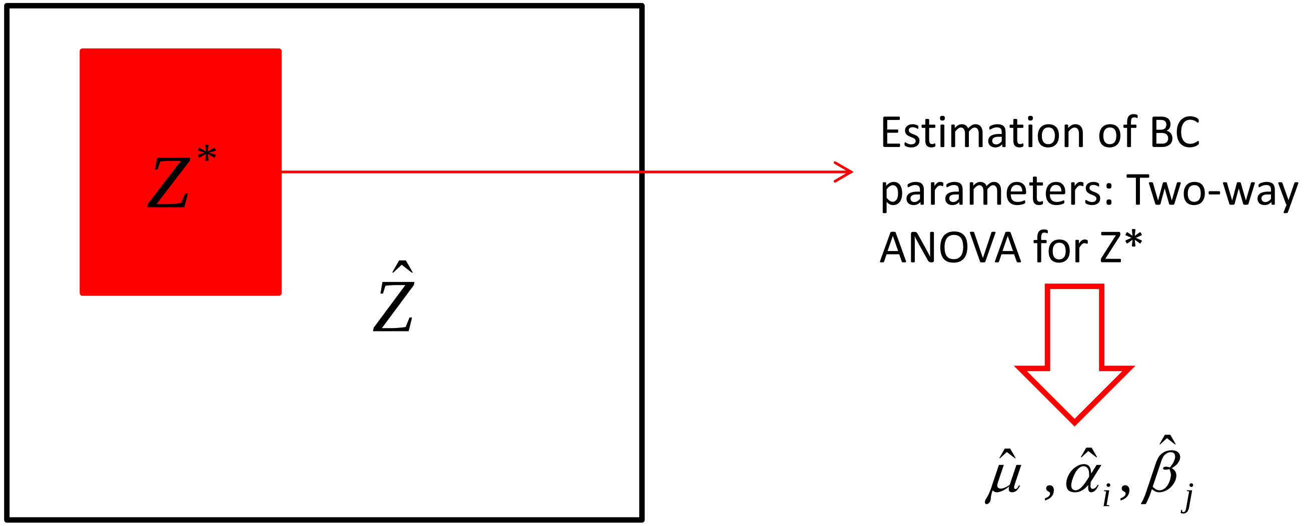
1. Compute $\hat{\mathbf{Z}}$: matrix of residuals from the current model.
2. Compute starting values or initial memberships $\hat{\rho}_i^0$ and $\hat{\kappa}_i^0$.
3. Set s=1.
4. Update the layer effects using $\mathbf{Z}^\star$: submatrix of $\hat{\mathbf{Z}}$ indicated by $\hat{\rho}_i^{(s-1)}$ and $\hat{\kappa}_j^{(s-1)}$: $\hat{\mu}^s$, $\hat{\alpha}_i^s$ and $\hat{\beta}_j^s$.
5. Update cluster membership parameters: $\hat{\rho}_i^s$ and $\hat{\kappa}_j^s$
6. Repeat steps 4 and 5 for $s = 2, \ldots, S$ iterations.
7. Compute $\hat{\mu}^{s+1}$, $\hat{\alpha}^{s+1}$, and $\hat{\beta}^{s+1}$ as in step 4.
8. Prune the bicluster to remove poor fitting rows and columns (see below).
9. Calculate layer sum of squares (LSS)
10. Permute $\hat{\mathbf{Z}}$ B times and follow steps 2 to 9 for each permutation.
11. Accept the bicluster if its LSS is greater than all permuted runs, otherwise stop.
12. sequentially, refit all layers in the model R times, then search for the next layer.

Search algorithm

Input for the analysis of the L’th BC $\hat{Z}$



4. Update the layer effects using $\mathbf{Z}^*$: submatrix of $\hat{\mathbf{Z}}$ indicated by $\hat{\rho}_i^{(s-1)}$ and $\hat{\kappa}_j^{(s-1)}$: $\hat{\mu}^s$, $\hat{\alpha}_i^s$ and $\hat{\beta}_j^s$.



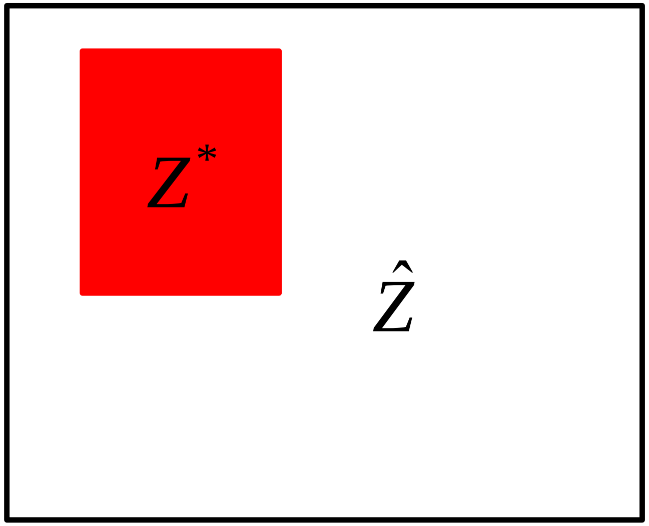
Estimation of BC parameters: Two-way ANOVA for Z*

$\hat{\mu}\ ,\hat{\alpha}_i,\hat{\beta}_j$

Search algorithm

Input for the analysis of the L’th BC $\hat{Z}$

5. Update cluster membership parameters: $\hat{\rho}_i^s$ and $\hat{\kappa}_j^s$

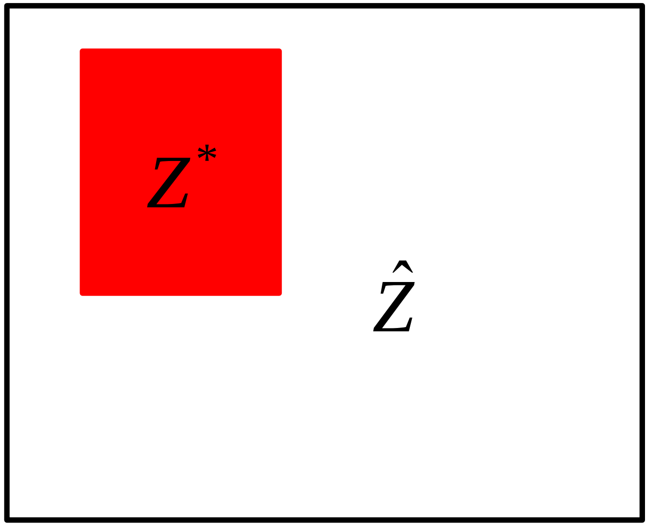


BC parameters are fixed form step 4

$$\hat{\mu}\ ,\hat{\alpha}_i,\hat{\beta}_j$$

Estimation the membership parameters: least squares (rows)

For the current BC:



$$Z_{ijk} = \rho_{ik}\left[\left(\underbrace{\hat{\mu}_k + \hat{\alpha}_{ik} + \hat{\beta}_{jk}}_{\hat{\theta}_{ijk}\ \text{fixed form step 4}}\right) \times \hat{\kappa}_{jk}\right] + \varepsilon_{ijk}$$

$$Z_{ijk} = \rho_{ik}\left[\hat{\theta}_{ijk} \times \hat{\kappa}_{jk}\right] + \varepsilon_{ijk}$$

$$Q = \sum\left(Z_{ijk} - \rho_{ik}\left[\hat{\theta}_{ijk} \times \hat{\kappa}_{jk}\right]\right)^2$$

Assume that k_j is known.

Least squares solution

$$Q = \sum\left(Z_{ijk} - \rho_{ik}\left[\hat{\theta}_{ijk} \times \hat{\kappa}_{jk}\right]\right)^2 \quad \Longrightarrow \quad \rho_i = \frac{\Sigma_j \kappa_j \theta_{ij} Z_{ij}}{\Sigma_j \kappa_j^2 \theta_{ij}^2};$$

Estimation the membership parameters: least squares

rows

Assume that k_j is known.

$$Q = \sum \left(Z_{ijk} - \rho_{ik} \left[\hat{\theta}_{ijk} \times \hat{\kappa}_{jk} \right] \right)^2$$

⇒

$$\rho_i = \frac{\Sigma_j \kappa_j \theta_{ij} Z_{ij}}{\Sigma_j \kappa_j^2 \theta_{ij}^2};$$

columns

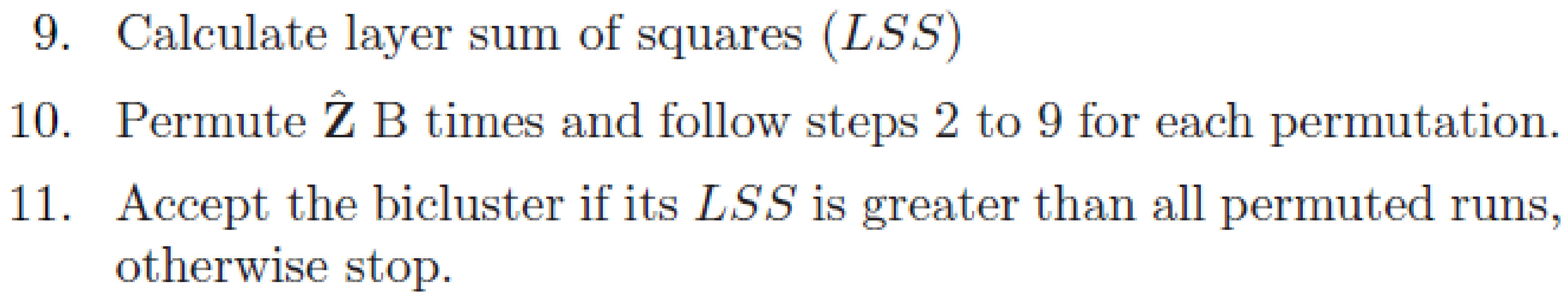
Assume that rho_i is known.

$$Q = \sum \left(Z_{ijk} - \kappa_{ik} \left[\hat{\theta}_{ijk} \times \hat{\rho}_{jk} \right] \right)^2$$

⇒

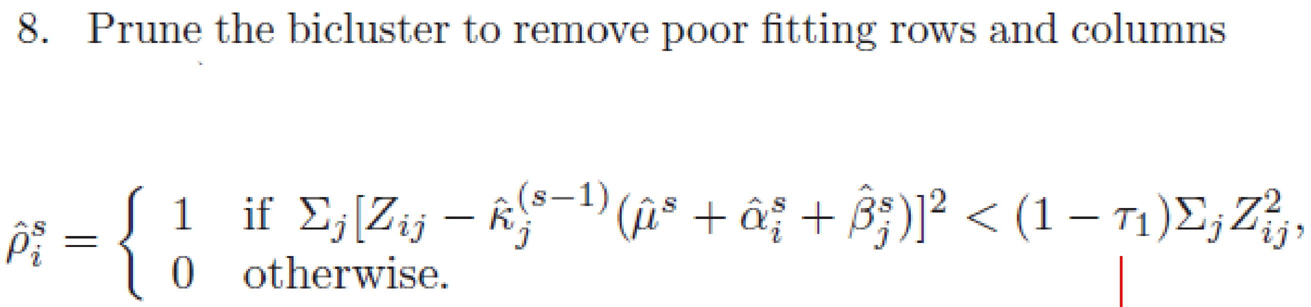
$$\kappa_j = \frac{\Sigma_i \rho_i \theta_{ij} Z_{ij}}{\Sigma_i \rho_i^2 \theta_{ij}^2}$$

Search algorithm



9. Calculate layer sum of squares (LSS)
10. Permute $\hat{\mathbf{Z}}$ B times and follow steps 2 to 9 for each permutation.
11. Accept the bicluster if its LSS is greater than all permuted runs, otherwise stop.

Search algorithm



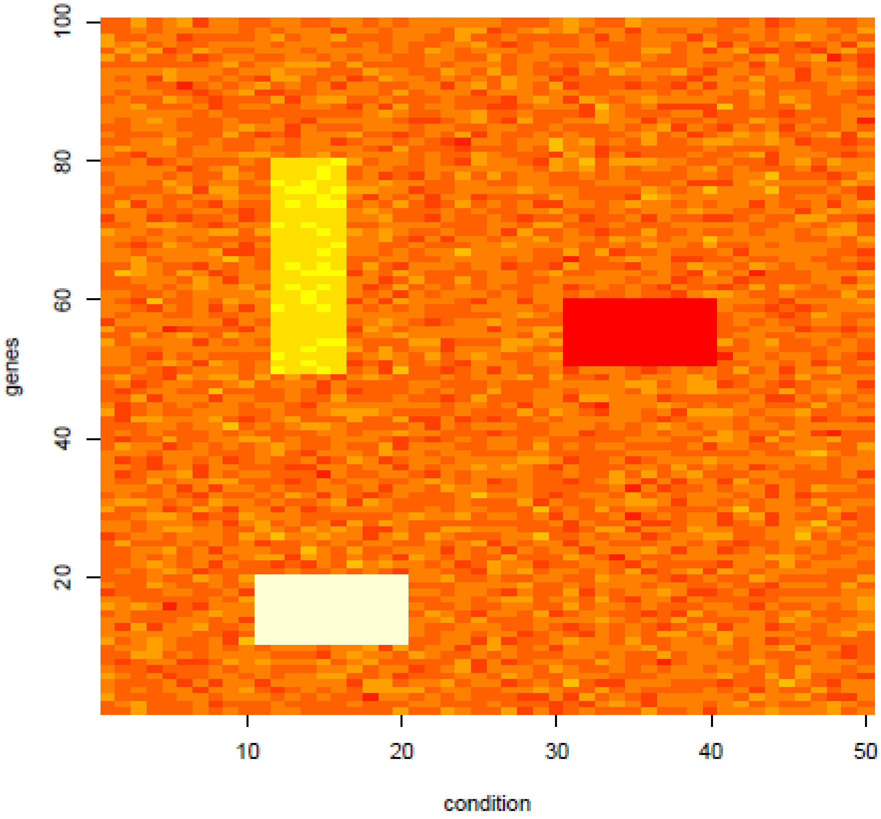
8. Prune the bicluster to remove poor fitting rows and columns

$$\hat{\rho}_i^s = \begin{cases} 1 & \text{if } \Sigma_j [Z_{ij} - \hat{\kappa}_j^{(s-1)}(\hat{\mu}^s + \hat{\alpha}_i^s + \hat{\beta}_j^s)]^2 < (1-\tau_1)\Sigma_j Z_{ij}^2, \\ 0 & \text{otherwise.} \end{cases}$$

$$0 \leq \tau_1 \leq 1$$

This means: a row is included is it leads to a reduction of τ_1 in the residuals sum of squares.

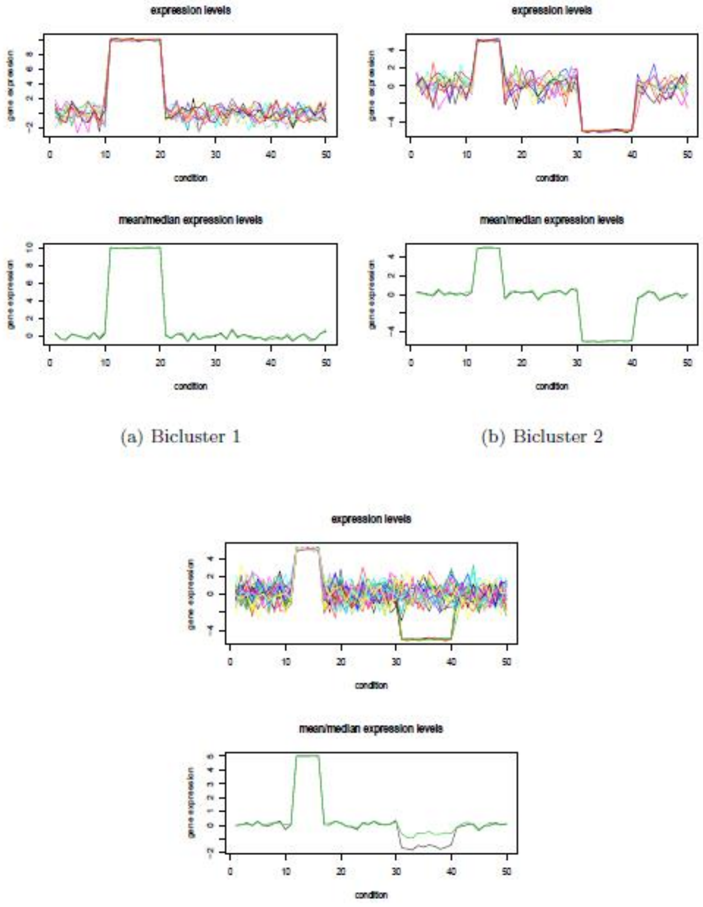
Example: the test data



A 100 X 50 data matrix with 3 BCs.

A group of rows are members in two BCs.

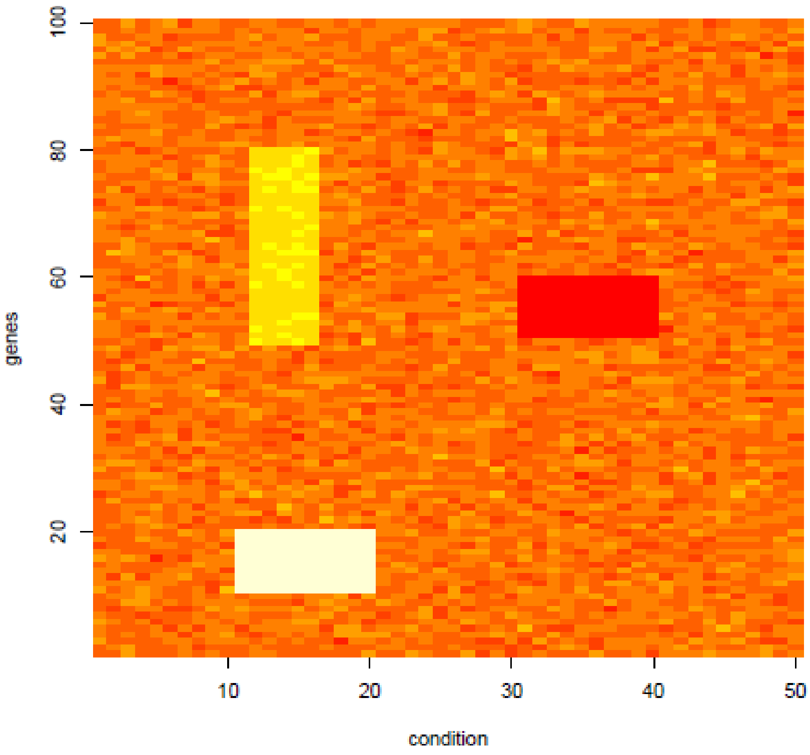
Response profiles in the three BCs



(a) Bicluster 1

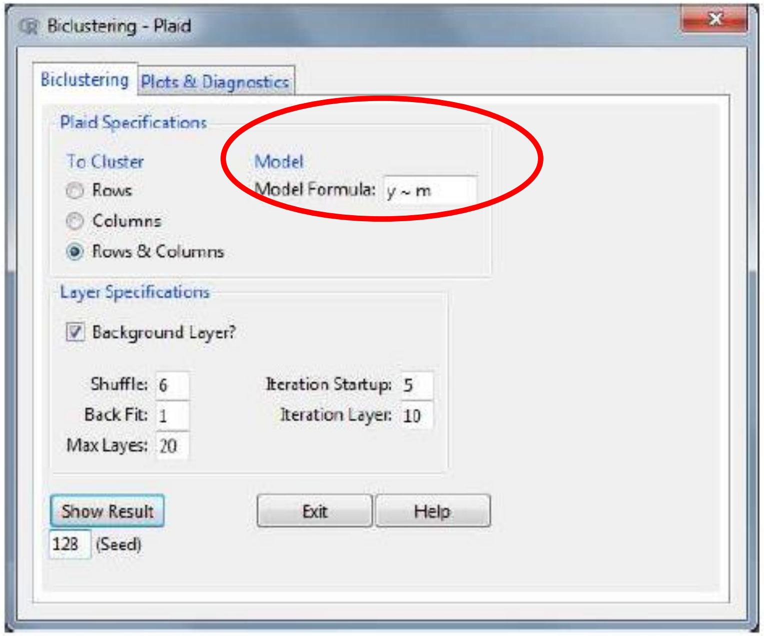
(b) Bicluster 2

(c) Bicluster 3



The plaid model in R (I)

Constant BC



Plaid Specifications

To Cluster

○ Rows

○ Columns

● Rows & Columns

Model

Model Formula: y ~ m

Layer Specifications

☑ Background Layer?

Shuffle: 6

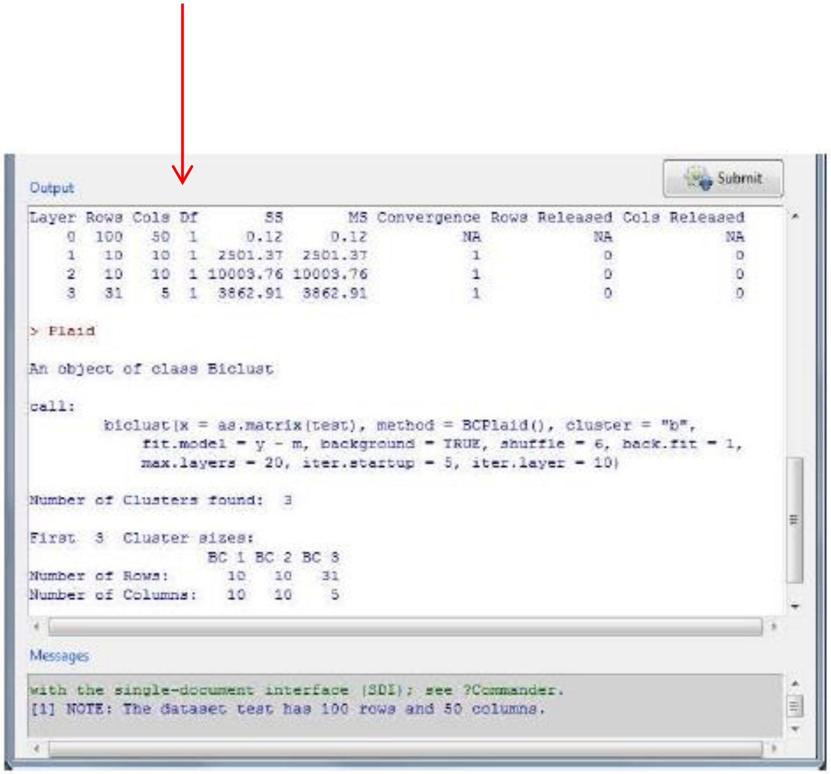
Back Fit: 1

Max Layes: 20

Iteration Startup: 5

Iteration Layer: 10

128 (Seed)



```
Layer Rows Cols Df       SS       MS Convergence Rows Released Cols Released
    0  100   50  1     0.12     0.12          NA            NA            NA
    1   10   10  1  2501.37  2501.37           1             0             0
    2   10   10  1 10003.76 10003.76           1             0             0
    3   31    5  1  3862.91  3862.91           1             0             0

> Plaid

An object of class Biclust

call:
	biclust(x = as.matrix(test), method = BCPlaid(), cluster = "b",
	    fit.model = y ~ m, background = TRUE, shuffle = 6, back.fit = 1,
	    max.layers = 20, iter.startup = 5, iter.layer = 10)

Number of Clusters found:  3

First  3  Cluster sizes:
                   BC 1 BC 2 BC 3
Number of Rows:      10   10   31
Number of Columns:   10   10    5
```

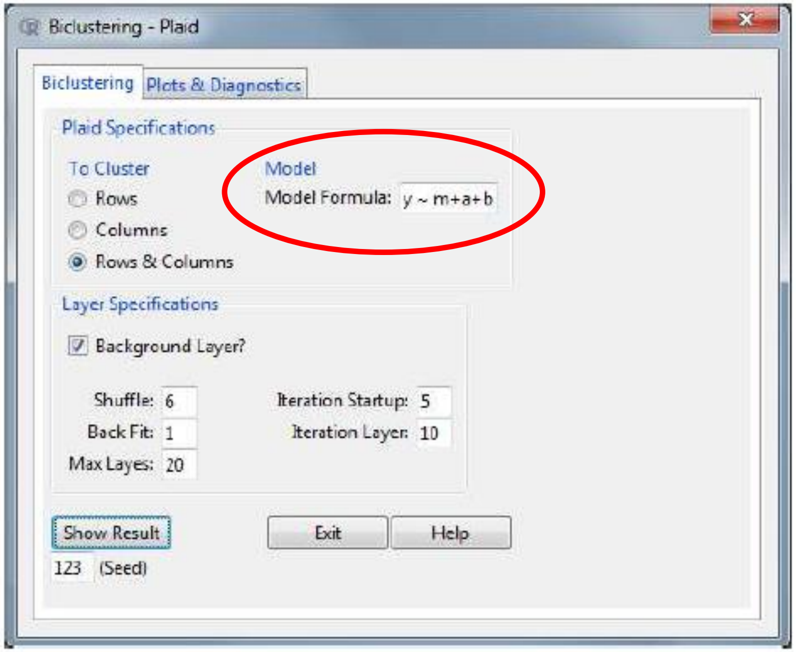
```
with the single-document interface (SDI); see ?Commander.
[1] NOTE: The dataset test has 100 rows and 50 columns.
```

$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \varepsilon_{ijk}$$

$$\theta_{ijk} = \mu_k$$

The plaid model in R (II)

Rows and columns effects



To Cluster
○ Rows
○ Columns
● Rows & Columns

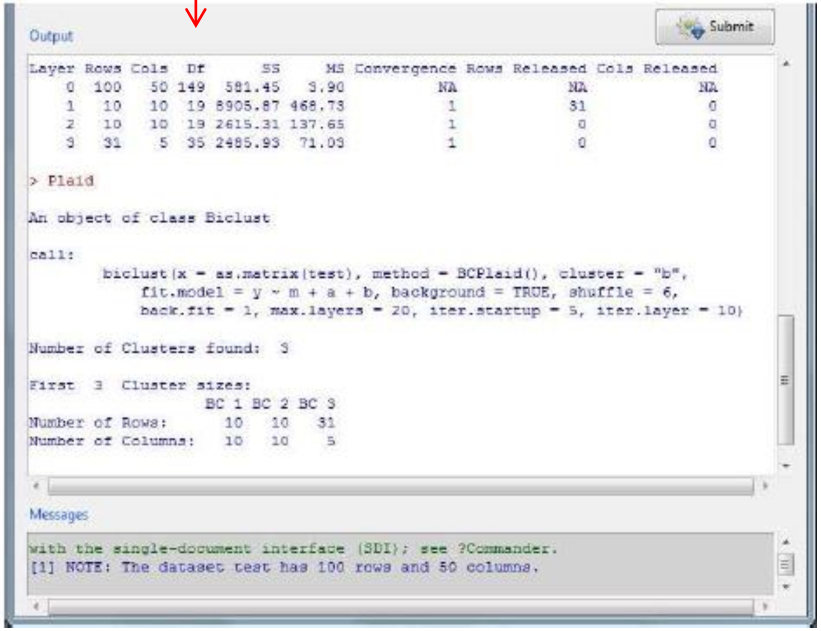
Model
Model Formula: y ~ m+a+b

☑ Background Layer?

Shuffle: 6
Back Fit: 1
Max Layes: 20

Iteration Startup: 5
Iteration Layer: 10

123 (Seed)



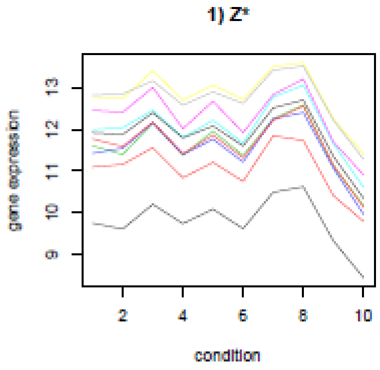
```
Layer Rows Cols  Df      SS     MS Convergence Rows Released Cols Released
    0  100   50 149  581.45   3.90          NA            NA            NA
    1   10   10  19 8905.87 468.73           1            31             0
    2   10   10  19 2615.31 137.65           1             0             0
    3   31    5  35 2485.93  71.03           1             0             0

> Plaid

An object of class Biclust

call:
        biclust(x = as.matrix(test), method = BCPlaid(), cluster = "b",
            fit.model = y ~ m + a + b, background = TRUE, shuffle = 6,
            back.fit = 1, max.layers = 20, iter.startup = 5, iter.layer = 10)

Number of Clusters found:  3

First  3  Cluster sizes:
                   BC 1 BC 2 BC 3
Number of Rows:      10   10   31
Number of Columns:   10   10    5
```

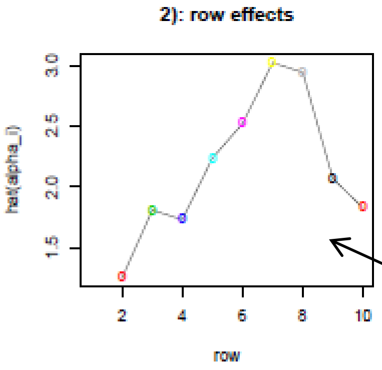
```
with the single-document interface (SDI); see ?Commander.
[1] NOTE: The dataset test has 100 rows and 50 columns.
```

$$Y_{ijk} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ijk}$$

$$\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$$

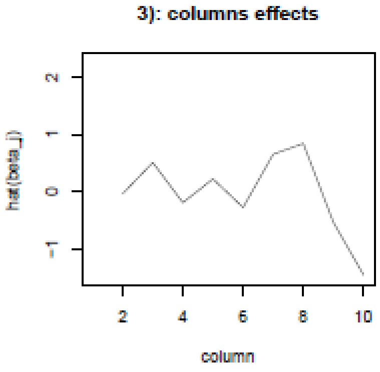
The mean structure within a BC (I)



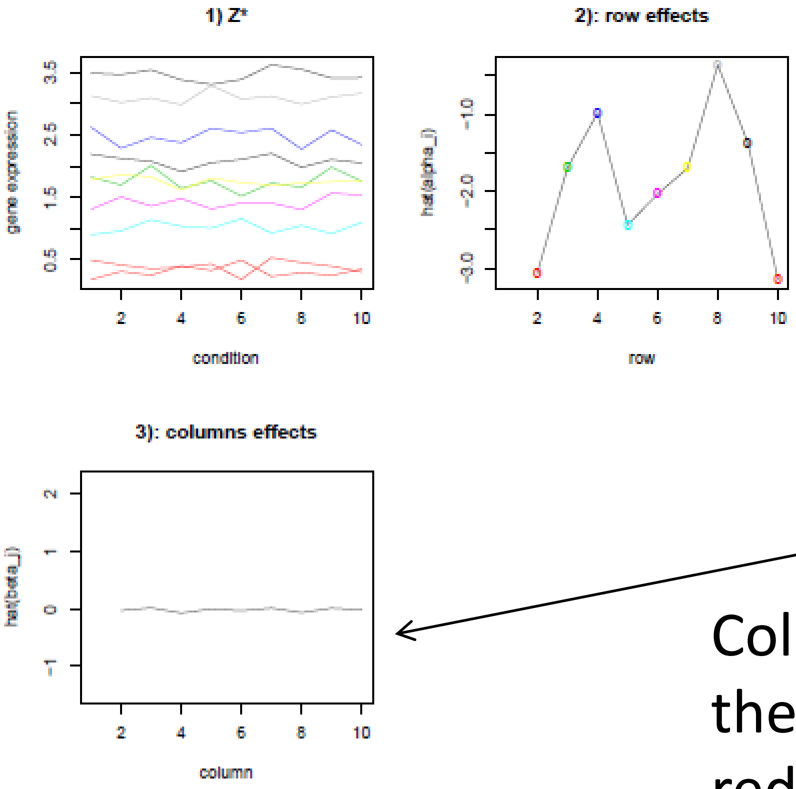


Rows and columns effects

$$Y_{ij1} = \mu_1 + \alpha_{i1} + \beta_{j1} + \varepsilon_{ij1}$$



The mean structure within a BC (II)



Only rows effects

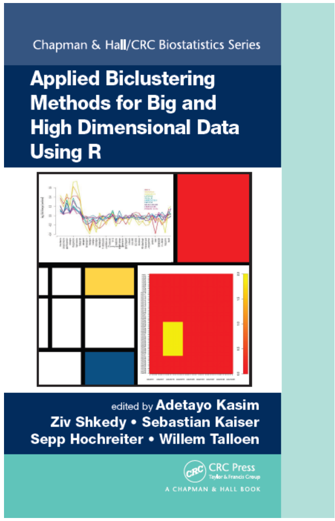
$$Y_{ij1} = \mu_1 + \alpha_{i1} + \beta_{j1} + \varepsilon_{ij1}$$

Columns effects are zero, the model can be reduced to

$$Y_{ij1} = \mu_1 + \alpha_{i1} + \varepsilon_{ij1}$$

Part 3.3

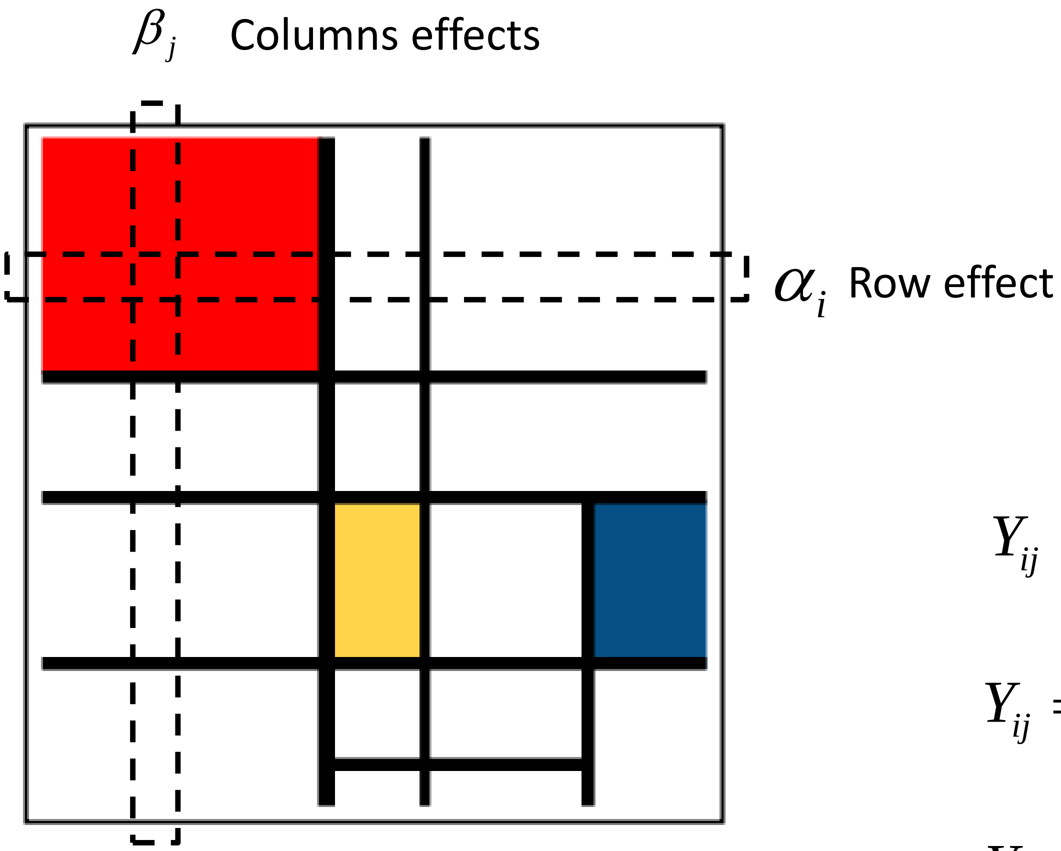
FABIA


Chapman & Hall/CRC Biostatistics Series
Applied Biclustering Methods for Big and High Dimensional Data Using R
edited by Adetayo Kasim
Ziv Shkedy • Sebastian Kaiser
Sepp Hochreiter • Willem Talloen
CRC Press
A CHAPMAN & HALL BOOK

Chapter 8

Paper:
Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., G¨ohlmann, H. W. H., Shkedy, Z. and Clevert, D.-A. (2010a) Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26, 1520–1527.

Multiplicative versus additive biclusters



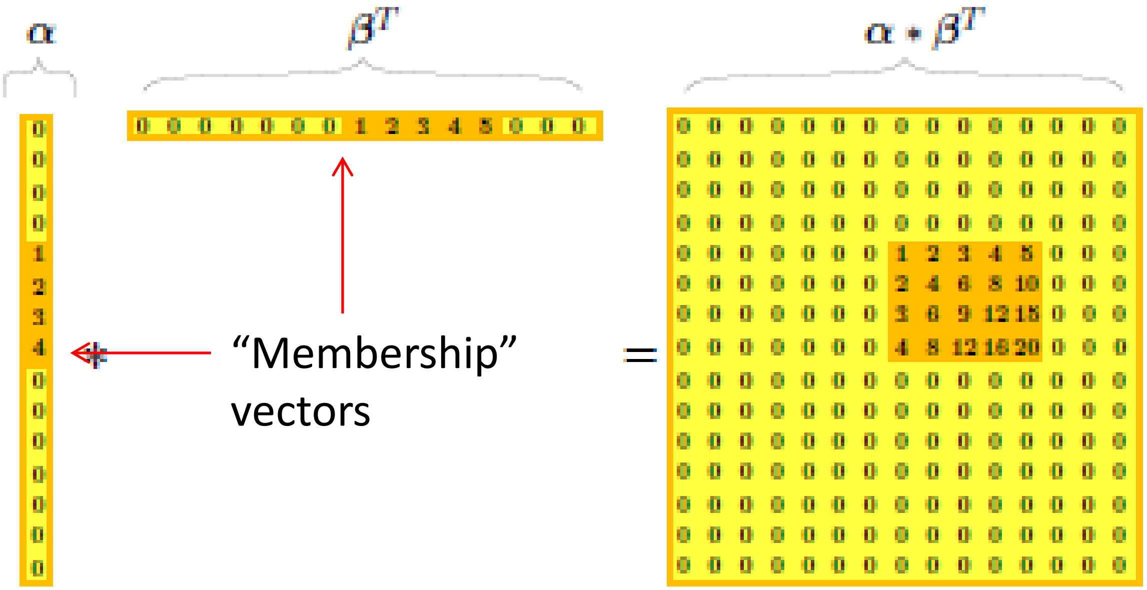
$$Y_{ij} = \mu + \alpha_i + \beta_j + error$$

$$Y_{ij} = \mu \times \alpha_i \times \beta_j + error$$

$$Y_{ij} = \alpha_i \times \beta_j + error$$

Multiplicative bicluster: signal structure

$$\beta_j = \begin{cases} \beta_j & C_j \in BC \\ 0 & C_j \notin BC \end{cases}$$



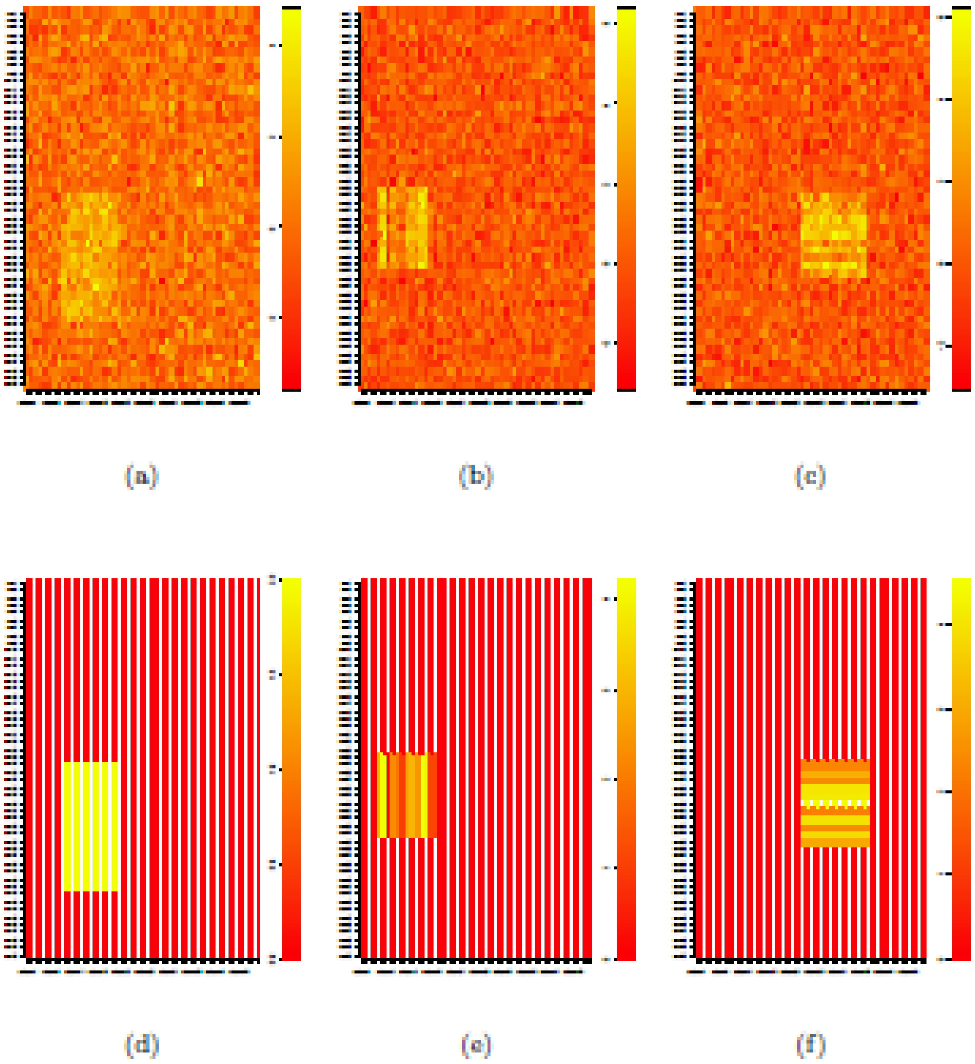
“Membership” vectors

$$\overbrace{\begin{bmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}}^{\alpha} * \overbrace{\begin{bmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{bmatrix}}^{\beta^T} = \overbrace{\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}}^{\alpha * \beta^T}$$

$$\alpha_i = \begin{cases} \alpha_i & R_i \in BC \\ 0 & R_i \notin BC \end{cases}$$

$$signal_{ij} = \alpha_i \times \beta_j$$

Examples



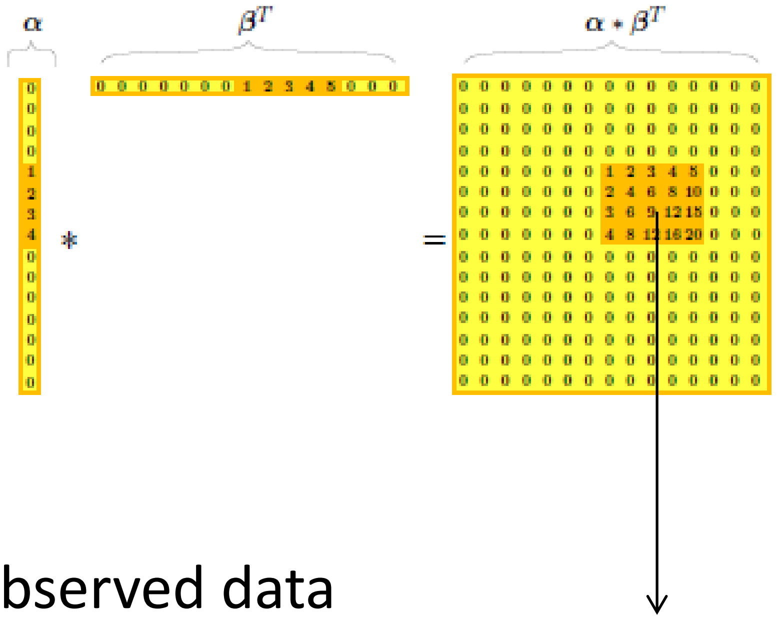
(a) (b) (c)

(d) (e) (f)

signal + noise

signal

Multiplicative model



$$\overbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}^{\alpha} * \overbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}^{\beta^T} = \overbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}^{\alpha * \beta^T}$$

A factor analysis model in which a BC is a factor.

Observed data

$$Y_{ij} = signal_{ij} + error = \alpha_i \times \beta_j + error$$

Signal structure:

$$Y \; = \underbrace{BC_1 + BC_2 +,\dots,+ BC_K}_{} + error$$

A factor analysis model with K factors.

FABIA: model formulation

$$Y = \sum_{k=1}^{K} \alpha_k \times \beta_k^T + Z$$

Rows scores

Columns scores

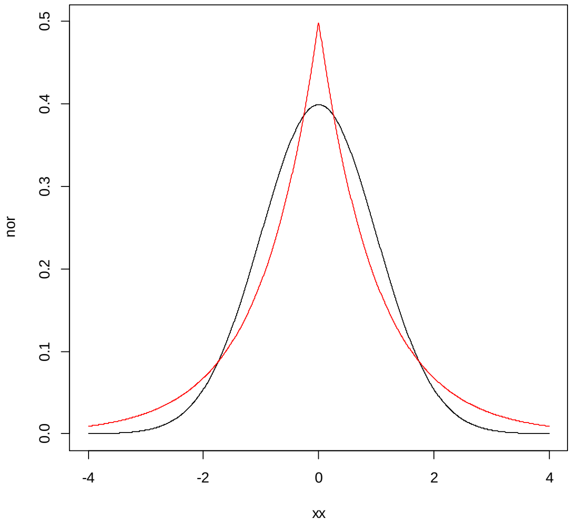
error

Model formulation of a factor analysis model with K factors.

Multiplicative signal.

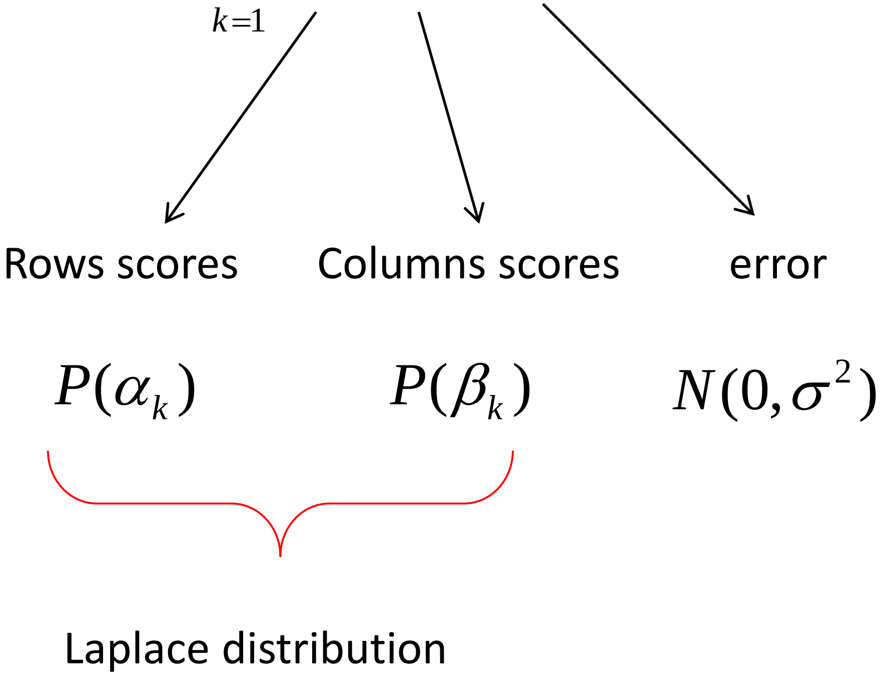
For FABIA, priors for factor loadings and scores:

$$\left.\begin{array}{l} P(\alpha_k) \\ P(\beta_k) \end{array}\right\}$$ Laplace distribution



FABIA: model formulation

$$Y = \sum_{k=1}^{K} \alpha_k \times \beta_k^T + Z$$



Rows scores

Columns scores

error

$P(\alpha_k)$ $P(\beta_k)$ $N(0,\sigma^2)$

Laplace distribution

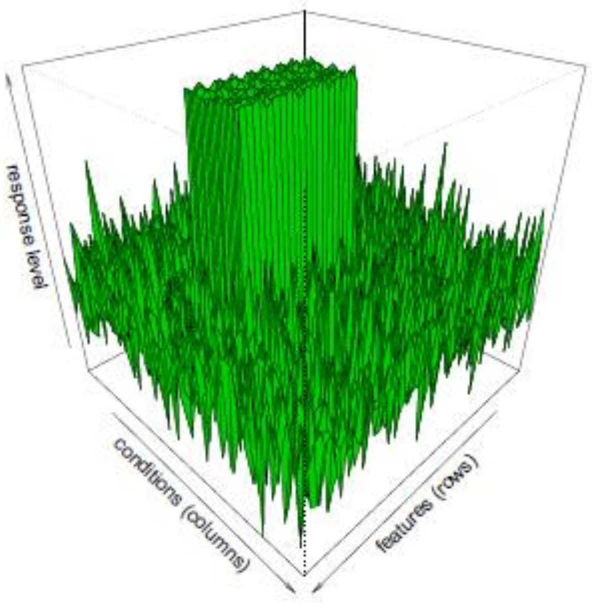
A factor analysis model:

Rows scores (membership vector for rows): factor loadings.

Columns scores (membership vector for columns): factor scores.

FABIA: example – a data matrix with one BC

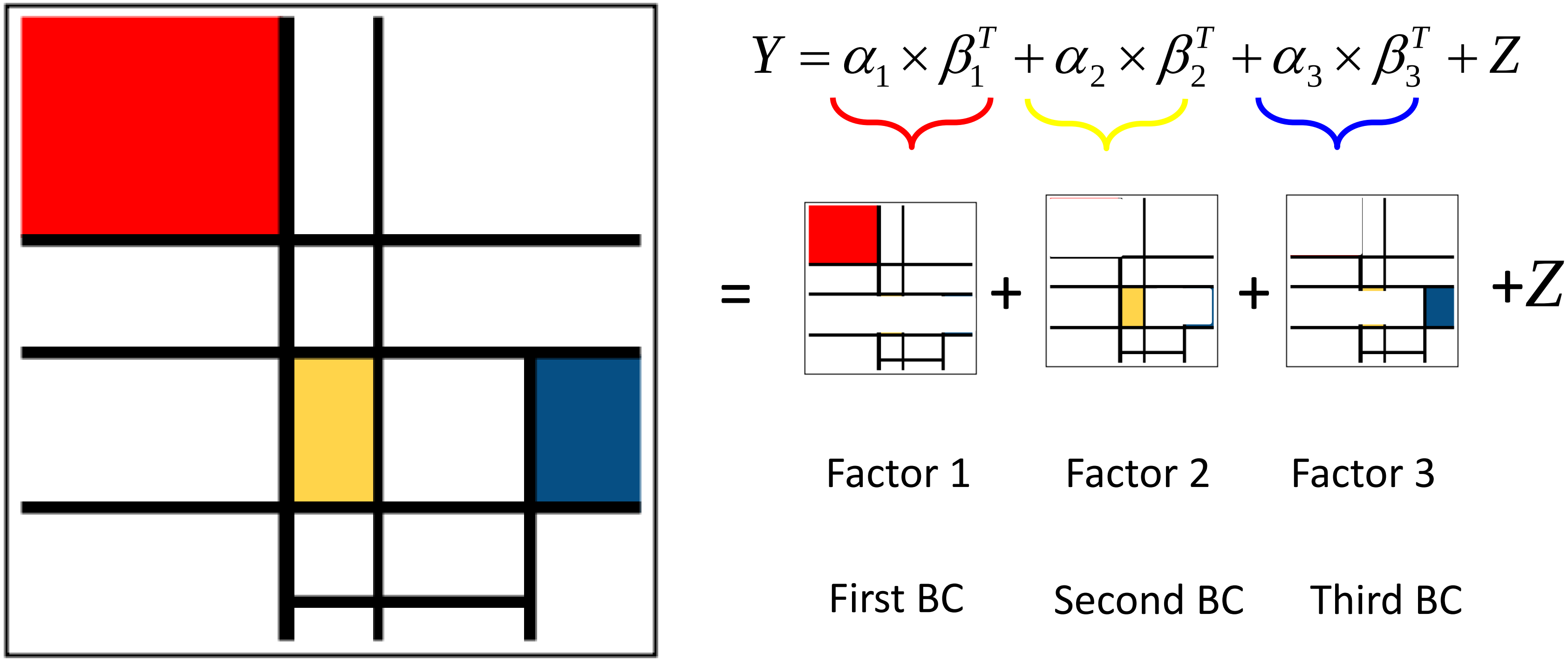
$$\overbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}^{\alpha} * \overbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}^{\beta^T} = \overbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}^{\alpha * \beta^T}$$



$$Y = \sum_{k=1}^{K} \alpha_k \times \beta_k^T + Z$$

Rows scores Columns scores error

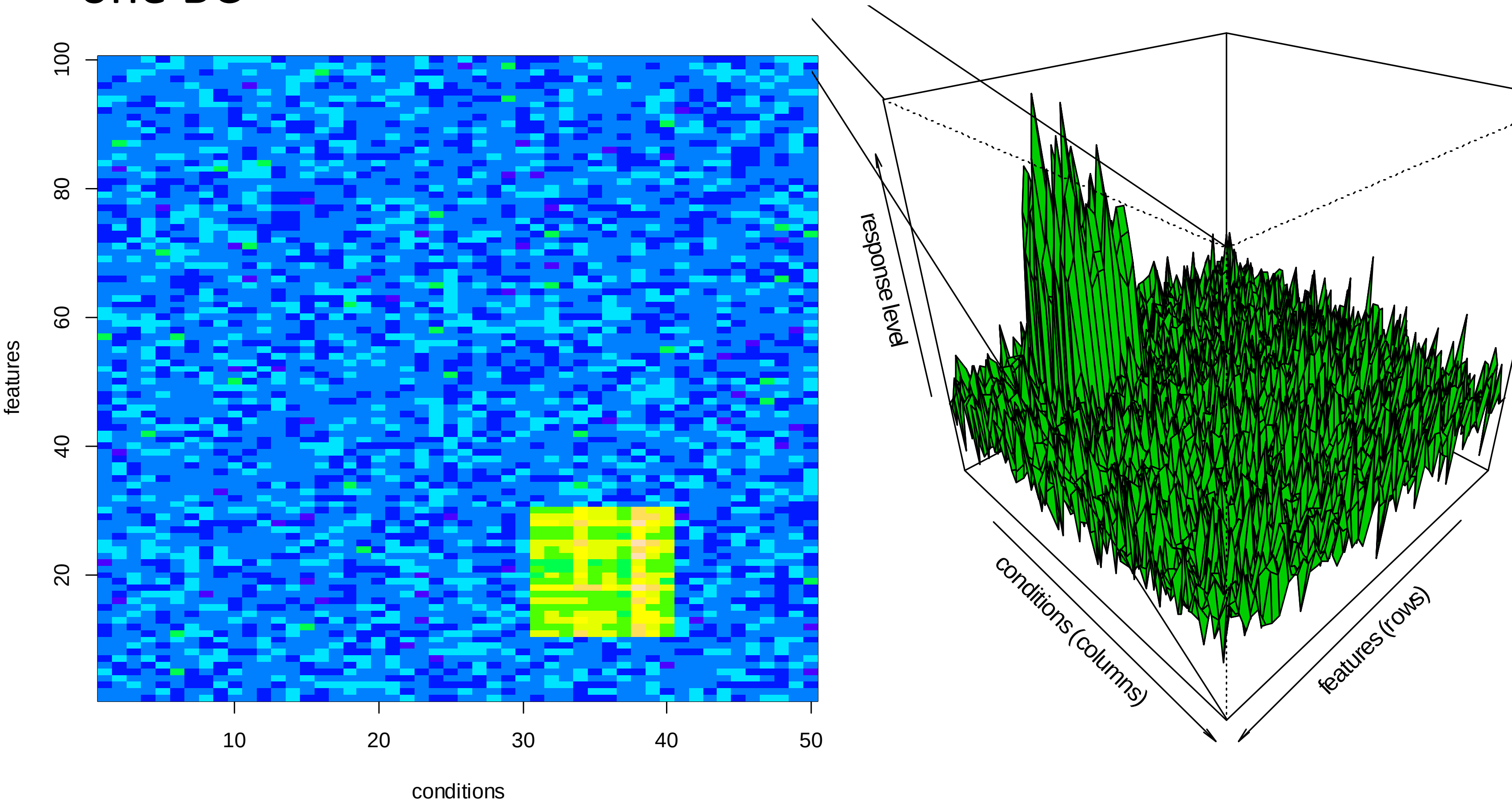
FABIA: example – a data matrix with three BCs



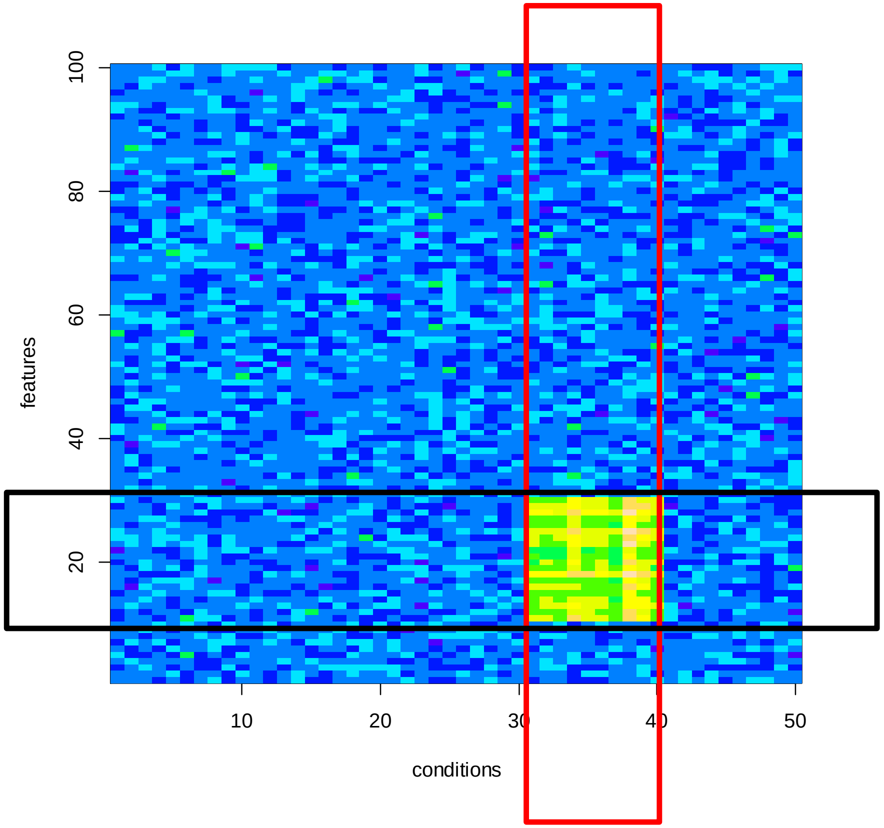
$$Y = \underbrace{\alpha_1 \times \beta_1^T}_{} + \underbrace{\alpha_2 \times \beta_2^T}_{} + \underbrace{\alpha_3 \times \beta_3^T}_{} + Z$$

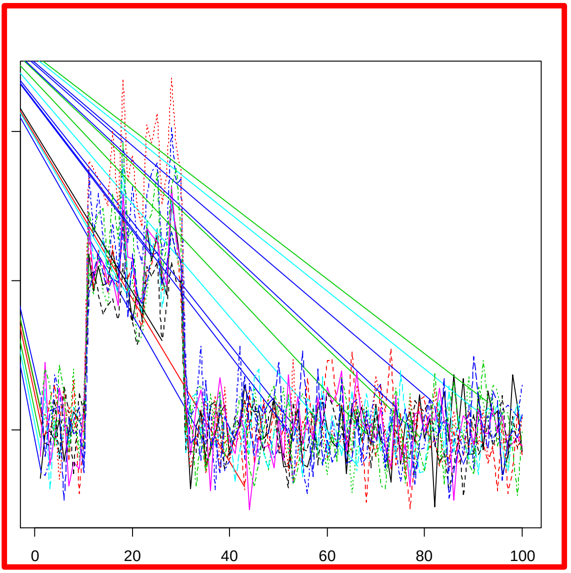
Example: one BC

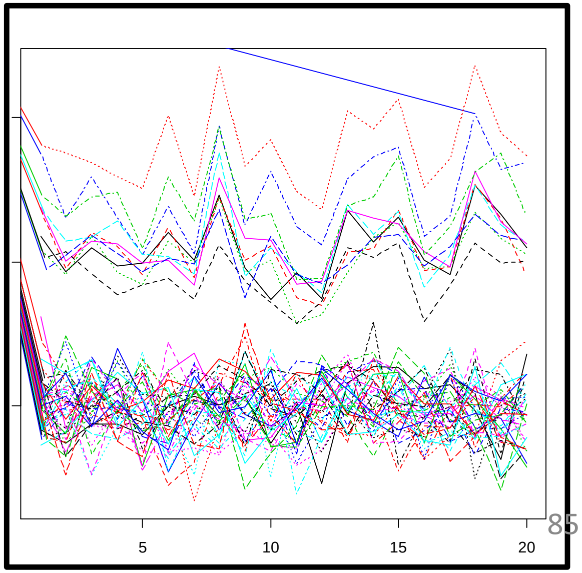
A 100X50 data matrix with one BC



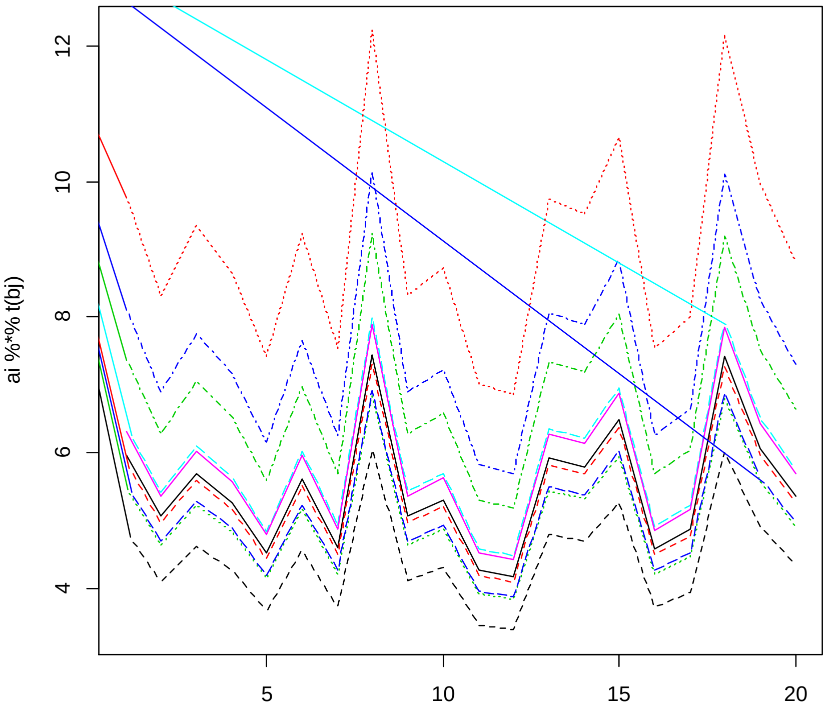
Example: one BC – rows and columns

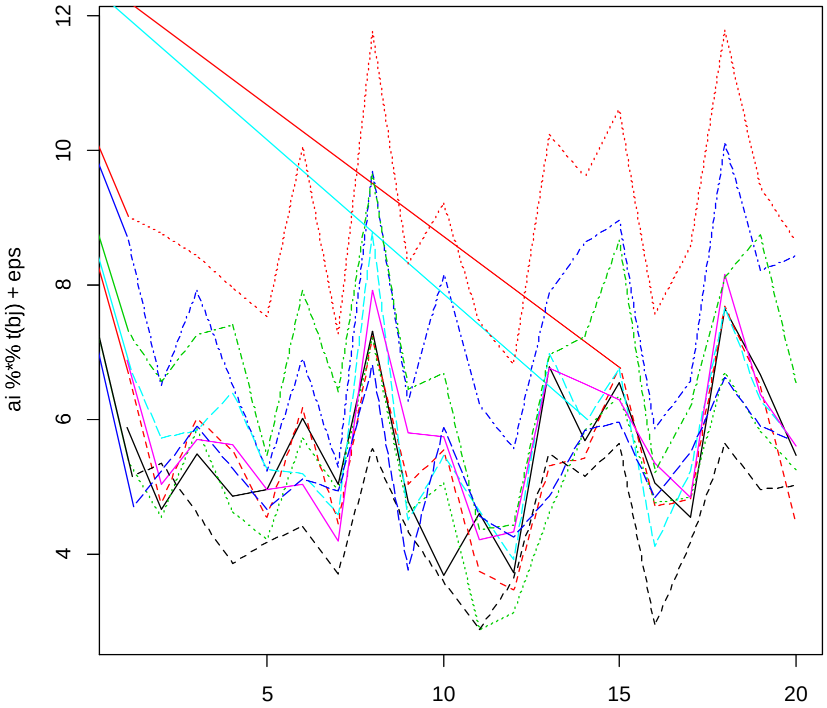






Example: signal, and signal+noise





Data analysis

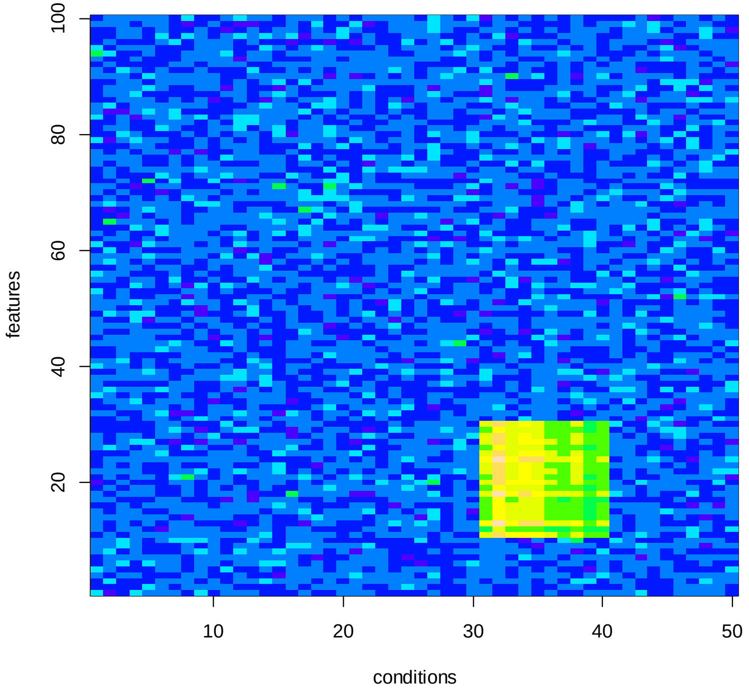
A factor analysis model with one factor:

$$Y = \alpha_1 \times \beta_2^T + Z$$

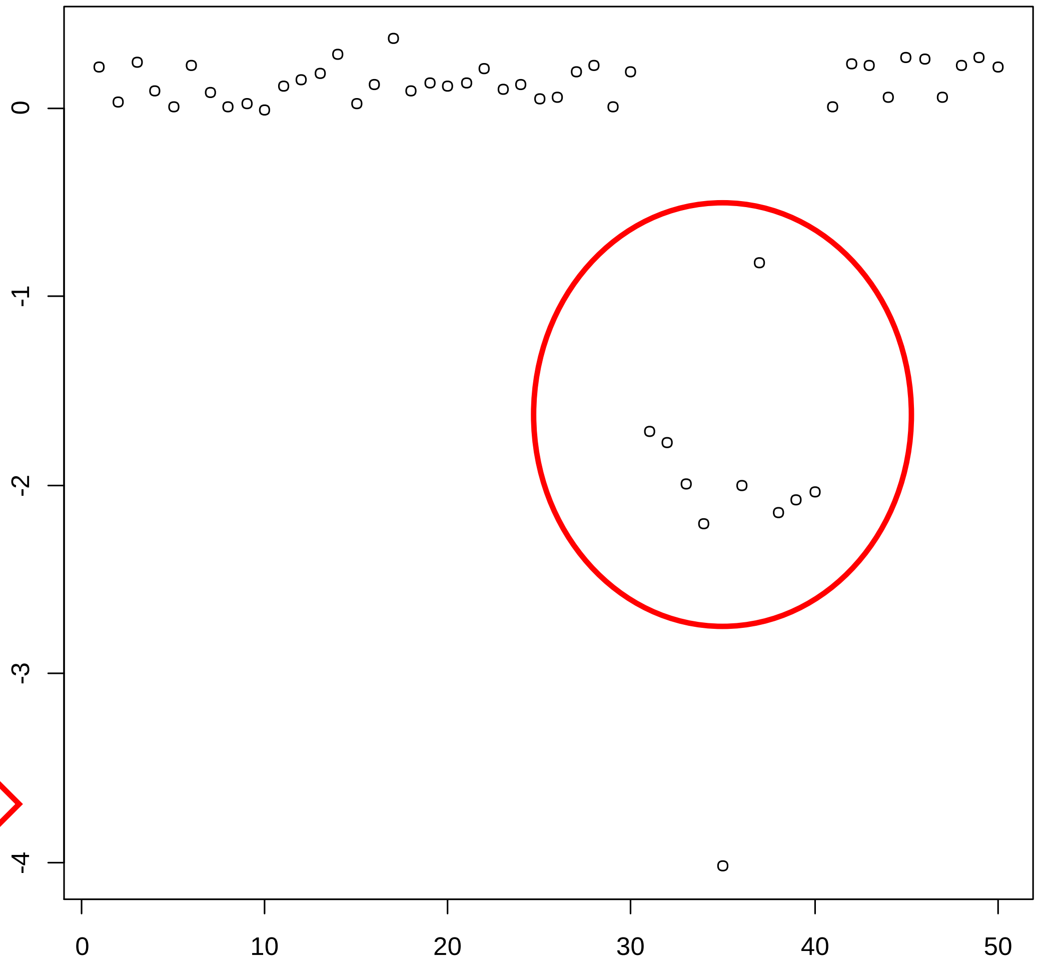
In R:

```
> fabRes <- fabia(mdat,p=1)
```

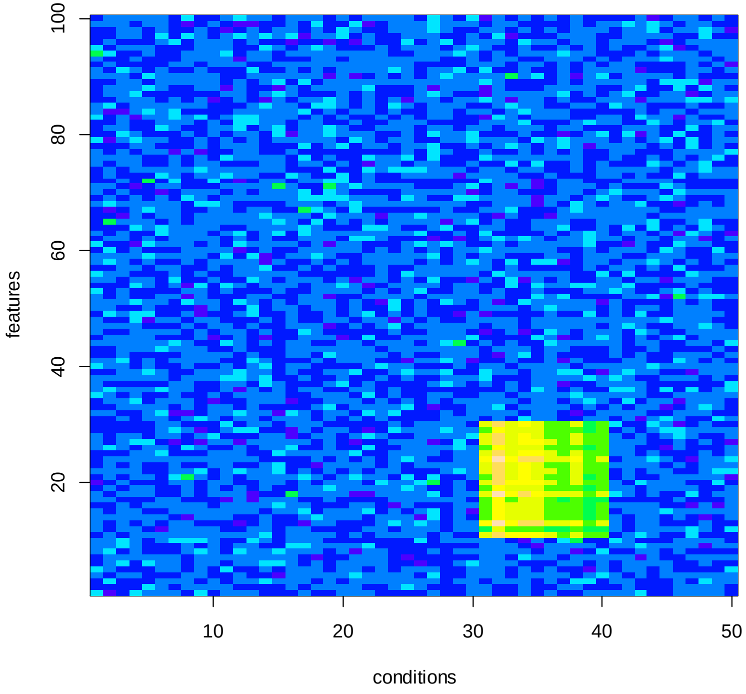
Results: factor scores (columns)

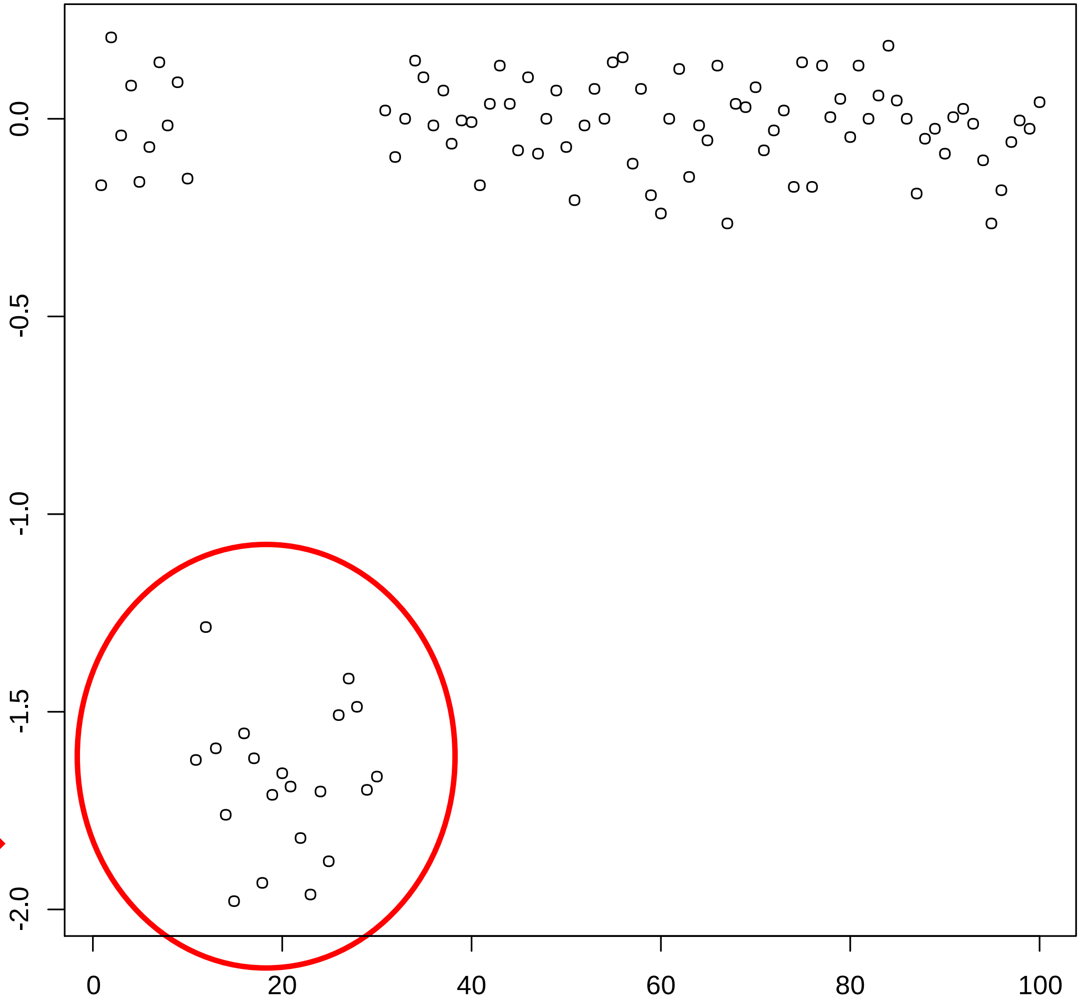


Estimates for betas



Results: factor loadings (rows)



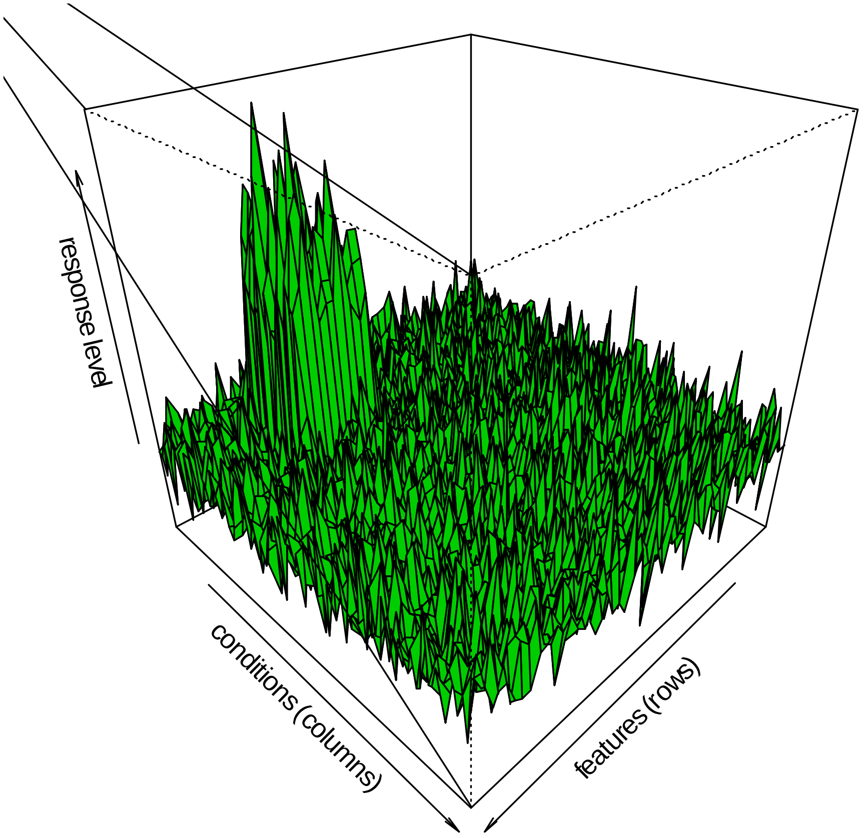


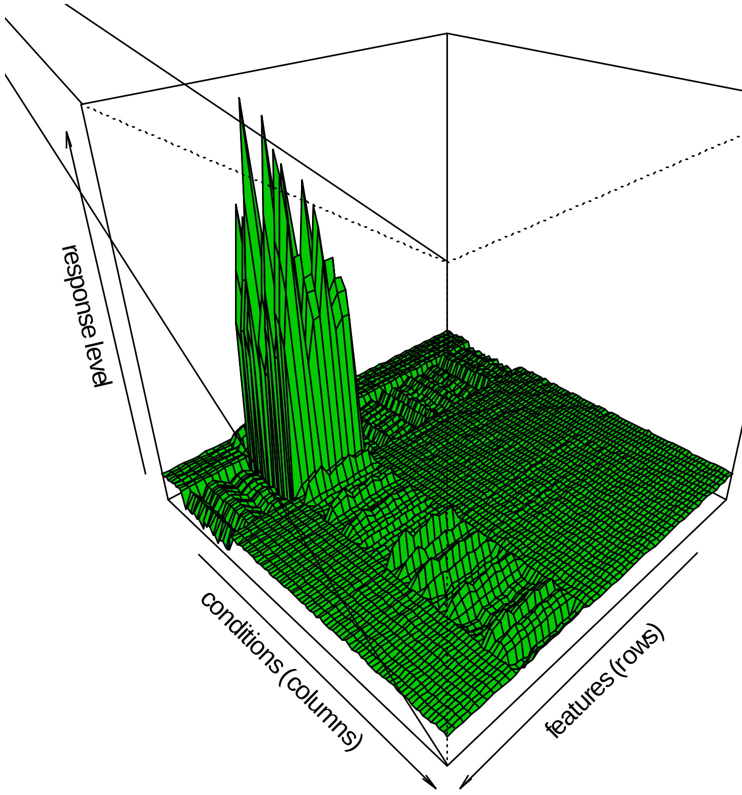
Estimates for alphas

Observed and predicted data

$$Y = \sum_{k=1}^{K} \alpha_k \beta_K^T + Z$$

$$\hat{Y} = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\beta}_K^T$$





Short summary: methods

- Many other methods were developed.
- Local patterns.
- Trying to discover the signal in a noisy data.
- Best method ? No, completely data dependent.
- For all method: subjective selection of parameter settings !
- For most of the methods, multiple runs leads to multiple results !!
- Robust analysis should be performed !!!

Short summary: software

- Method specific (many methods and packages are avilable).
- Genreal:
  - biclust.
  - biclsutGUI.
  - biclust shiny App.
  - online and cloud products.

Part 4

Case Studies

Part 4.1

Biclustering for Market segmentation

Market segmentation

- Market segmentation is essential for marketing success.
- The most successful firms drive their businesses based on segmentation.
- In tourism:
  - identify groups of tourists who share common characteristics.
  - Make it possible to develop a tailored marketing mix to most successfully attract such subgroups of the market.
  - Focusing on subgroups increases the chances of success within the subgroup.

Dimensionality problem

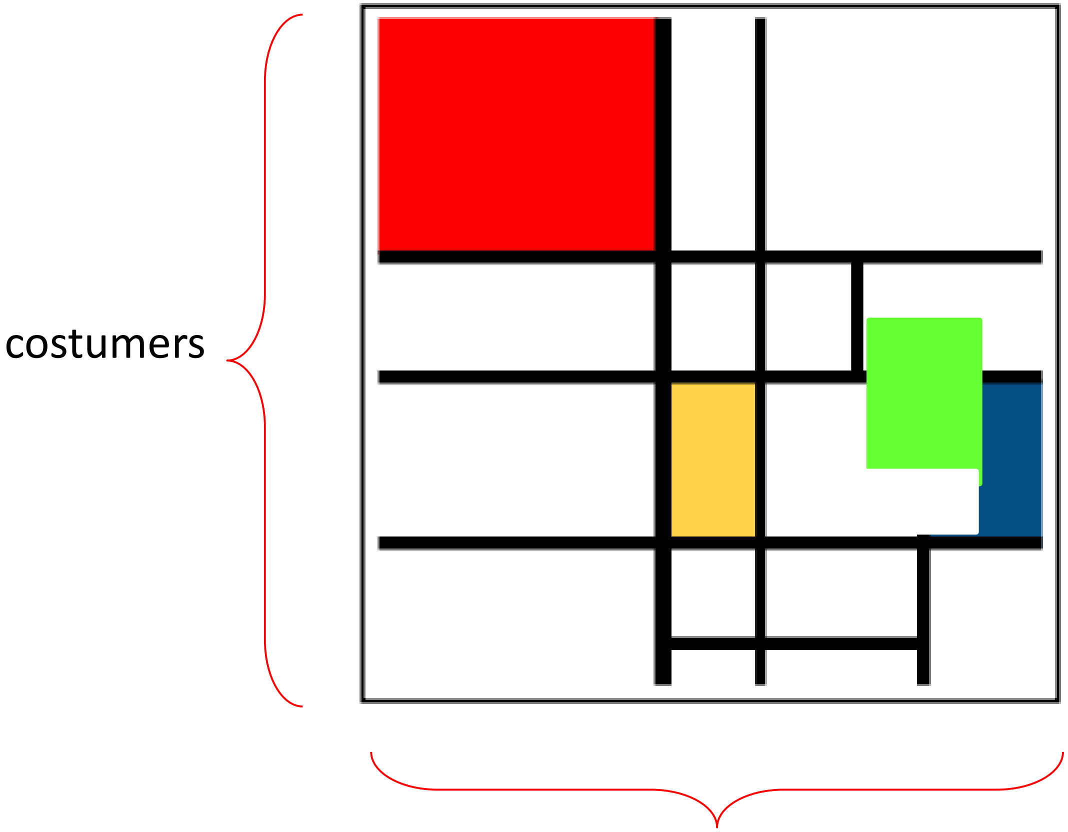
- One of the typical methodological challenge:
  - large amount of information (responses to many survey questions) is available from tourists….
  - …. But typically the sample sizes are too low given the number of variables used to conduct segmentation analysis.

- Solution: collect large samples that allow segmentation with a large number of variables.

The tourism survey

- The data set used for this illustration is a tourism survey of adult Australians (internet based survey).
- Participants were asked questions about their general travel behavior, their travel behavior on their last Australian vacation, benefits they perceive of undertaking travel, and image perceptions of their ideal tourism destination.
- Information was also collected about the participants age, gender, annual household income, marital status, education level, occupation, family structure, and media consumption.

Data structure



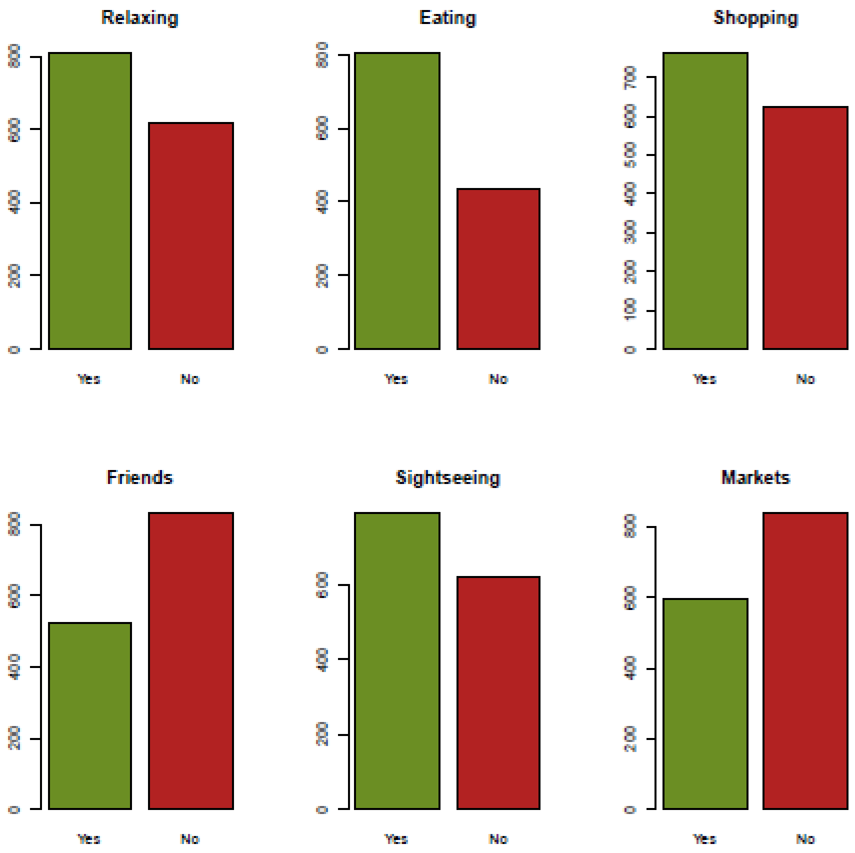
Holiday items = vacation activities

A biclsuter:
A group of tourists that share the same vacation activities.

Consumption of holiday activities

- In the present data set 1,003 respondents were asked to state for 44 vacation activities whether they engaged in them during their last vacation.
- Activities includes: relaxing, eating in reasonably priced eateries, shopping, sightseeing, visiting industrial attractions (such as wineries, breweries, mines, etc.), going to markets, scenic walks, visiting museums and monuments, botanic and public gardens, and the countryside/farms.

Consumption of holiday activities



Distribution per item gives information how popular is an item among the costumers but…

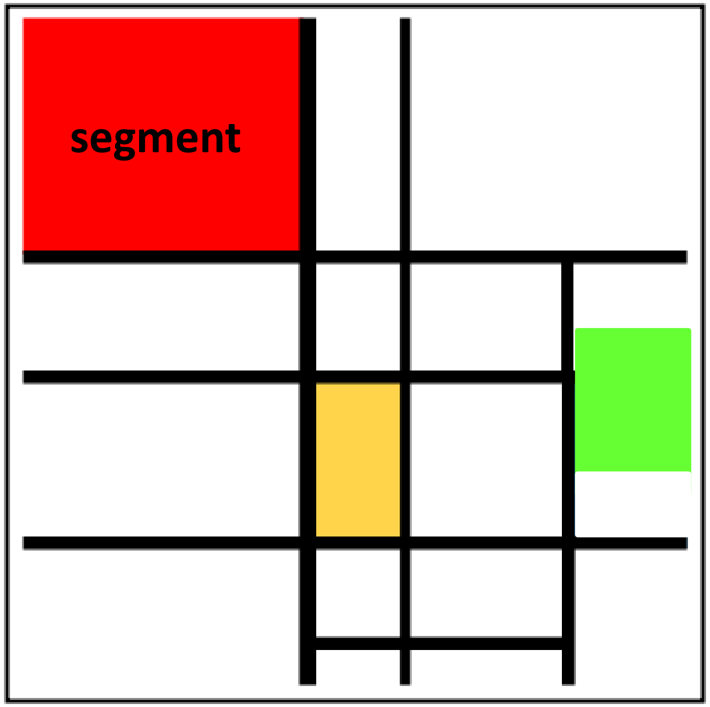
..we do not know which items are consumed together.

Bicluster configuration: market segmentation

vacation activities

costumers

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

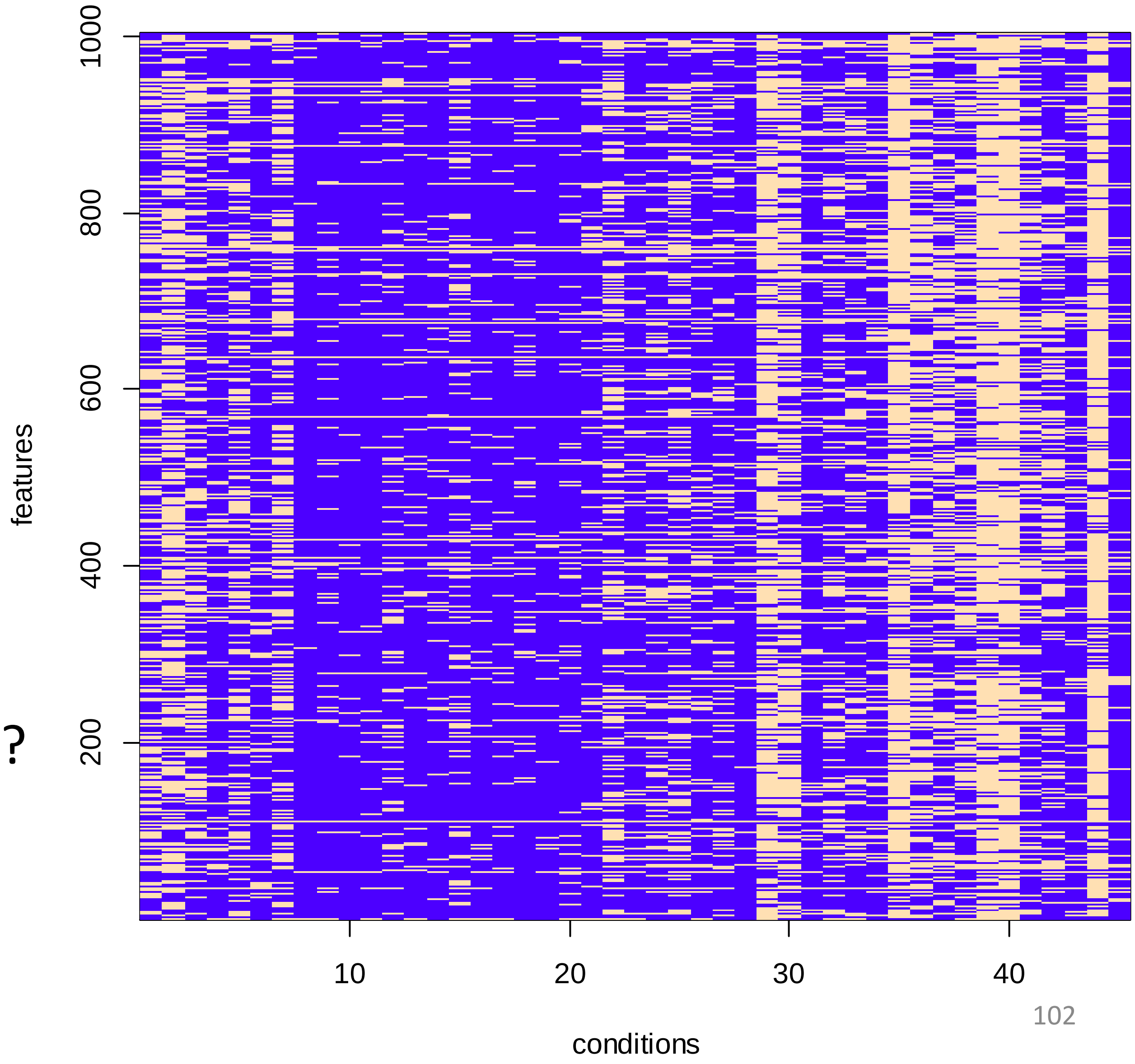


Observed data

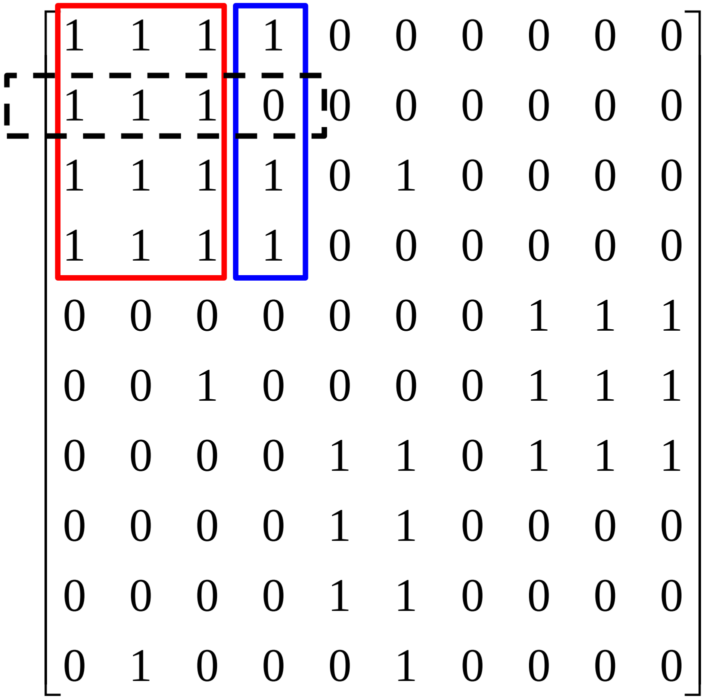
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A 1003X45 binary data.

Observed patterns ?



Data analysis using Bimax

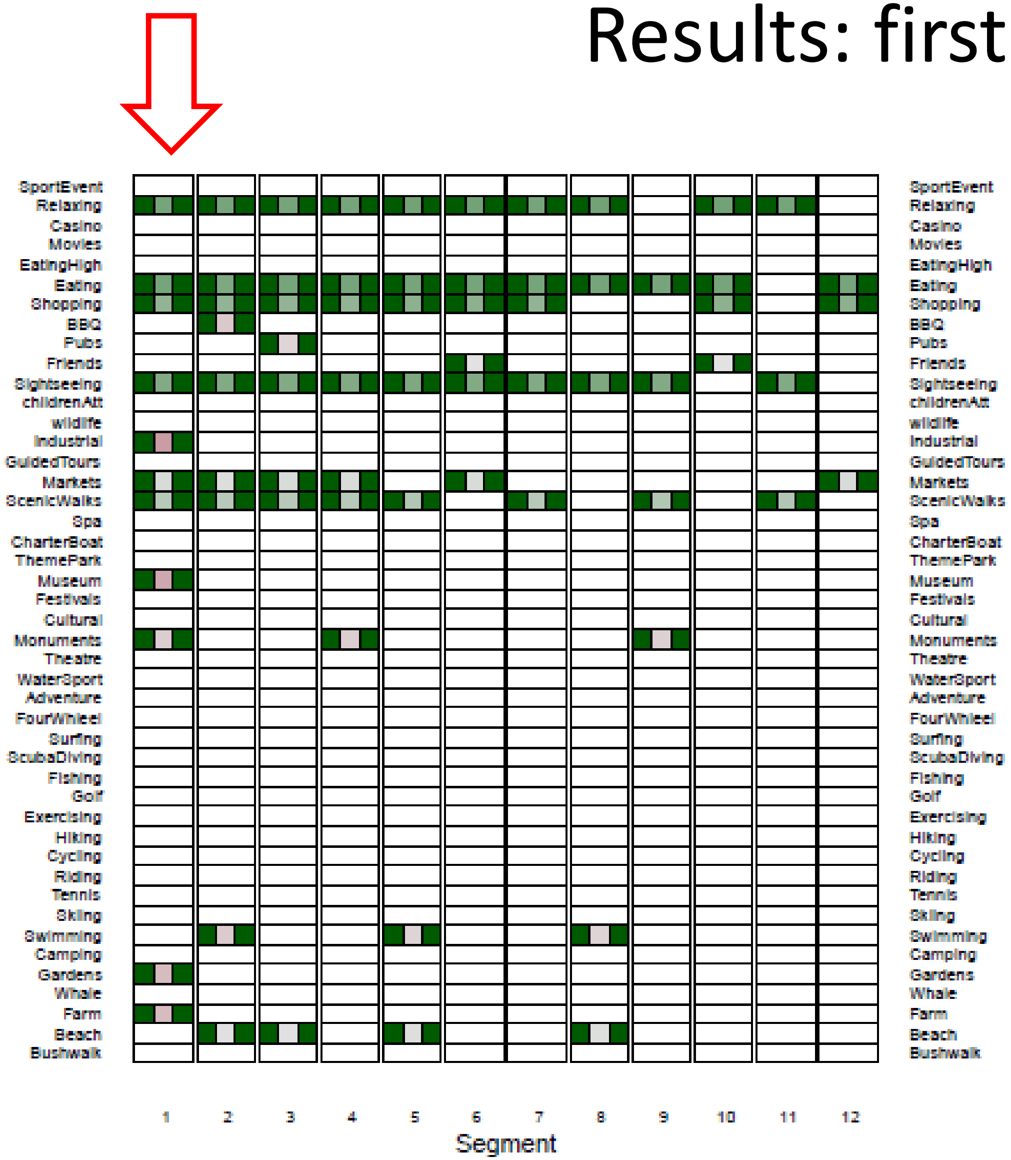


$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Find group od subjects with the same sequence of 1s.

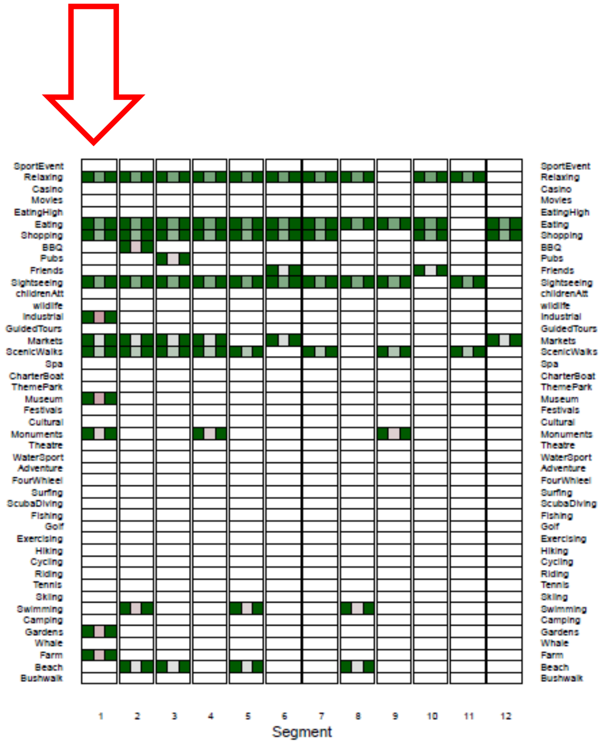
Minimal size of BC ?

Results: first BC



- Number of rows: 74.
- Number of columns: 11
- 7% of the sample:
  1. relaxing
  2. eating in reasonably priced eateries
  3. shopping
  4. sightseeing
  5. visiting industrial attractions (such as wineries, breweries, mines, etc.)
  6. going to markets
  7. scenic walks
  8. visiting museums
  9. visiting monuments
  10. visiting public and botanic gardens
  11. visiting countryside/farms.

Results: first BC

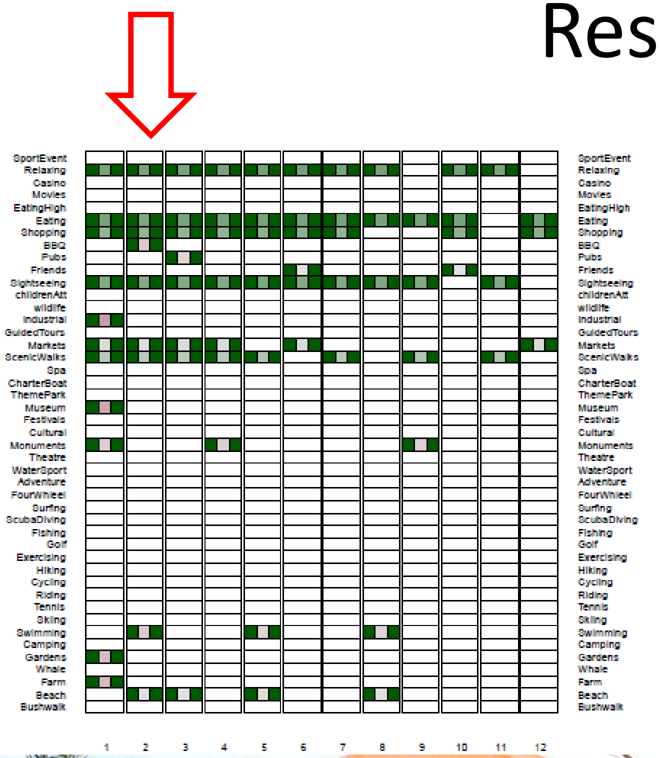


- Number of rows: 74.
- Number of columns: 11
- 7% of the sample:
  1. relaxing
  2. eating in reasonably priced eateries
  3. shopping
  4. sightseeing
  5. visiting industrial attractions (such as wineries, breweries, mines, etc.)
  6. going to markets
  7. scenic walks
  8. visiting museums
  9. visiting monuments
  10. visiting public and botanic gardens
  11. visiting countryside/farms.

Eating

Shopping

Results: second BC



- Number of rows: 87.
- Number of columns: 9
- 8.6% of the sample:
  1. relaxing
  2. eating in reasonably priced eateries
  3. Shopping
  4. BBQ
  5. sightseeing
  6. going to markets
  7. scenic walks
  8. Swimming
  9. Beach

Software

R packages:

- biclust (CRAN)
- biclustGUI (CRAN)

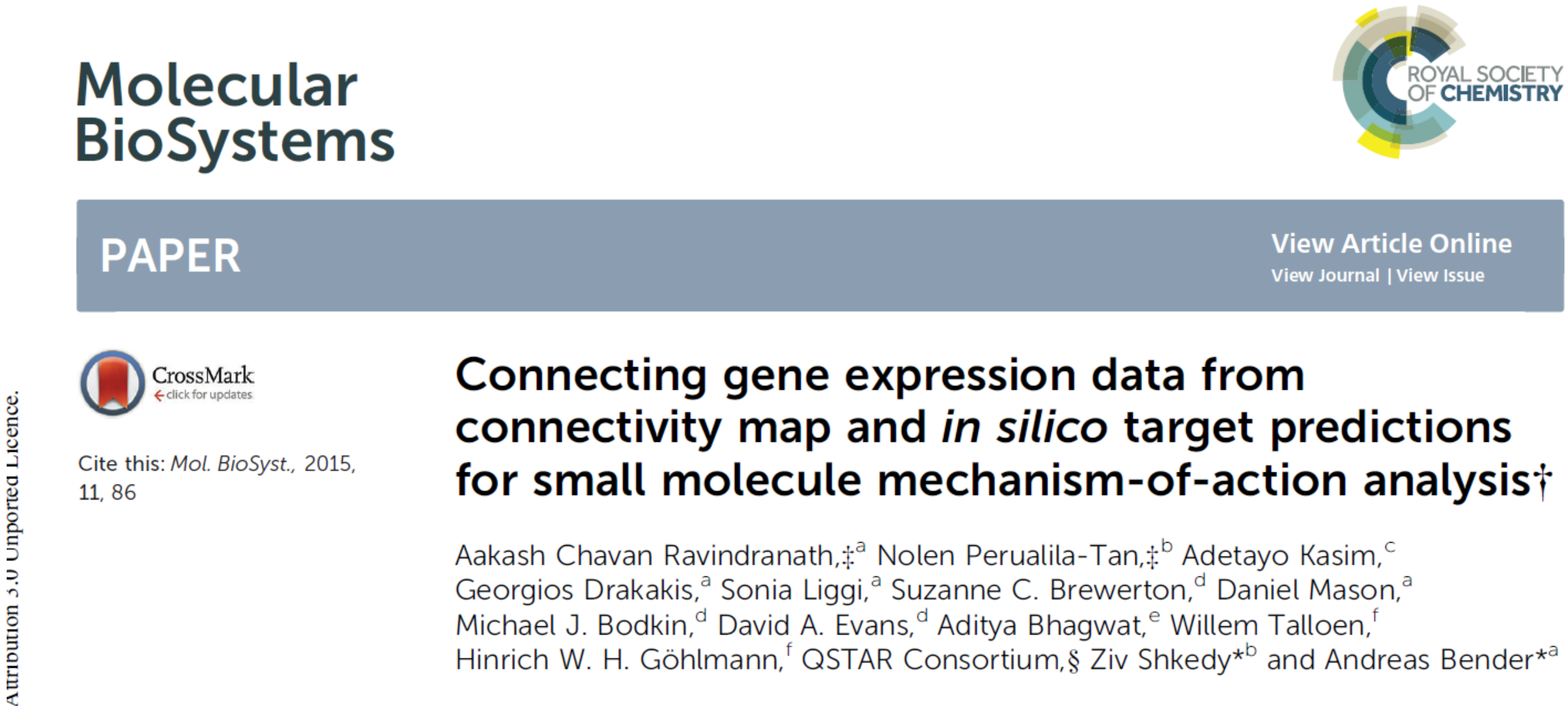
Website:

https://ewouddt.github.io/RcmdrPlugin.BiclustGUI/2016/09/27/biclustGUI/

Part 4.2

Drug Discovery (I):

Biclustering methods for chemoinformatics



Molecular BioSystems

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

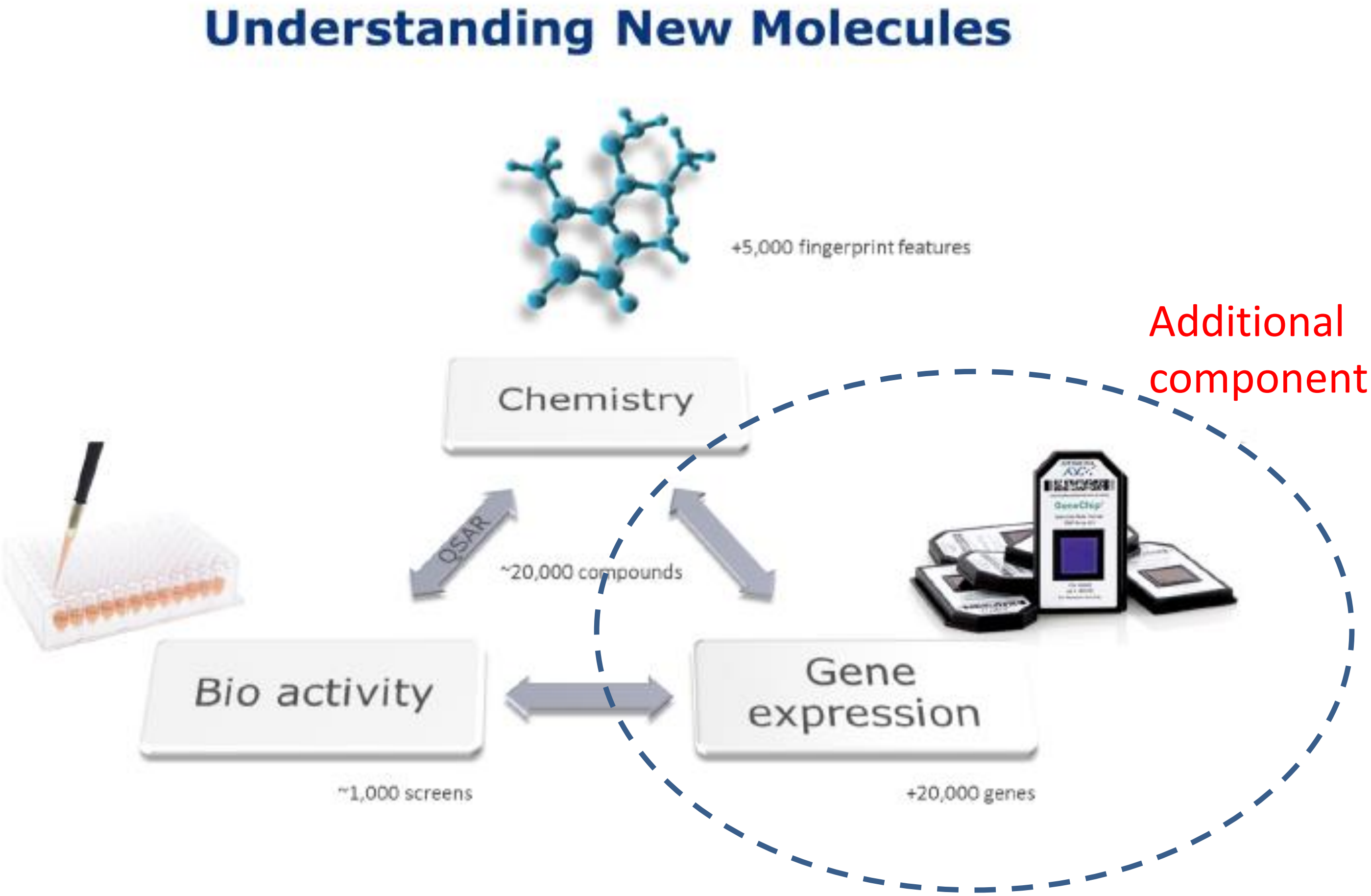
CrossMark click for updates

Cite this: *Mol. BioSyst.*, 2015, 11, 86

Connecting gene expression data from connectivity map and *in silico* target predictions for small molecule mechanism-of-action analysis†

Aakash Chavan Ravindranath,‡[a] Nolen Perualila-Tan,‡[b] Adetayo Kasim,[c] Georgios Drakakis,[a] Sonia Liggi,[a] Suzanne C. Brewerton,[d] Daniel Mason,[a] Michael J. Bodkin,[d] David A. Evans,[d] Aditya Bhagwat,[e] Willem Talloen,[f] Hinrich W. H. Göhlmann,[f] QSTAR Consortium,§ Ziv Shkedy*[b] and Andreas Bender*[a]

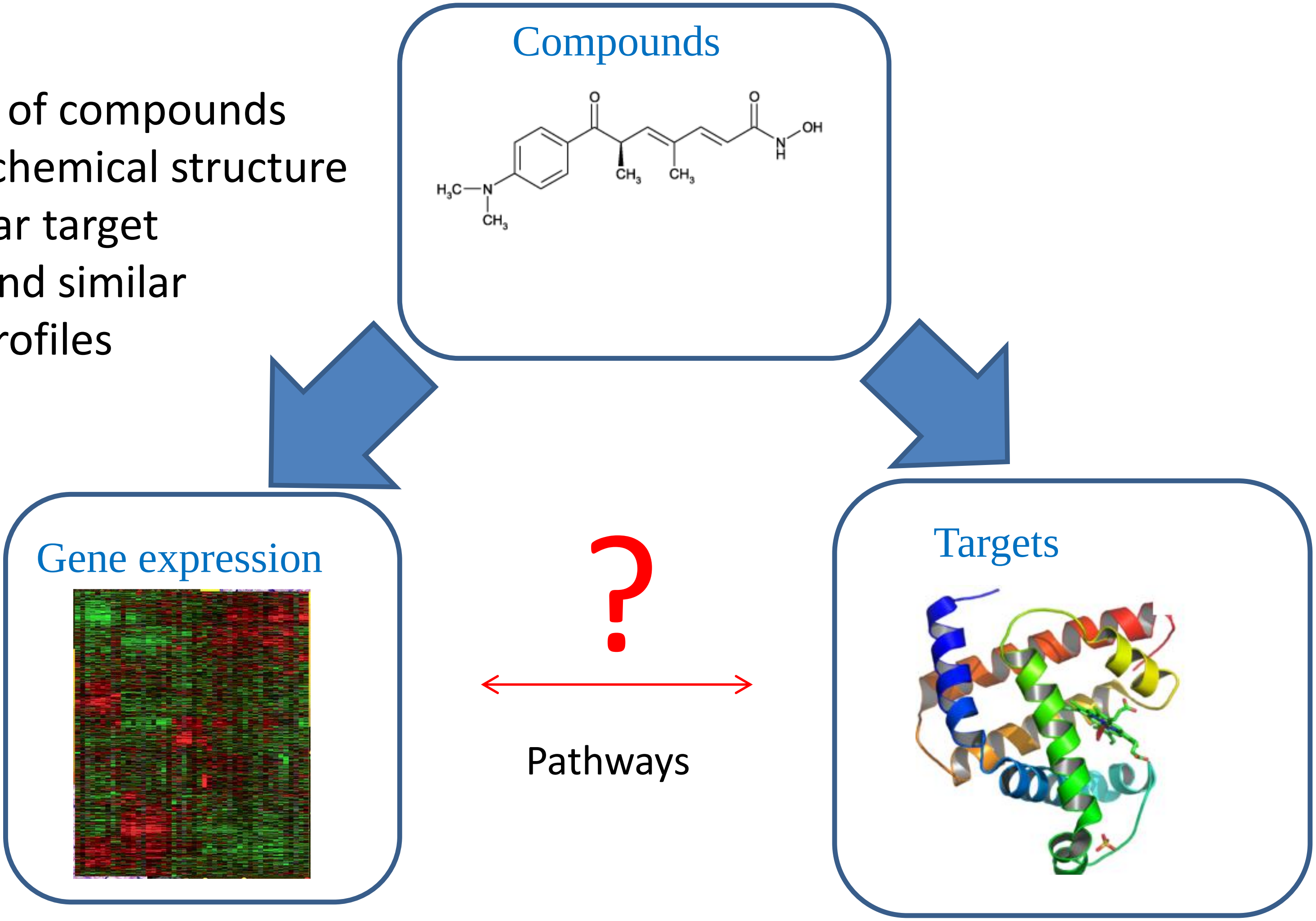
Quantifying S-Transcription-A-R



Additional component

Relating gene expression profiles to Protein targets via compounds

Aim:
find a subset of compounds with similar chemical structure (= have similar target prediction) and similar expression profiles



Compounds

Gene expression

?

Pathways

Targets

Protein Targets and Target prediction

- Many candidate molecules
  - With unknown mechanism of action

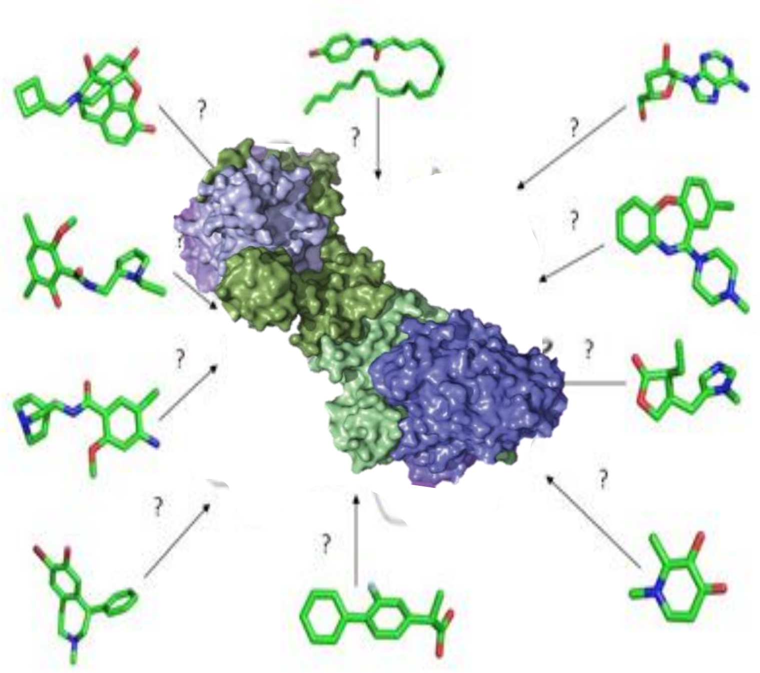
- One drug – many targets
- One target – many active sites (for binding)

- Difficult (expensive) to measure activity of a molecule in all assays

- Predict if drugs will bind to a target given its chemical structure and already known drug-target associations?

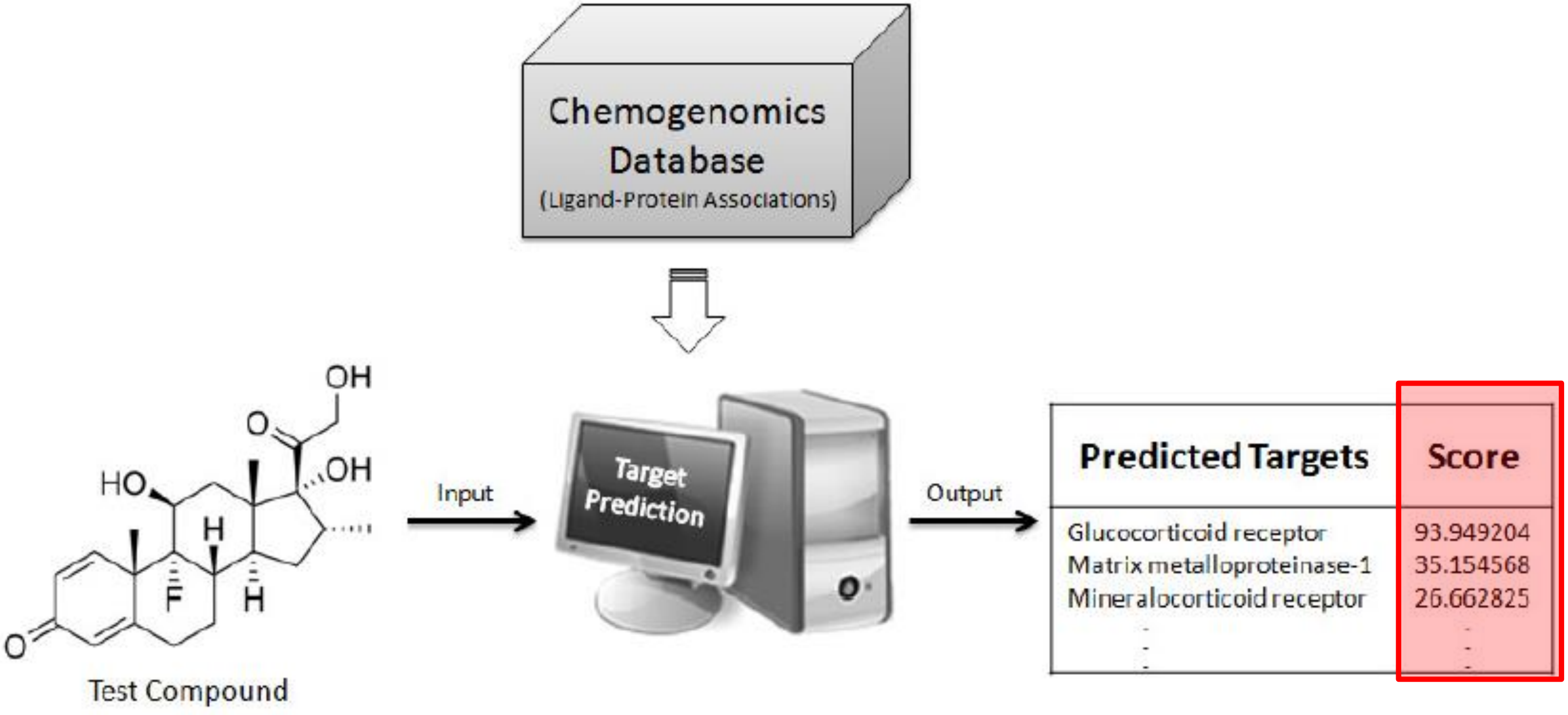
- Target prediction

e.g. Histone deacetylase enzyme



Target Prediction Score

- Likelihood of binding of a compound to every protein target (Koutsoukas,2011)

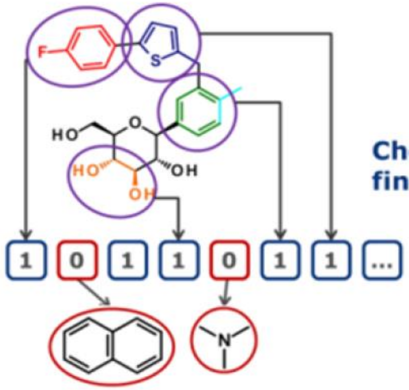


Predicted Targets	Score
Glucocorticoid receptor	93.949204
Matrix metalloproteinase-1	35.154568
Mineralocorticoid receptor	26.662825
-	-
-	-
-	-

Protein Targets and Target prediction

- Predict if drugs will bind to a target given its chemical structure and already known drug-target associations?

- Target prediction



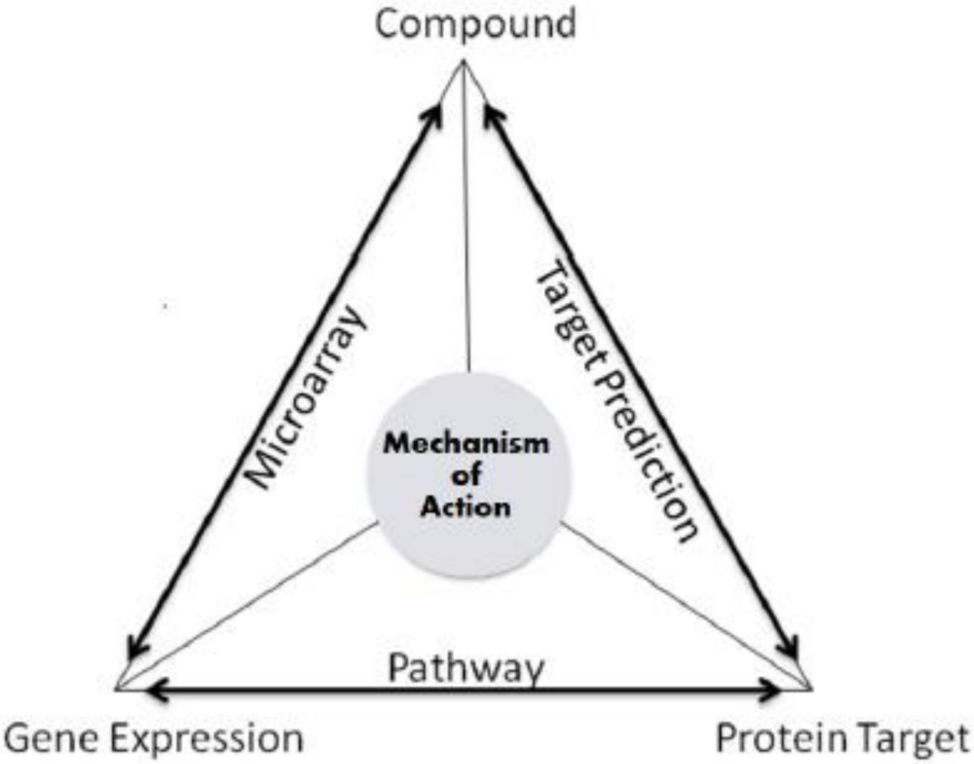
Chemical fingerprint

$$Z_j = \begin{bmatrix} z_{1j} \\ z_{2j} \\ \\ \\ z_{Kj} \end{bmatrix}$$

$$P(TARGET_j) \quad = f\left(FP_1, FP_2, \ldots, FP_K\right)$$

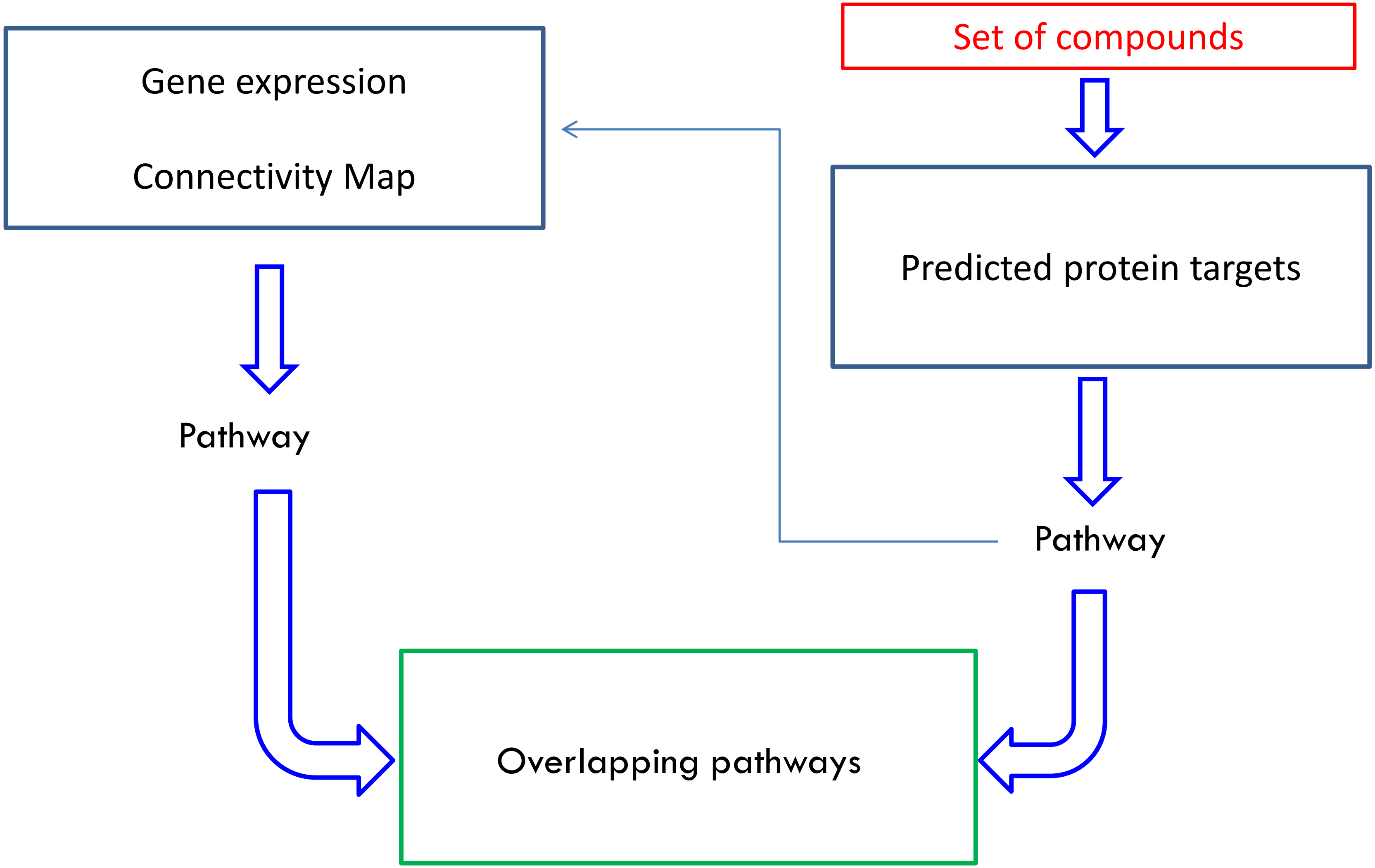
The setting

Mechanism of Action of compound



- Drugs regulating similar protein targets (similar structure) affects similar set of genes

Genes and protein targets pathway overlap



Set of compounds

Gene expression

Connectivity Map

Predicted protein targets

Pathway

Pathway

Overlapping pathways

Gene expression profiles

- Connectivity map data :
  - ❖4 cell lines(MCF7,PC3,HL60 and SKMEL5)
  - ❖After pre-processing ~2400 genes
  - ❖1309 drug like compounds
  - ❖Similar concentration and time of compound exposure

X = Gene Expression

J = 2340 genes

I = 36 compounds

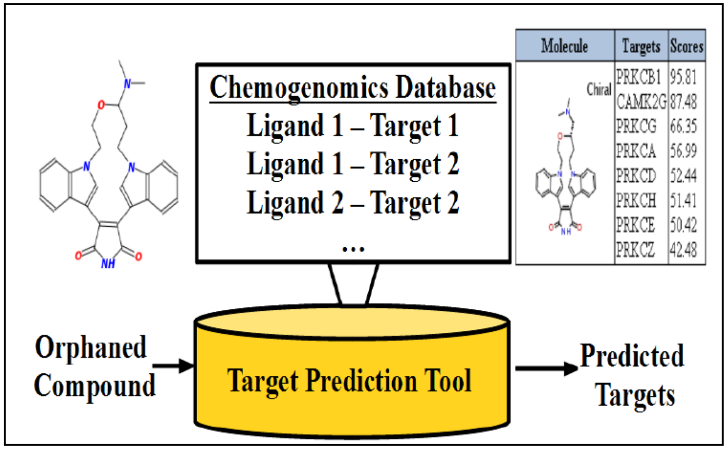
(MCF7 cell line, 6 hours,

10micromolars)

$$X = \underbrace{\begin{bmatrix} x_{11} & x_{21} & . & . & x_{n1} \\ x_{12} & x_{22} & . & . & x_{n2} \\ . & . & . & . & \\ . & . & . & . & \\ x_{1m} & x_{2m} & . & . & x_{nm} \end{bmatrix}}_{\text{compounds}} \left.\vphantom{\begin{matrix} x \\ x \\ . \\ . \\ x \end{matrix}}\right\}\text{genes}$$

Target prediction: binary scores

- Biosar (Naïve Bayes) used to predict targets
- Individual cut off used for each target



Molecule	Targets	Scores
Chiral	PRKCB1	95.81
	CAMK2G	87.48
	PRKCG	66.35
	PRKCA	56.99
	PRKCD	52.44
	PRKCH	51.41
	PRKCE	50.42
	PRKCZ	42.48

Target scores matrix

$$T = \left.\underbrace{\begin{bmatrix} t_{11} & t_{21} & . & . & t_{n1} \\ t_{12} & t_{22} & . & . & t_{n2} \\ . & . & . & . & \\ . & . & . & . & \\ t_{1m} & t_{2m} & . & . & t_{nm} \end{bmatrix}}_{\text{compounds}}\right\}\text{targets}$$

$$t_{ip} = \begin{cases} 1 & \text{Comp } i \text{ hit on target } p \\ 0 & \end{cases}$$

$$T_{C_i} = (0,1,1,0,0,0,.....1,0)$$

Pathways

- Specific group of compounds
- A group of genes and targets that share:

  - ➢ a biological pathway
  - ➢ a statistical pathway

Biological pathways

- KEGG (Koyoto Encyclopedia of Genes and Genomes)- is a freely available information repository of the network of genes and molecules for practical analysis of the gene functions
- GO (Gene Ontology)- is a bioinformatics project that is the largest repository for catalogue gene function that unifies the representation of gene and gene product attribute across all the species.
- KEGG and Go pathways were annotated to proteins and genes.

Part II
Data analysis

Data analysis steps

- Target based clustering.

  ⇨ similarity matrix based on target prediction scores.

- Gene expression profiling.
- Enrichment of the gene set.
- Pathway identification.

Correlation between compounds

Target scores matrix

$$T = \left.\underbrace{\begin{bmatrix} t_{11} & t_{21} & . & . & t_{n1} \\ t_{12} & t_{22} & . & . & t_{n2} \\ . & . & . & . & \\ . & . & . & . & \\ t_{1m} & t_{2m} & . & . & t_{nm} \end{bmatrix}}_{\text{compounds}}\right\}\text{targets}$$

- N_{c1} and N_{c2} are the number of fingerprint features present in compound 1and compound 2.
- N_{c12} is the number of features common to both compounds.

Tanimoto scores

$$TC = \frac{N_{C12}}{N_{C1} + N_{C2} - N_{C12}}$$

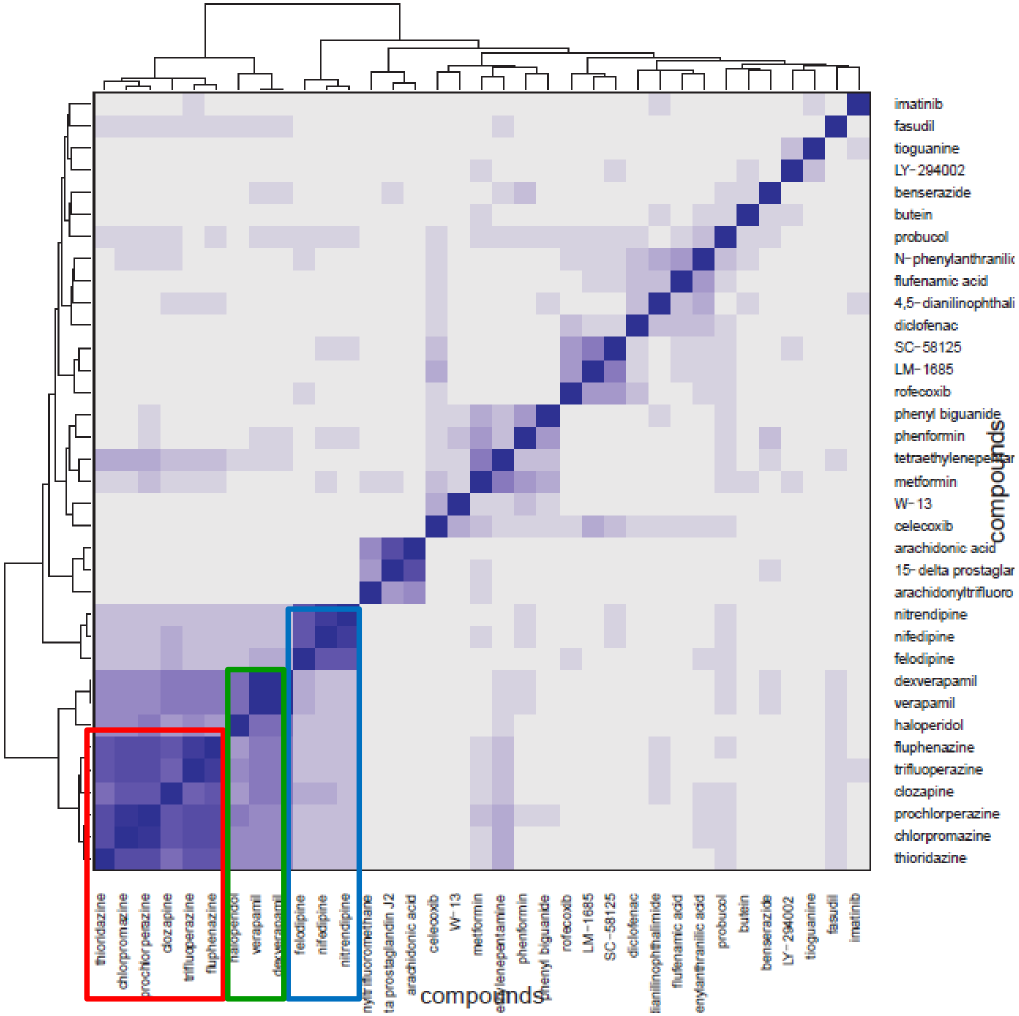
$$TC = 1$$

2 compounds are identical given the set of chemical structures

$$TC = 0$$

2 compounds do not share any chemical structure

Target similarity matrix



Similarity matrix based on Tanimoto scores

$$TC = \frac{N_{C12}}{N_{C1} + N_{C2} - N_{C12}}$$

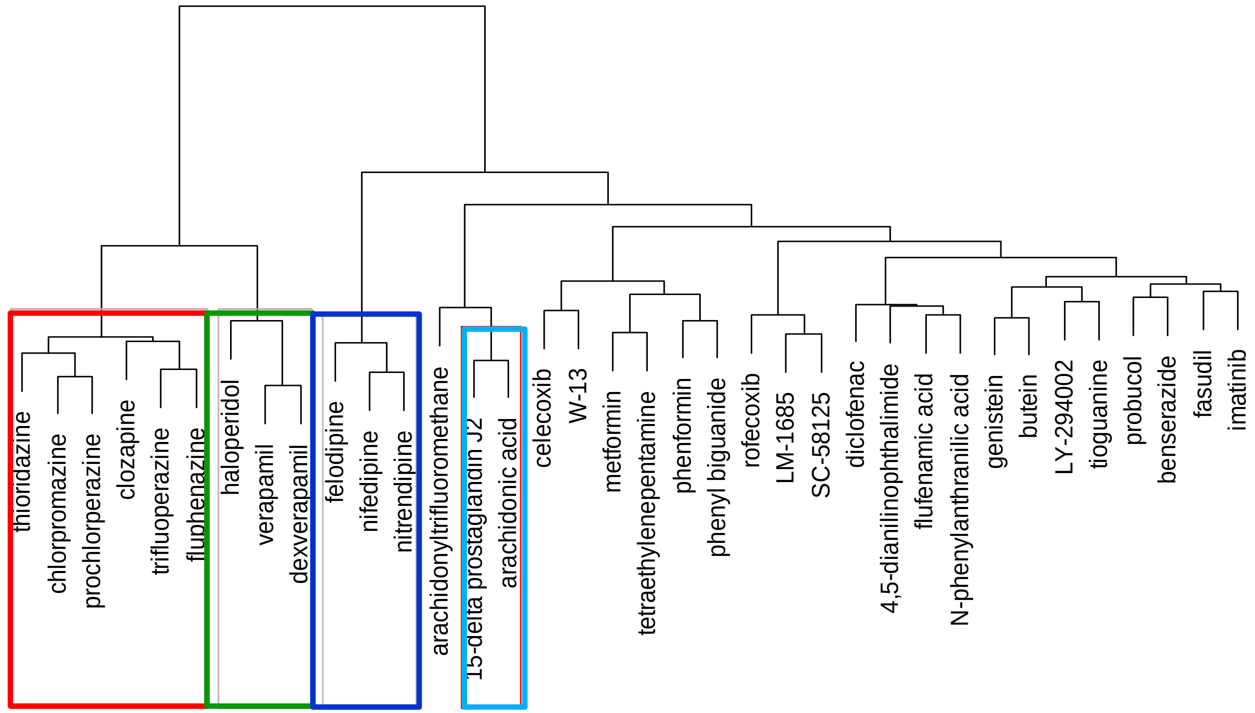
Cluster compound based on similarity scores

Hierarchical clustering

- Input: Similarity matrix
- Start: each compound is a cluster
- Merge compounds according to a criterion
- Ward’s distance
- End: single cluster of all compounds

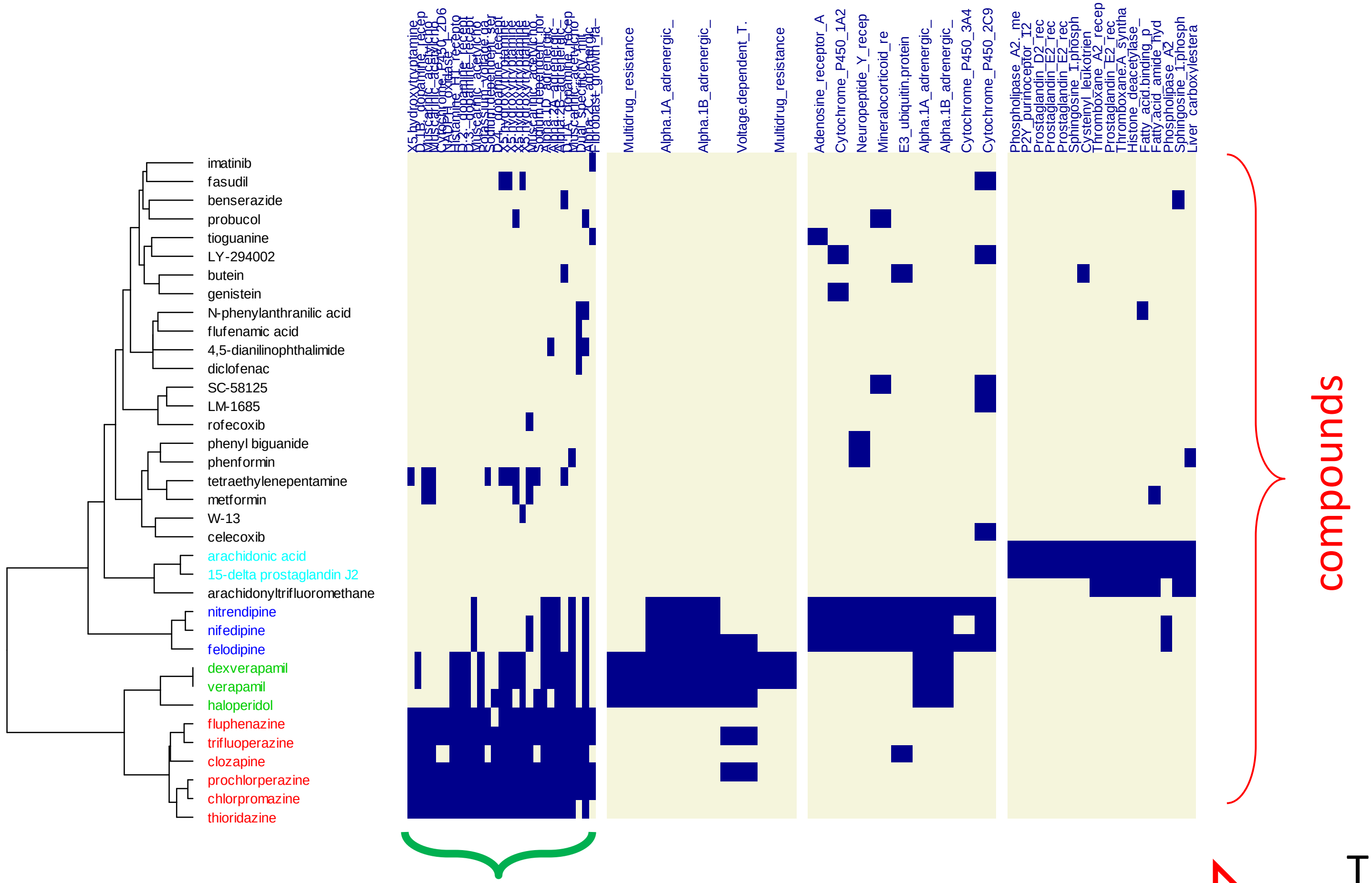
Target prediction based clustering

For each cluster, identify target scores in common for all compounds in the cluster.



Target prediction based clustering

Identification of target prediction scores which are in common for the compounds in a cluster.

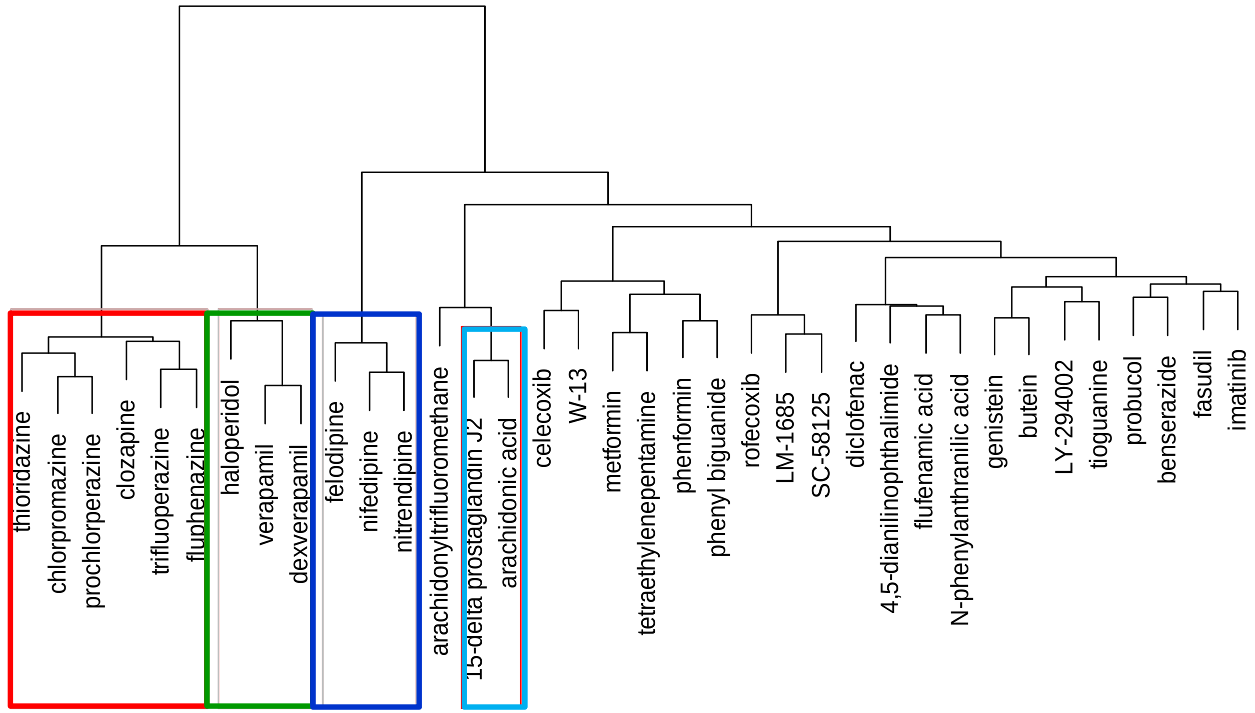


Subset of target prediction scores related to cluster 1

The targets in the pathway

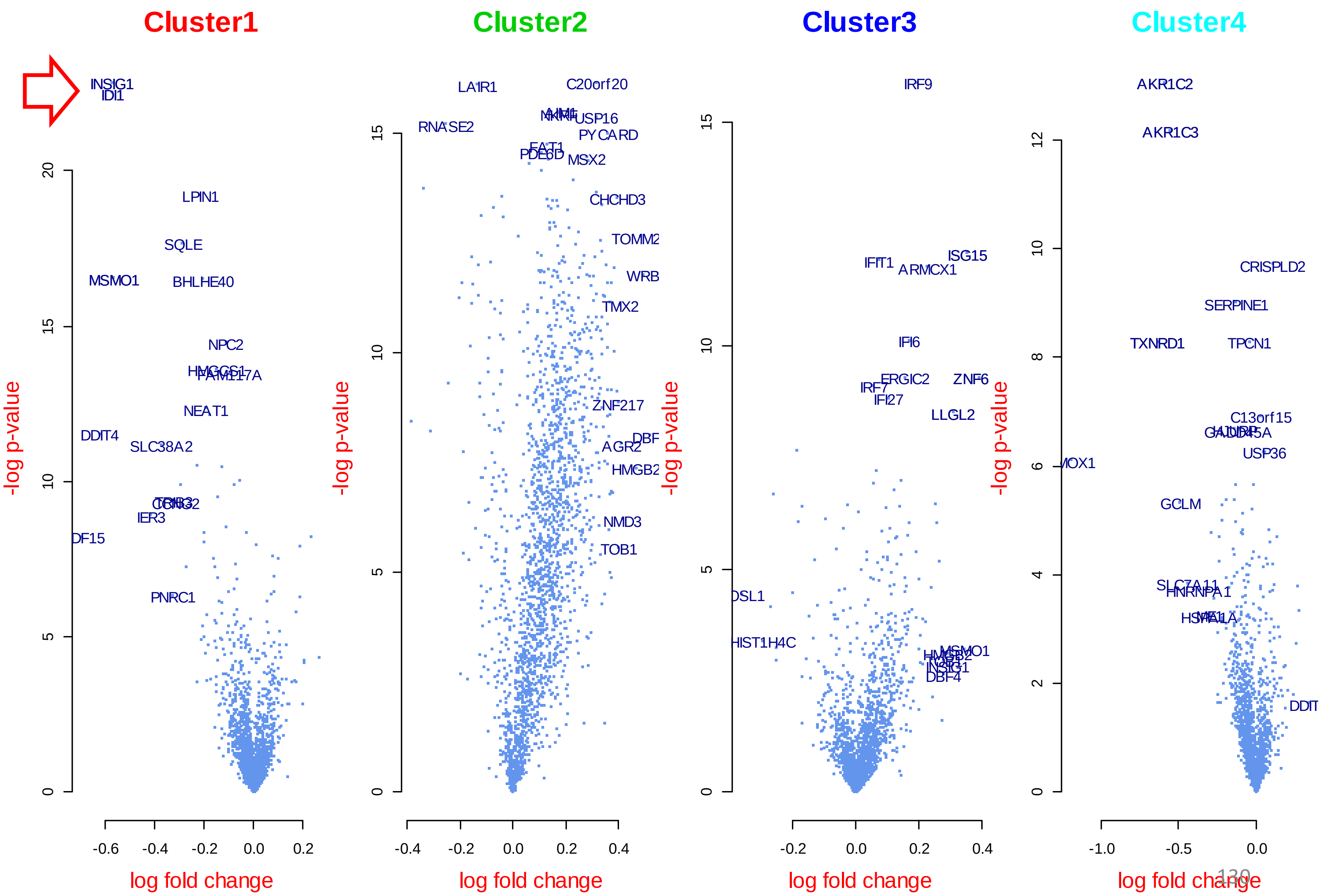
Target-based clustering

Identify genes which have different expression profile between a cluster of interest and the rest of the compounds.

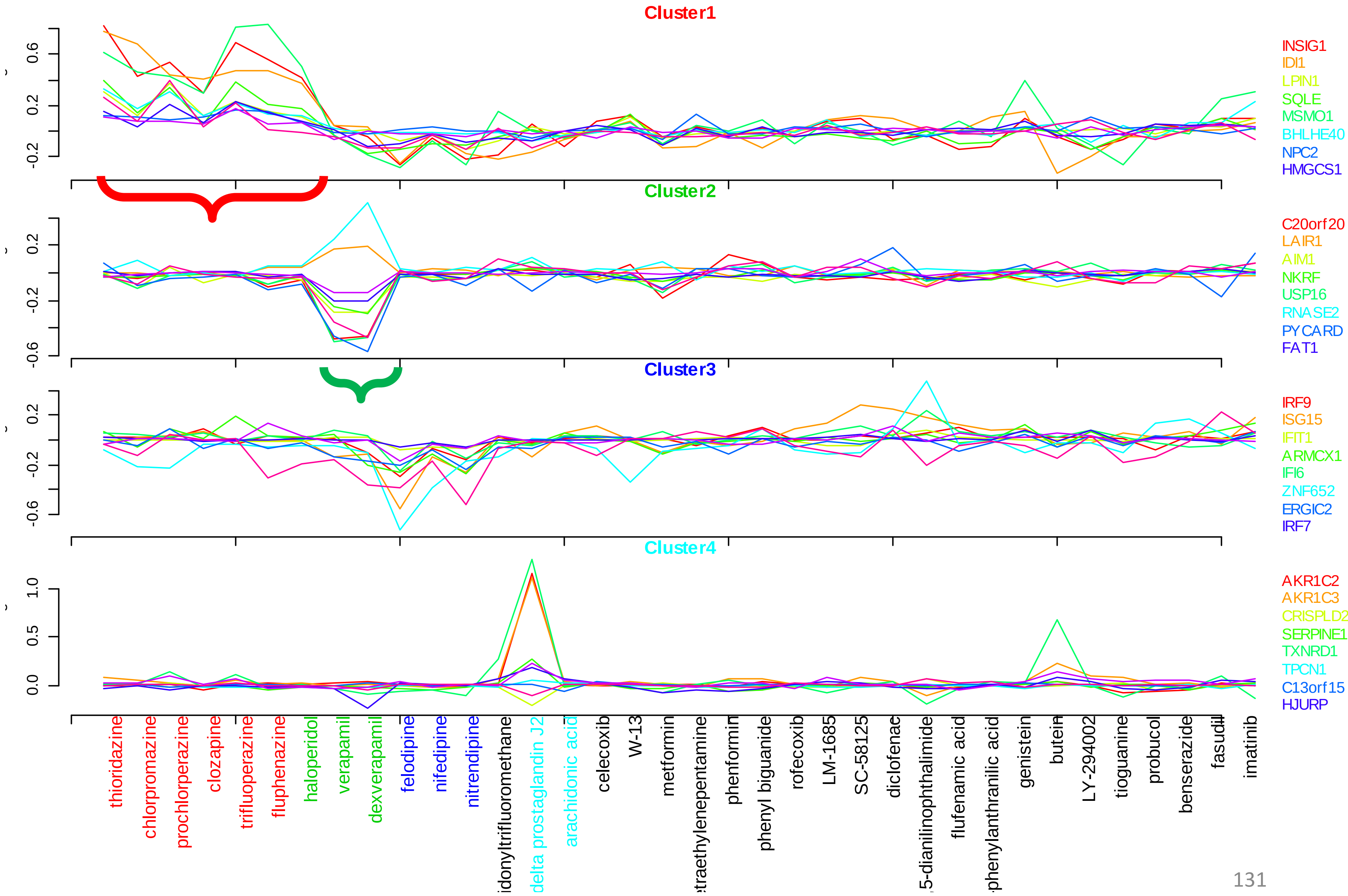


Which genes are related to this specific cluster ?

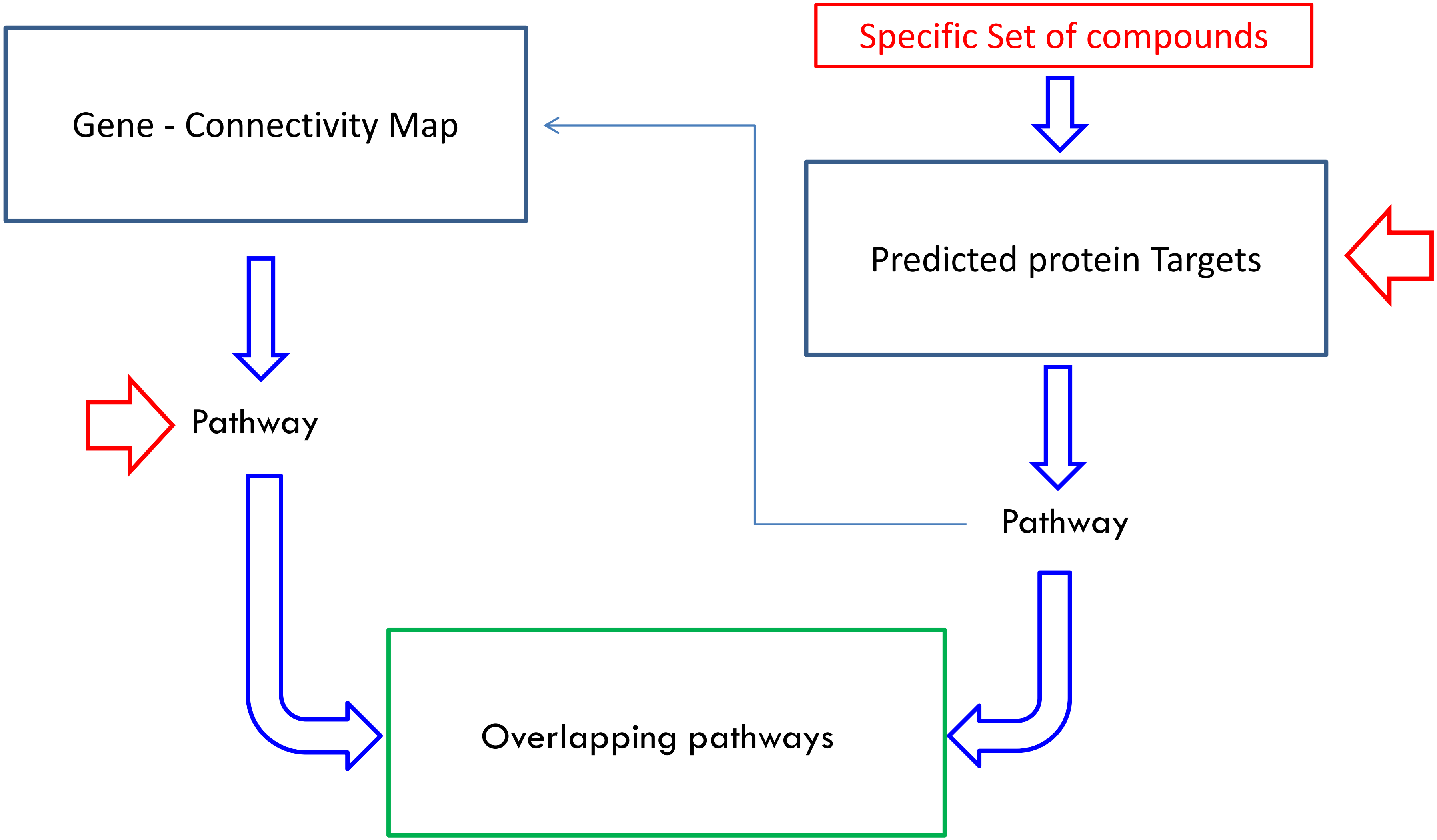
Differentially expressed genes



Profiles plots for top 8 genes by cluster



Genes and protein targets pathway overlap



Specific Set of compounds

Gene - Connectivity Map

Predicted protein Targets

Pathway

Pathway

Overlapping pathways

Biological pathways: cluster 1

Use:

top K genes.

Target scores which are in common among compounds in cluster 1 (search in KEGG , GO)

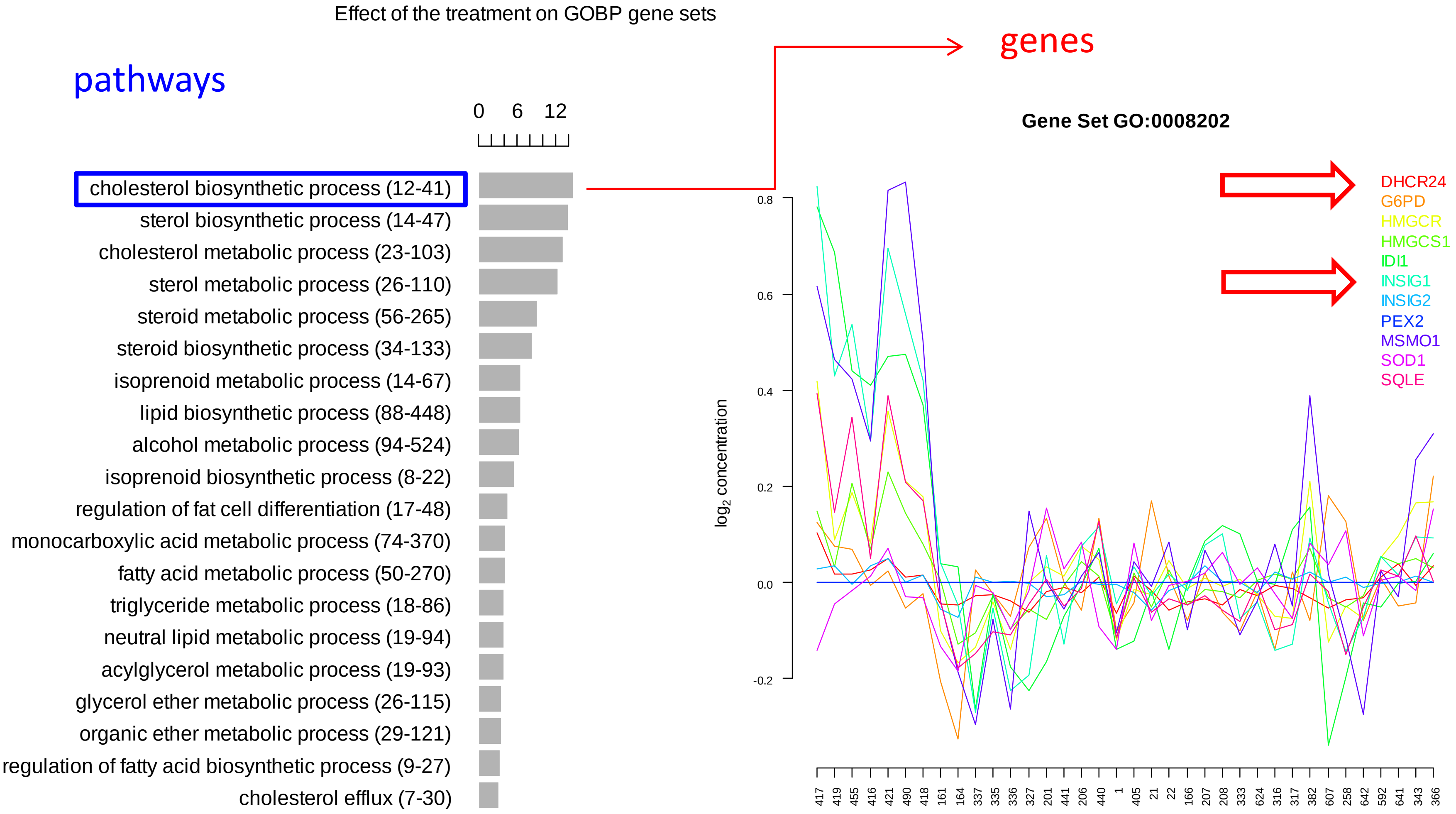
Identify :

biological pathways

Gene set analysis with MLP was done as well (to discover more genes).

Compound Clusters	Compound Names	Targets	Genes	Pathways
antipsychotic	"clozapine" "thioridazine" "chlorpromazine" "trifluoperazine" "prochlorperazine" "fluphenazine"	CytochromeP 4502D6	INSIG1; LDLR	GO:0008202; P:steroid metabolic process; IMP:BHF-UCL
		Dual specificity mitogen: activated protein kinase kinase1	DUSP4	hsa04010: MAPKsignalingpathway
		Fibroblast growth f actor receptor1		
		Dual specificity mitogen: activated protein kinase kinase1	LAMA3	hsa04510: Focal adhesion
		Dual specificity mitogen: activated protein kinase kinase1	TUBA1A	hsa04540: Gapjunction

MLP: cluster 1

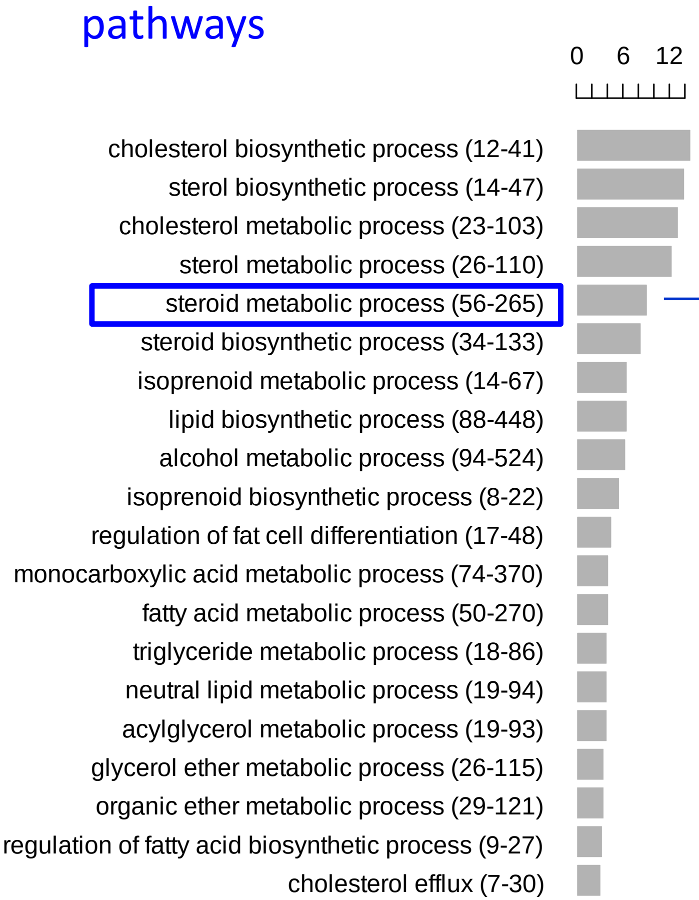


pathways

genes

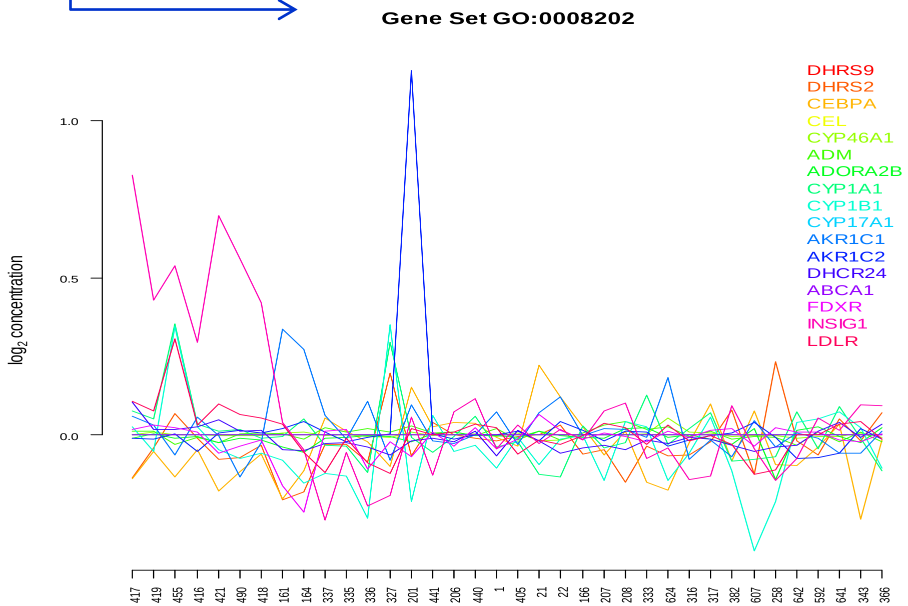
MLP: cluster 1

Effect of the treatment on GOBP gene sets

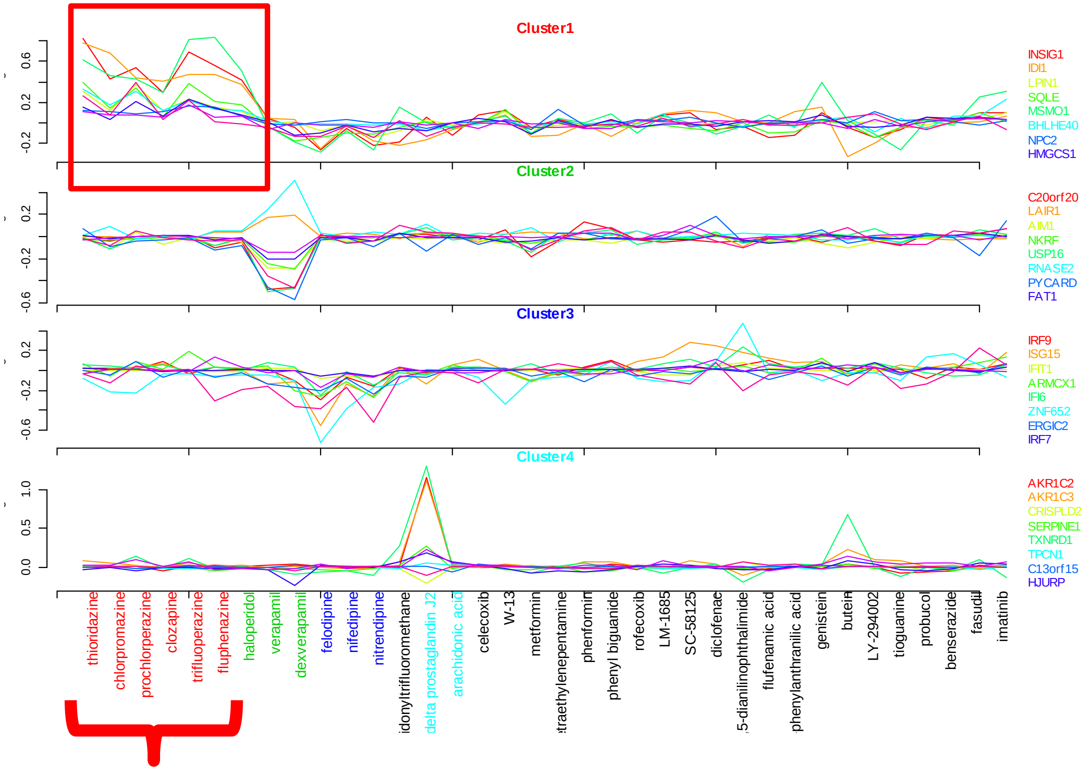


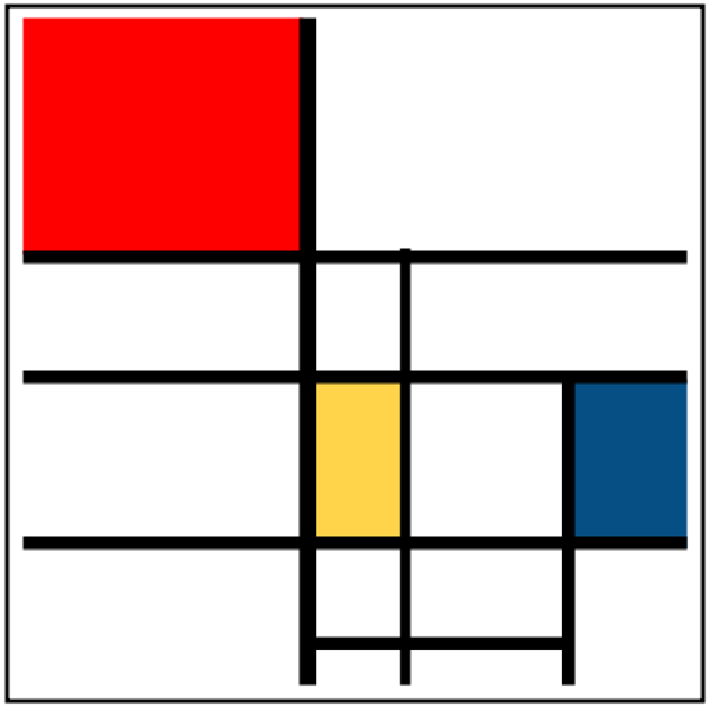
The genes: INSIG1 & LDLR are related to this pattern.

More genes ?

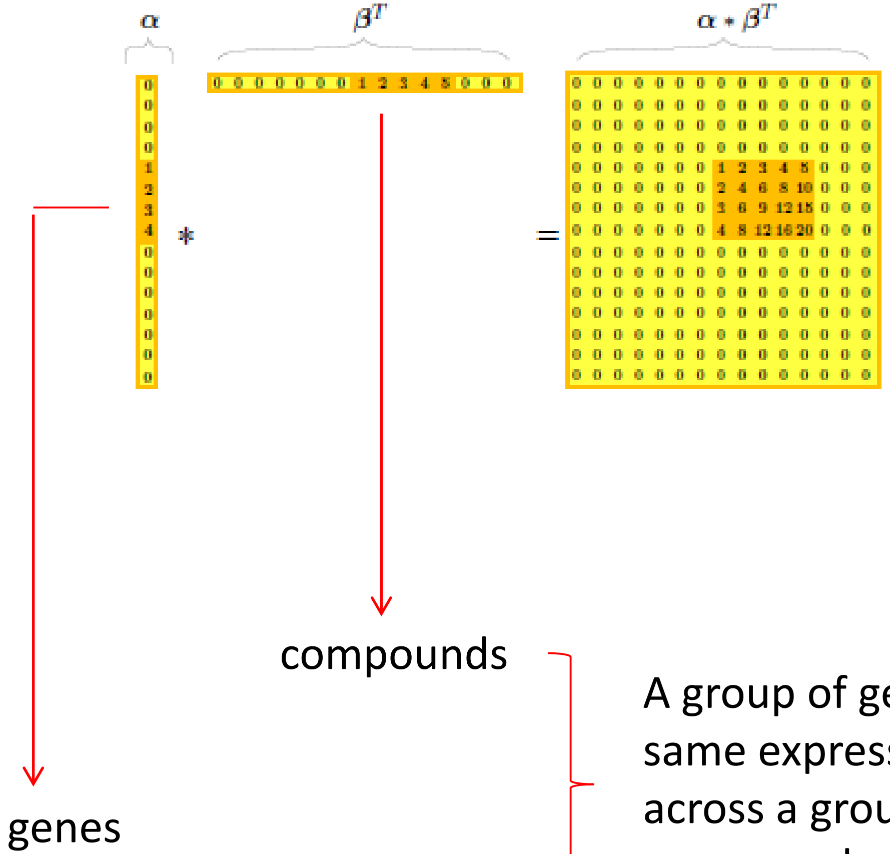


Why this is a bicluster ?





Applying FABIA: data structure + model



$$\underbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}_{\alpha} * \underbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}_{\beta^T} = \underbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}_{\alpha * \beta^T}$$

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{G1} & X_{G2} & \ldots & X_{Gn} \end{pmatrix}.$$

$$\mathbf{X} = \sum_{i=1}^{p} \lambda_i {\gamma_i}^T + \Upsilon,$$

A group of genes with the same expression profiles across a group of compounds.

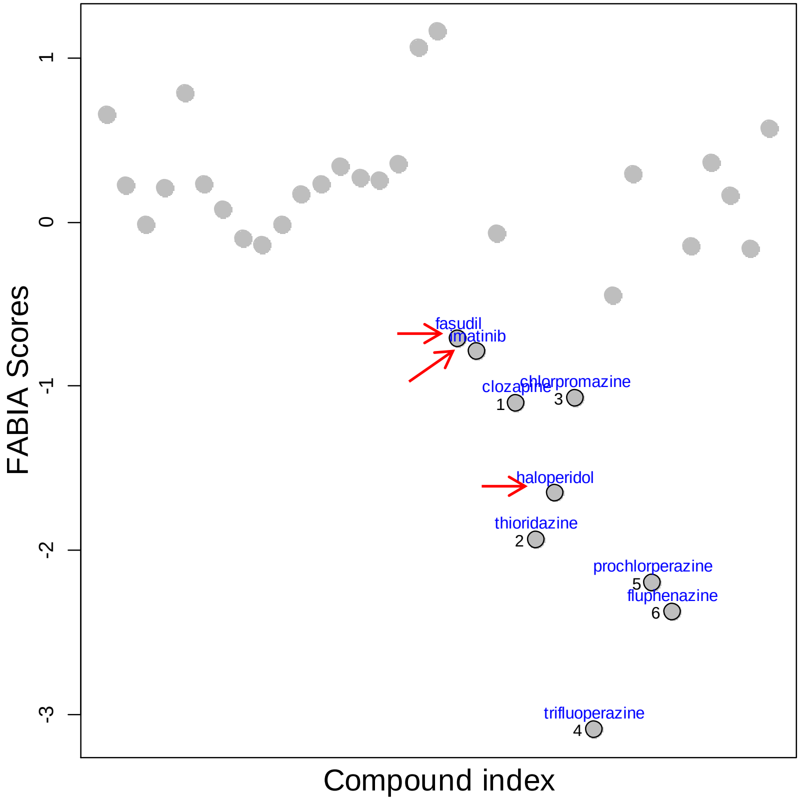
Software: biclustering using FABIA

$$\text{gMat} = \boxed{\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1B} \\ X_{21} & X_{22} & \dots & X_{2B} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{B1} & X_{B2} & \dots & X_{GB} \end{pmatrix}}$$

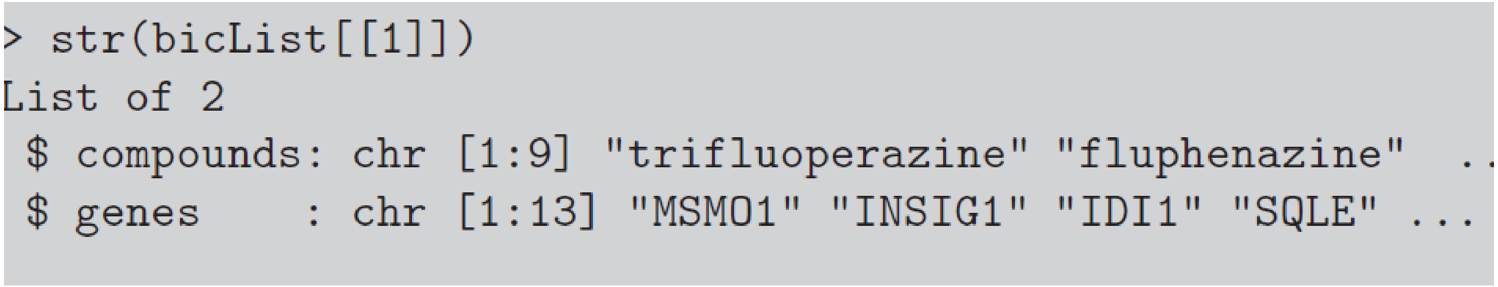
G

```
>fabRes <- fabia(gMat, alpha=0.1, p=20, cyc=1000, spl=1,
                 spz=0.5)
>rb <- extractBic(fabRes)
```

Results: biclustering using FABIA- compound scores

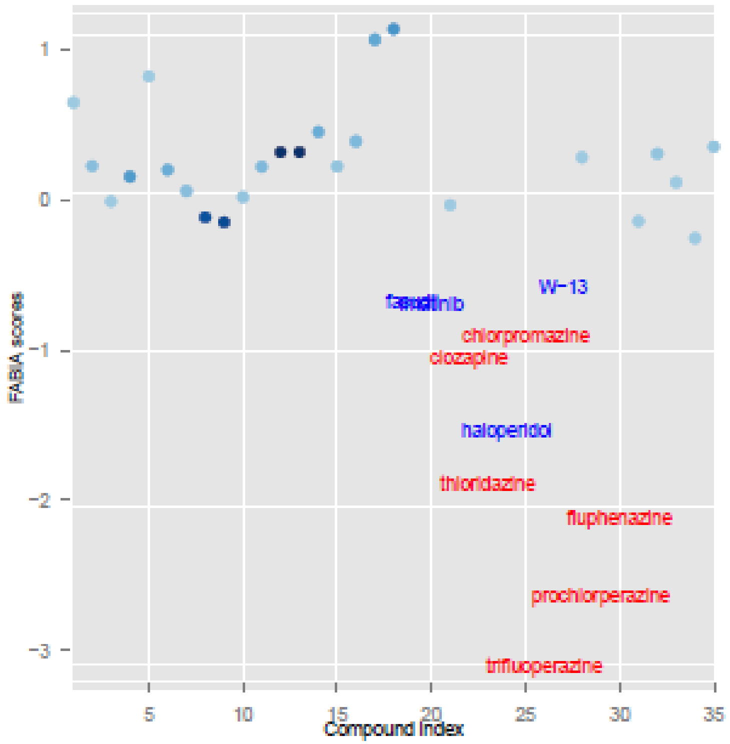


- Bicluster 1 is similar to cluster 1 compound set
- With 3 extra compounds

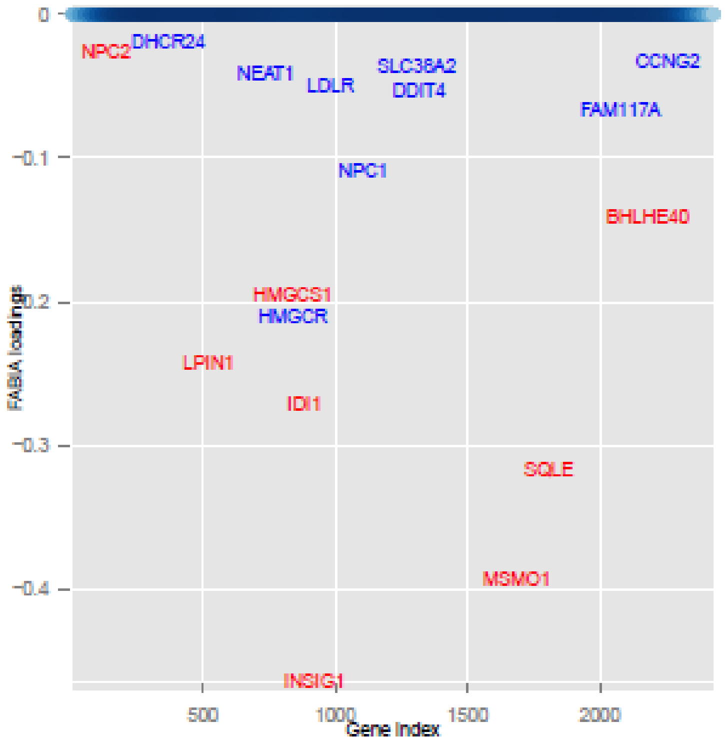


```
> str(bicList[[1]])
List of 2
 $ compounds: chr [1:9] "trifluoperazine" "fluphenazine"  ..
 $ genes    : chr [1:13] "MSMO1" "INSIG1" "IDI1" "SQLE" ...
```

Factor scores (compounds) and factor loadings (genes)

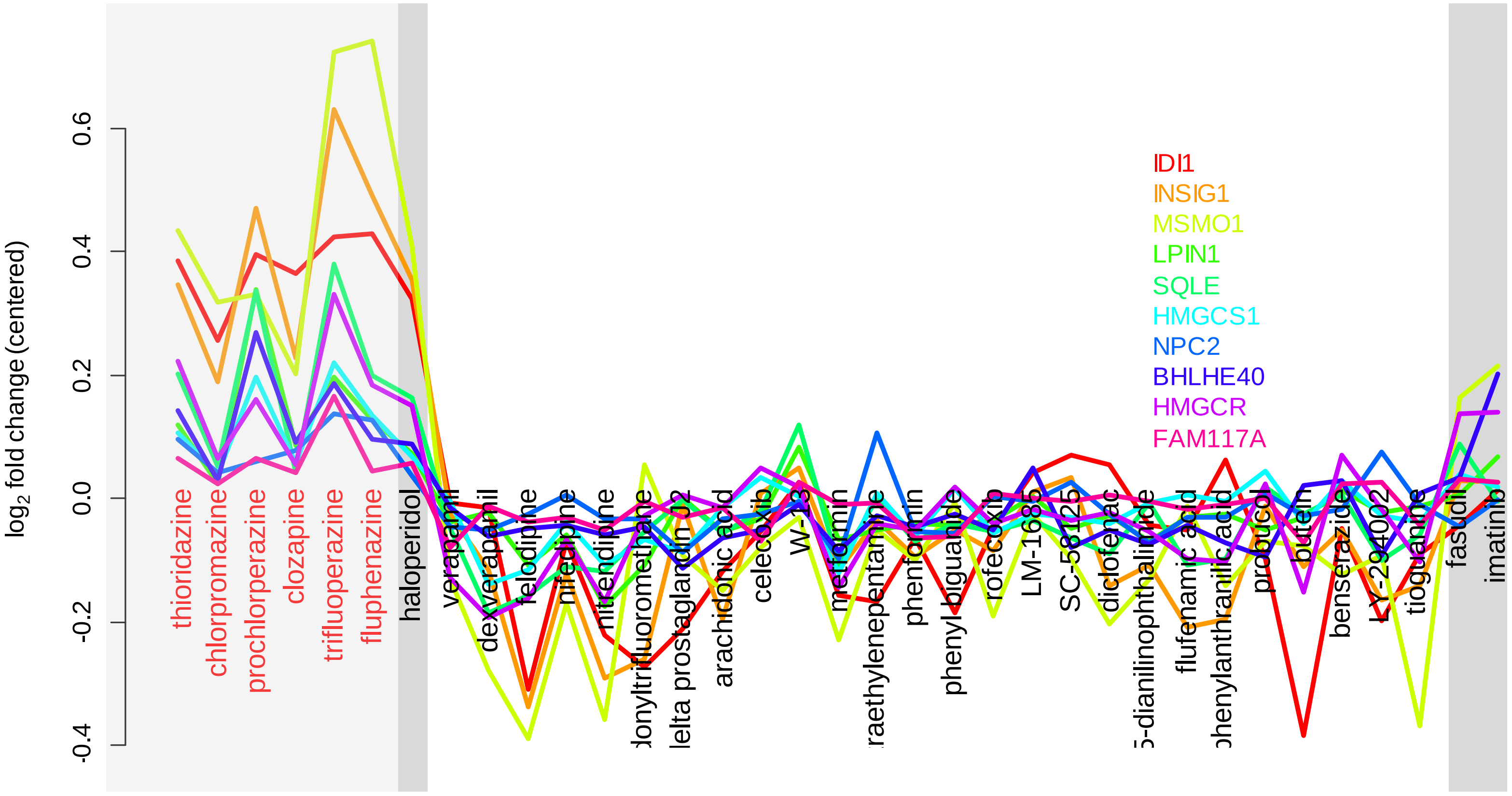


(a) FABIA BC1: Compound scores.



(b) FABIA BC1: Gene Loadings.

A bicluster (FABIA)



Discussion

- An exploratory tool for discovering subgroups with aligned multiple properties
- Could be applicable in other research fields

- One of the integrative clustering approaches included in the package *IntClust.*

Part 4.3

Sport:

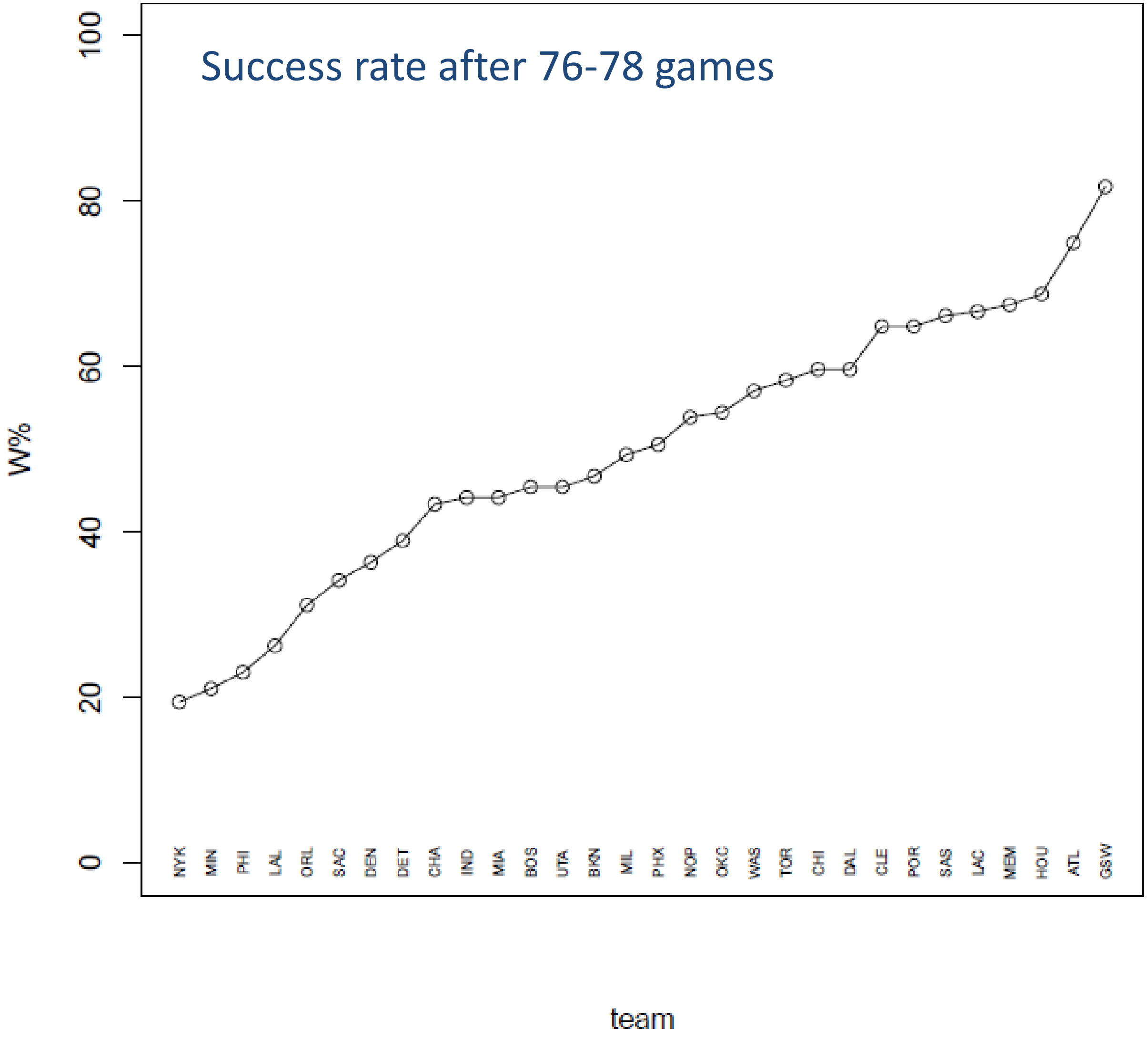
Using biclustering method to detection of local patterns in NBA data

NBA

- 30 teams
- Regular season : 82 games per team
- 16 teams go to the Play-offs
- Performance Statistics (teams and individuals) is well developed

Our aim: to develop a multivariate performance indicator

Success rate of NBA teams in the regular season of 2014/2015


Success rate after 76-78 games

$$\%W = \frac{\#\, games \quad won}{\#\, games}$$

Commonly used performance indicators in NBA

- 2-pt / 3-pt Successful
- 2-pt / 3-pt Unsuccessful
- Free Throw Successful / Unsuccessful
- Defensive / Offensive Rebounds
- Assists
- Turnovers
- Steals
- Dunks
- Blocks Committed / Received
- Fouls Committed / Received

Garcia et al (2013) showed that these variables are good performance indicators for Regular Season as well as for Playoff Games

Data Structure

- 7 online databases in the NBA website:
  - Traditional Stats
  - Advanced Stats
  - Four Factors
  - Misc. Stats
  - Scoring
  - Opponent
  - Shooting

Updated after each game

Data structure

$$X = \left[\underbrace{X_1, X_2, X_3, X_4, X_5, X_6, X_7}\right]$$

1. Advanced Stat
2. Four Factors
3. Misc. Stats
4. Scoring
5. Opponent
6. Shooting

Each matrix is a *30 X* n_i

Example :

Traditional Stats

- 2-pt / 3-pt Successful
- 2-pt / 3-pt Unsuccessful
- Free Throw Successful / Unsuccessful
- Defensive / Offensive Rebounds
- Assists
- Turnovers
- Steals
- Dunks
- Blocks Committed / Received
- Fouls Committed / Received

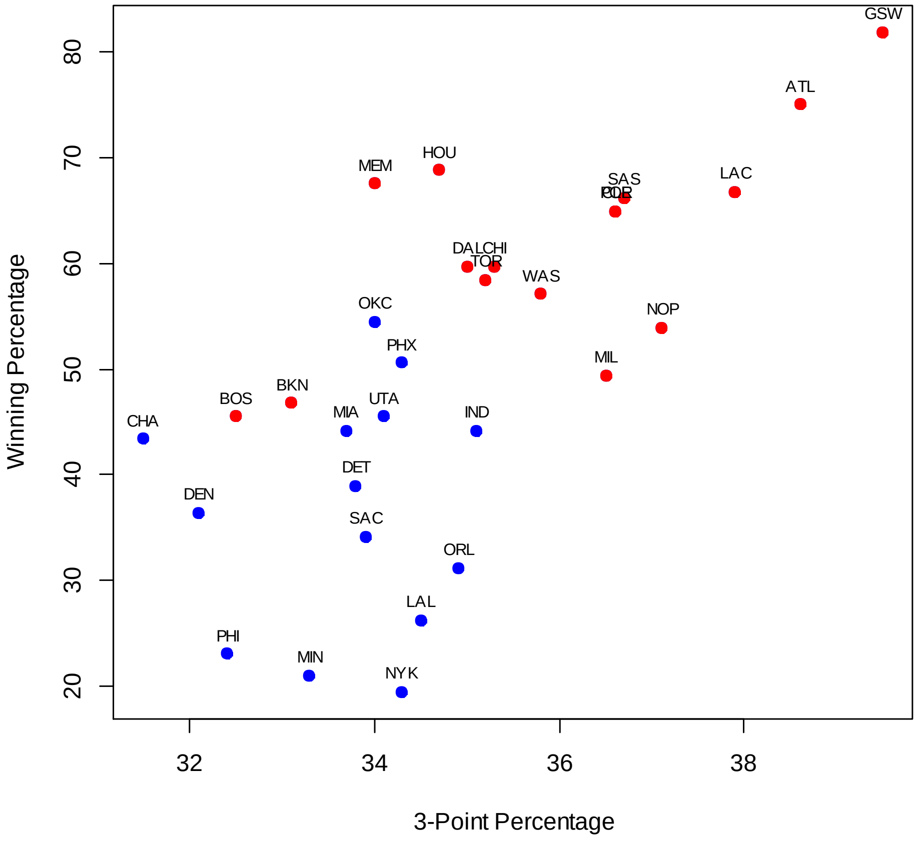
Analysis plan

- Step 1: PCA for the Traditional Stats:
  - 2-pt / 3-pt Successful
  - 2-pt / 3-pt Unsuccessful
  - Free Throw Successful / Unsuccessful
  - Defensive / Offensive Rebounds
  - Assists
  - Turnovers
  - Steals
  - Dunks
  - Blocks Committed / Received
  - Fouls Committed / Received
- Step 2: Multiple factor analysis

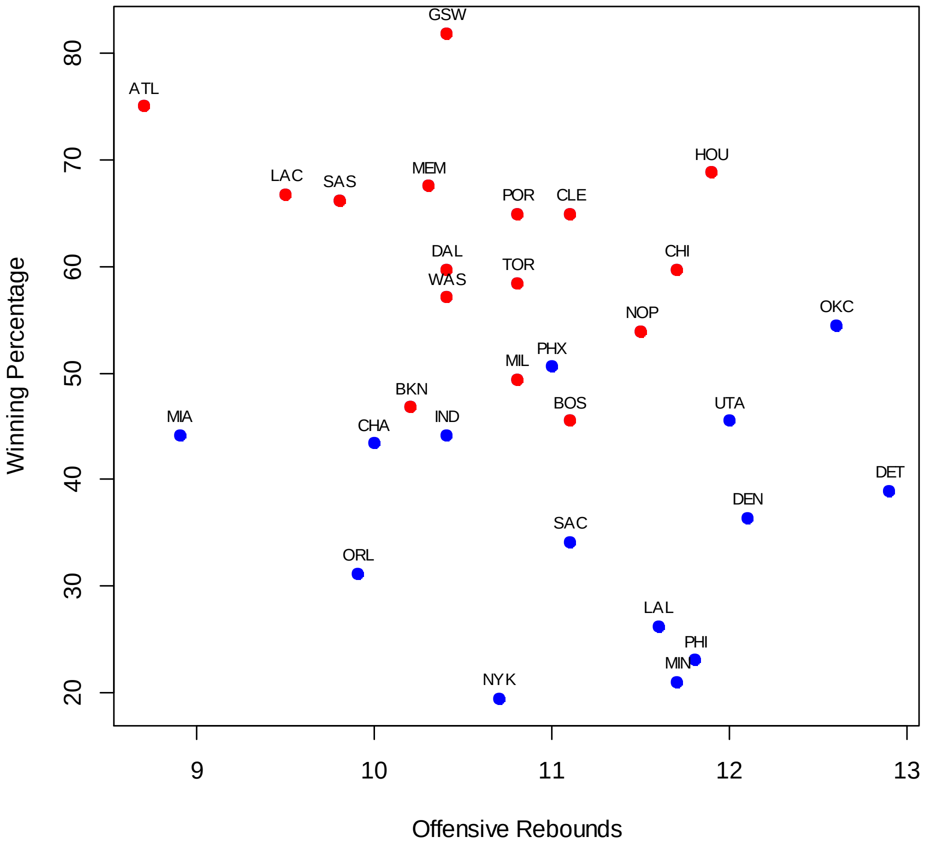
Can we find patterns among these indicators ??

Example: traditional performance indicator in NBA

% 3 points



Number of offensive rebounds



What do we want to do ?

1. Develop a performance score that will tell us who are the "best teams" (in terms of performance, i.e % Win).
2. Multivariate performance score.

3. Analysis in two steps:
   1. First step: PCA for Traditional Stats
   2. Second step: multiple factor analysis for all data (data integration).

MFA: data structure

$$X = \left[X_1 \mid \underbrace{X_2 X_3 X_4 X_5 X_6 X_7}\right]$$

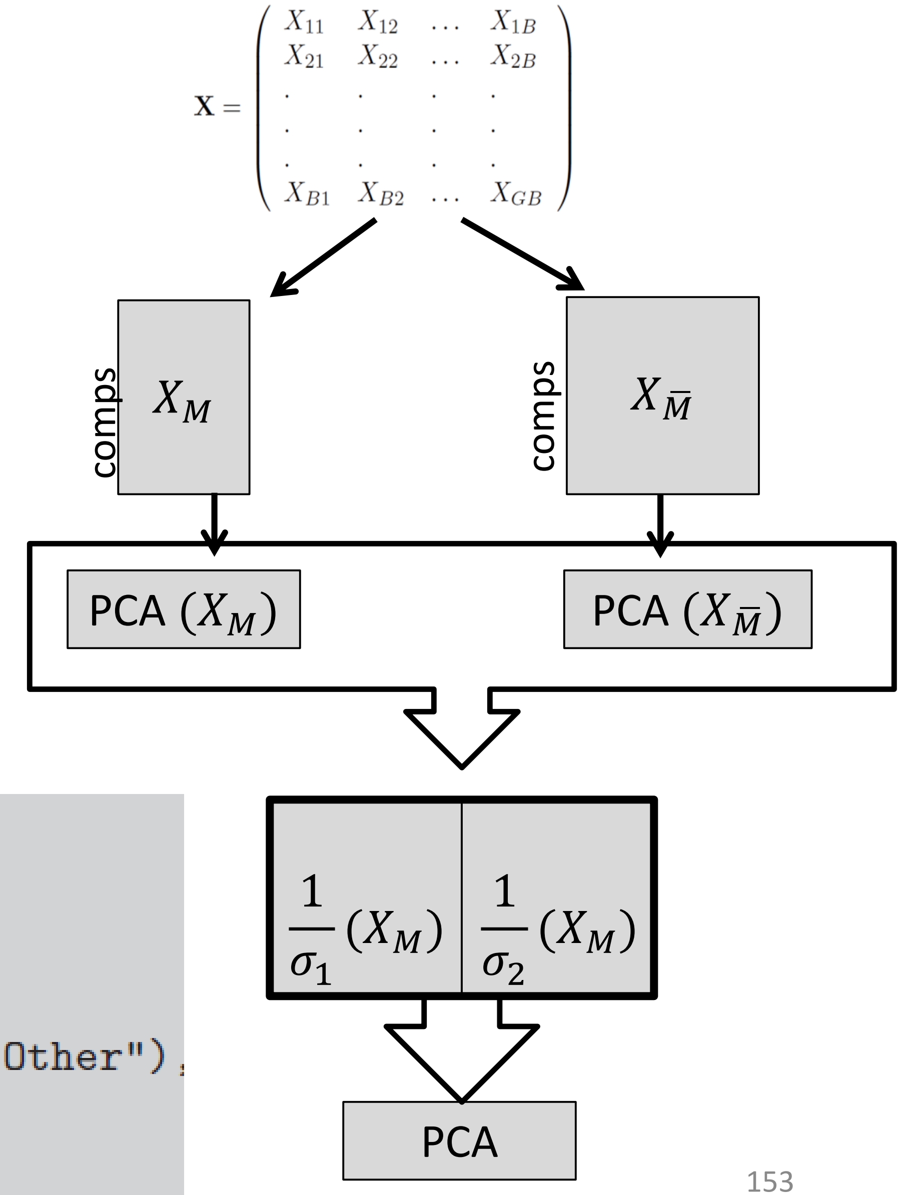
Set of the traditional stats indicators

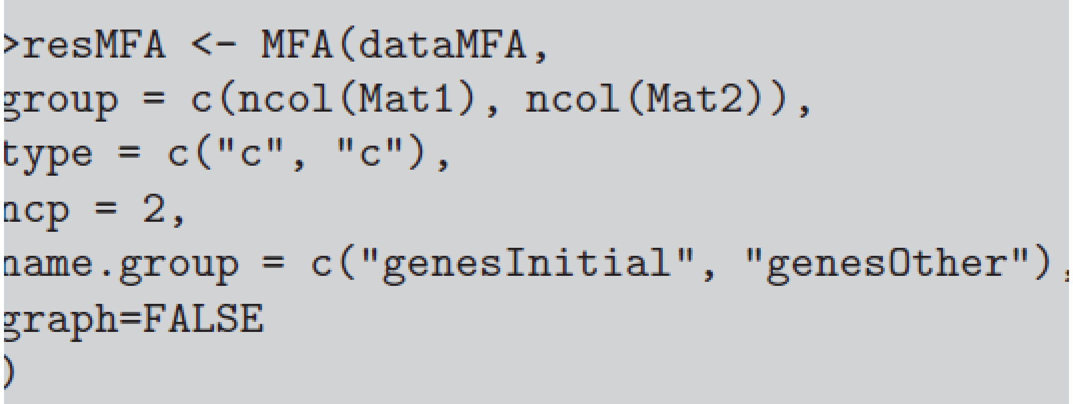
- 3-pt Percentage
- Free Throws Percentage
- Defensive Rebounds
- Offensive Rebounds
- Assists
- Turnovers
- Steals
- Field Goals Percentage
- Blocks Committed / Received
- Fouls Committed / Received

- Advanced Stats
- Four Factors
- Misc. Stats
- Scoring
- Opponent
- Shooting

Multiple factor analysis

- All variables standardized
- Normalize each data matrix
- Concatenate all normalized datasets and perform PCA on the combined weighted data
  - Factor scores describe compounds
  - Factor loadings describe variables


$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1B} \\ X_{21} & X_{22} & \dots & X_{2B} \\ \vdots & \vdots & \vdots & \vdots \\ X_{B1} & X_{B2} & \dots & X_{GB} \end{pmatrix}$
comps
X_M
comps
$X_{\bar{M}}$
PCA (X_M)
PCA $(X_{\bar{M}})$
$\frac{1}{\sigma_1}(X_M)$
$\frac{1}{\sigma_2}(X_M)$
PCA



```
>resMFA <- MFA(dataMFA,
group = c(ncol(Mat1), ncol(Mat2)),
type = c("c", "c"),
ncp = 2,
name.group = c("genesInitial", "genesOther"),
graph=FALSE
)
```

Step 1:
PCA for the leading performance indicator

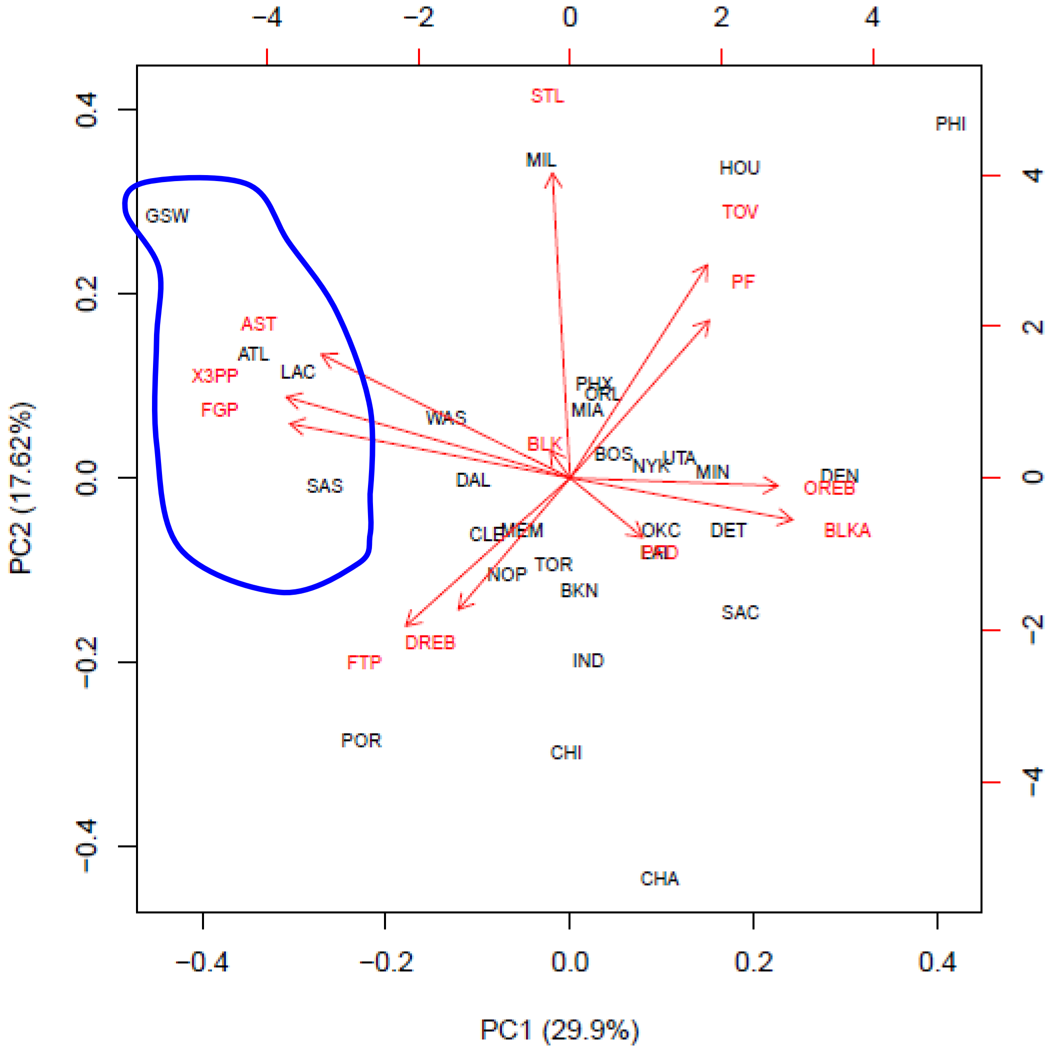
Leading Performance Indicators:

- 3-pt Percentage
- Free Throws Percentage
- Defensive Rebounds
- Offensive Rebounds
- Assists
- Turnovers
- Steals
- Field Goals Percentage
- Blocks Committed / Received
- Fouls Committed / Received

PCA:

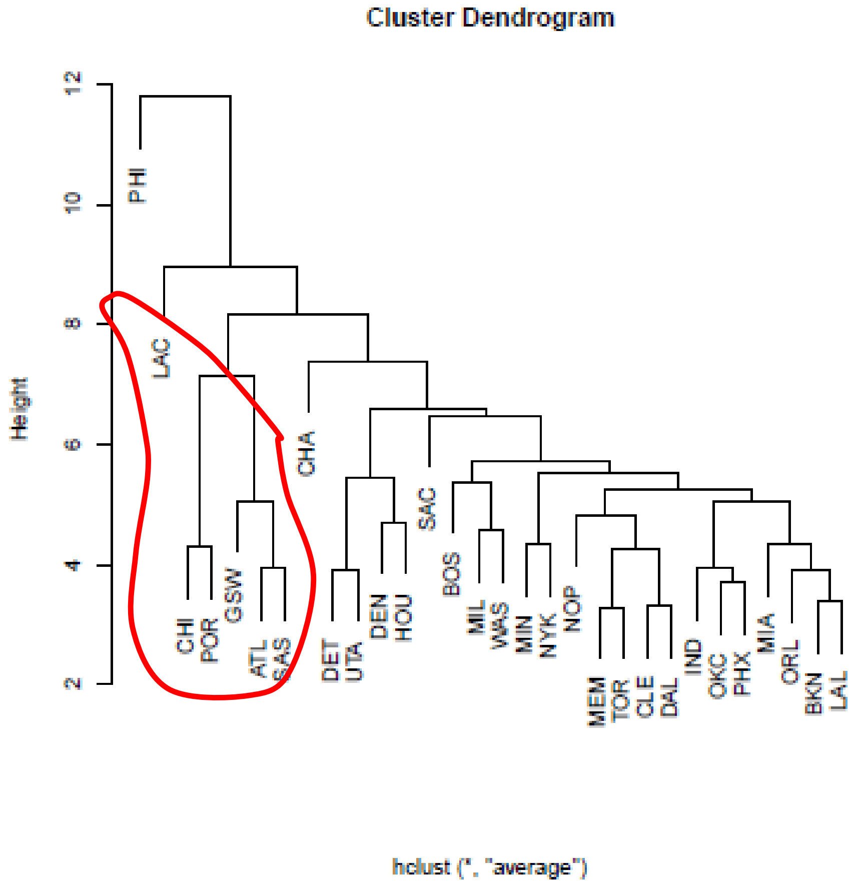
To convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables

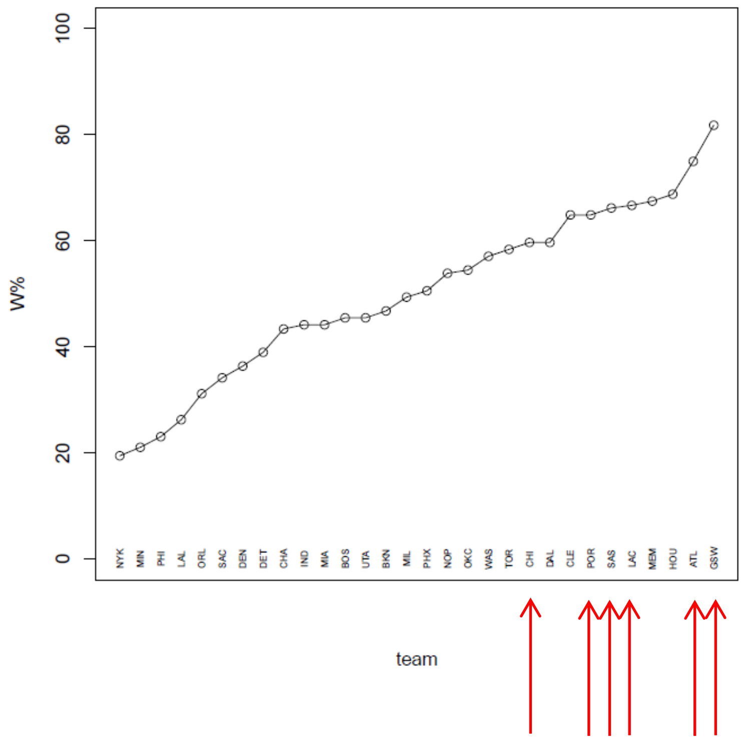
PCA for the leading performance indicator



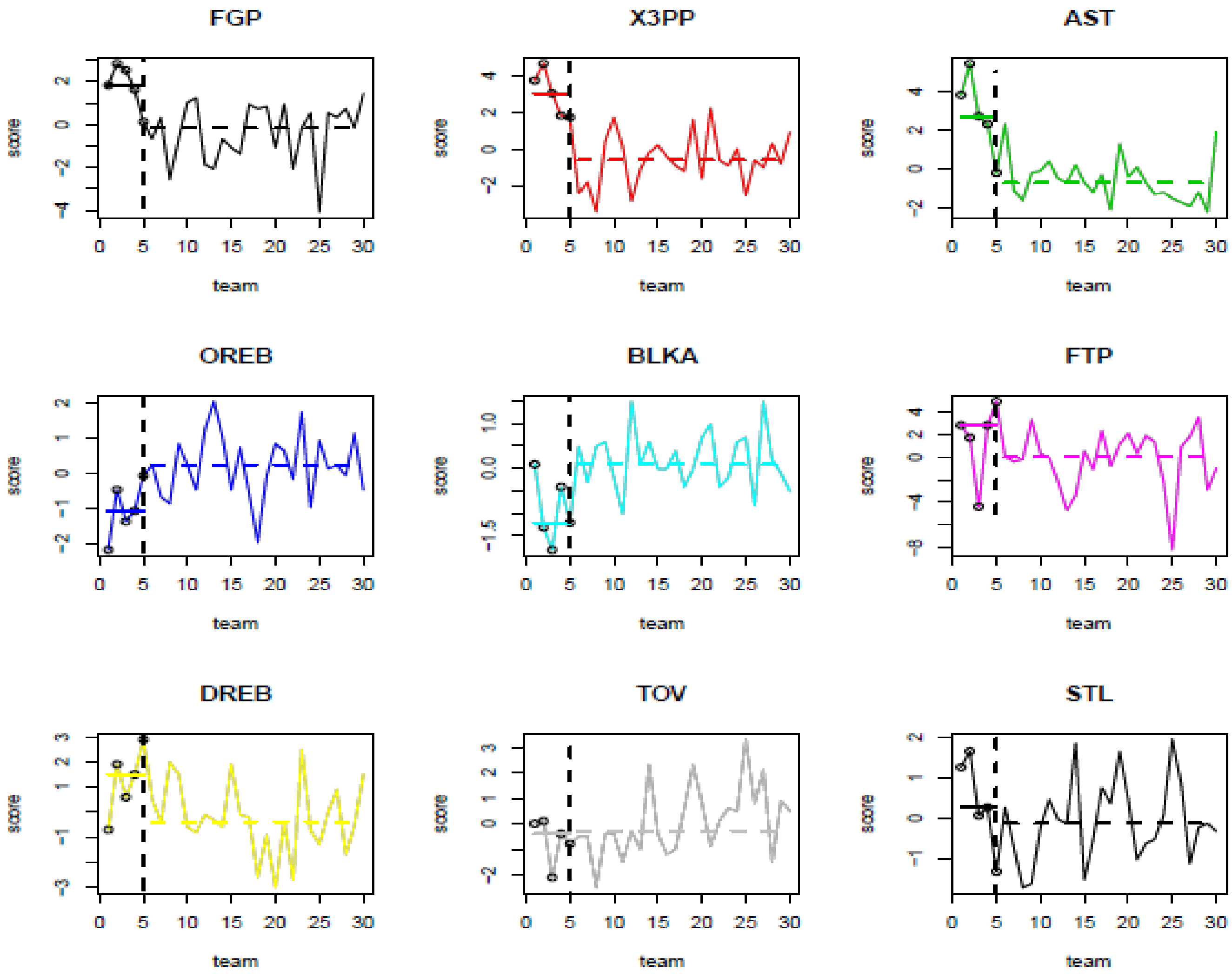
- Clear cluster of teams based on PC1
- Similar pattern was detected by Hierarchical clustering
- Indicators:
  - AST
  - X3PP
  - FGP

Hierarchical clustering

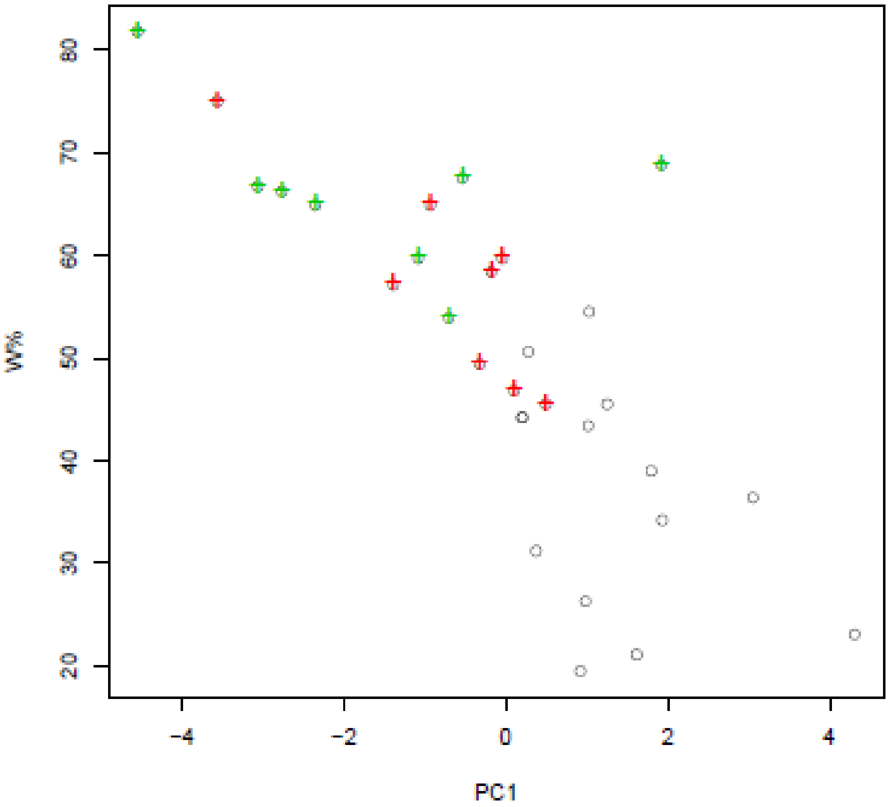




PCA for the leading performance indicator



Performance score



First PC versus % win.

Correlation.

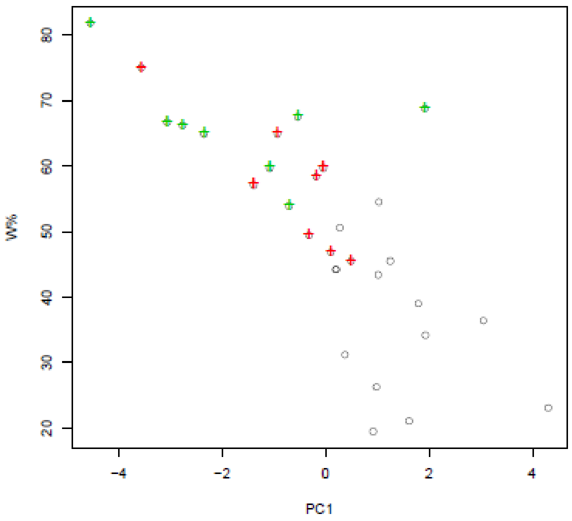
Performance score:

$$PC_1 = \sum \ell_j X_j$$

Data integration: MFA

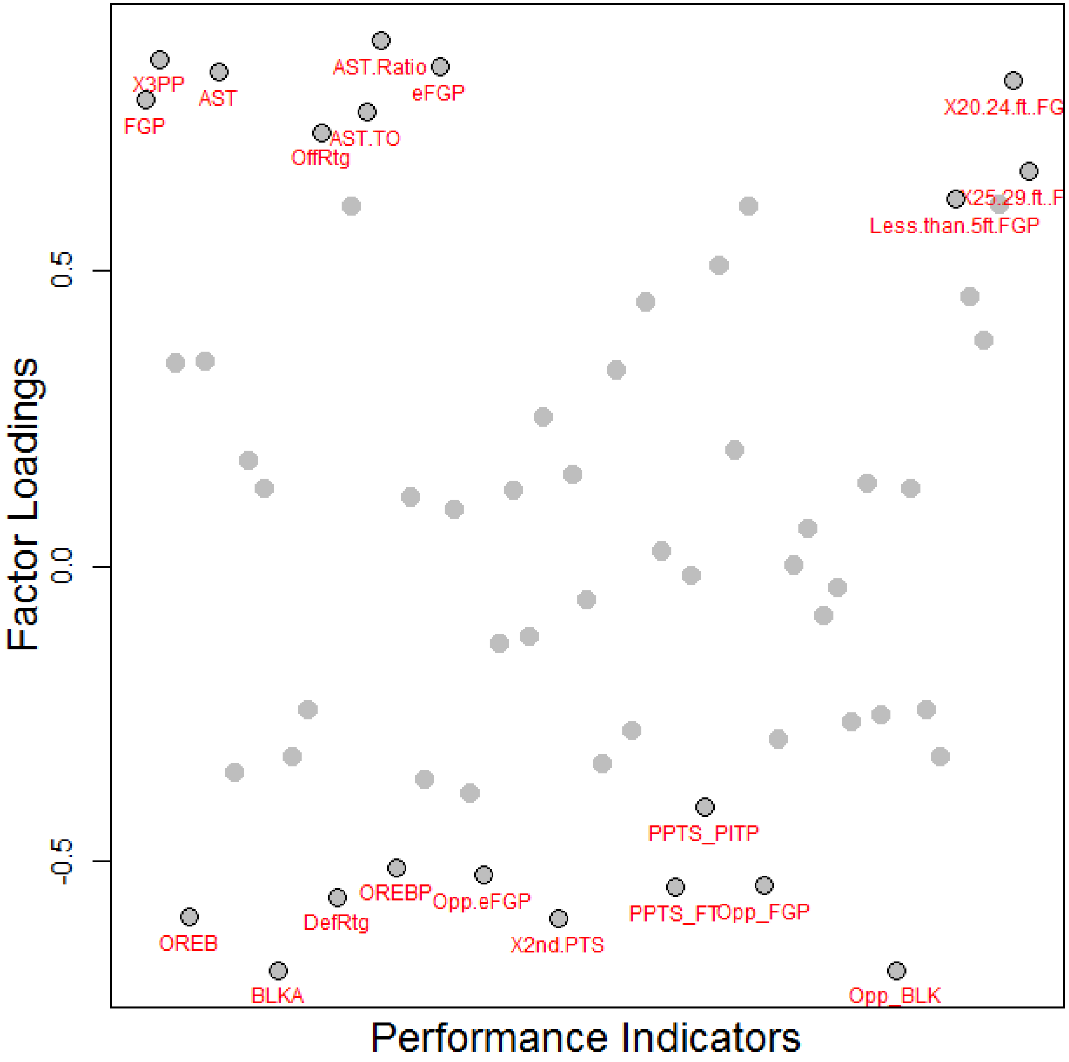
$$X = \left[X_1 \mid \underbrace{X_2 X_3 X_4 X_5 X_6 X_7}\right]$$

traditional stats indicators



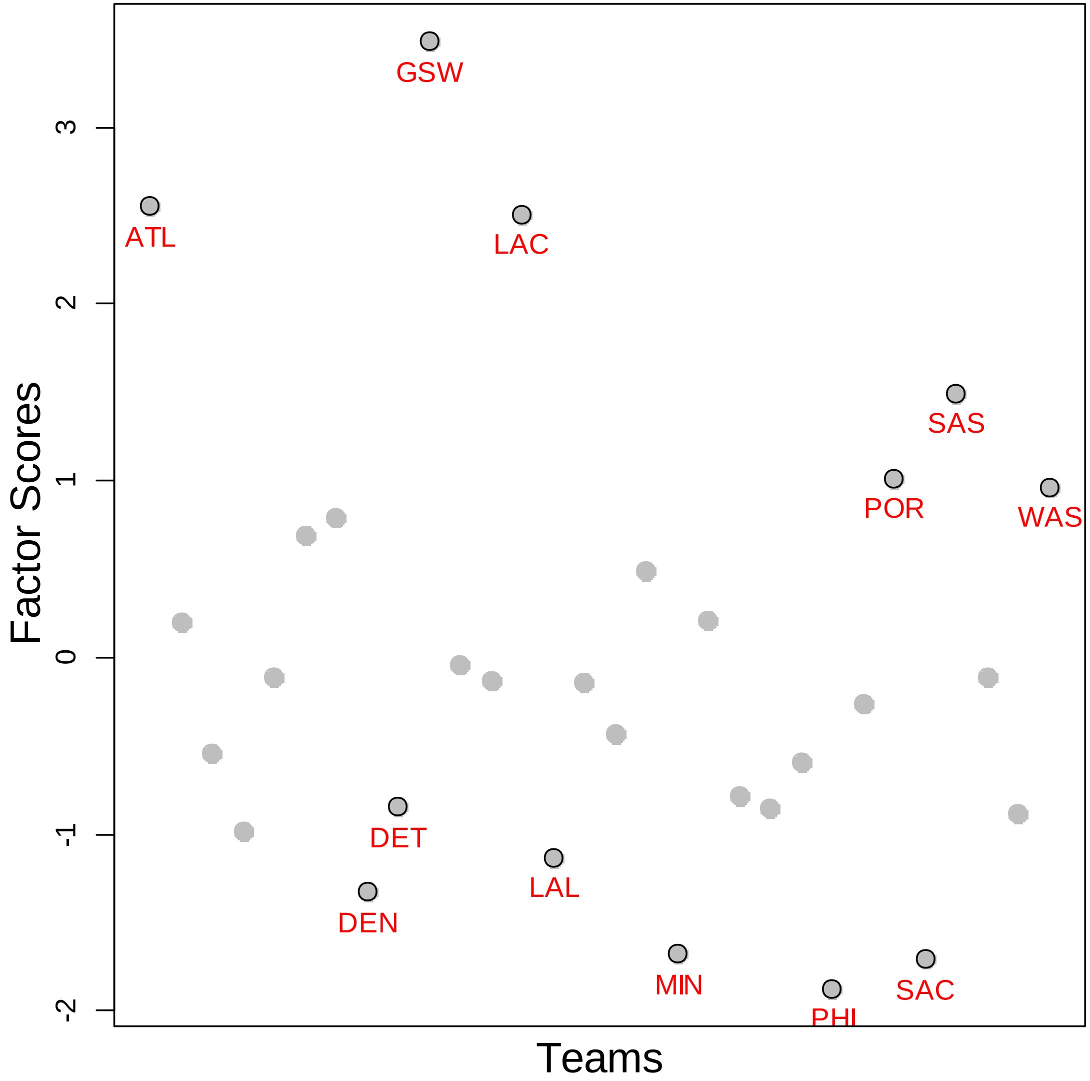
- Advanced Stats
- Four Factors
- Misc. Stats
- Scoring
- Opponent
- Shooting

MFA: factor loadings (variables)



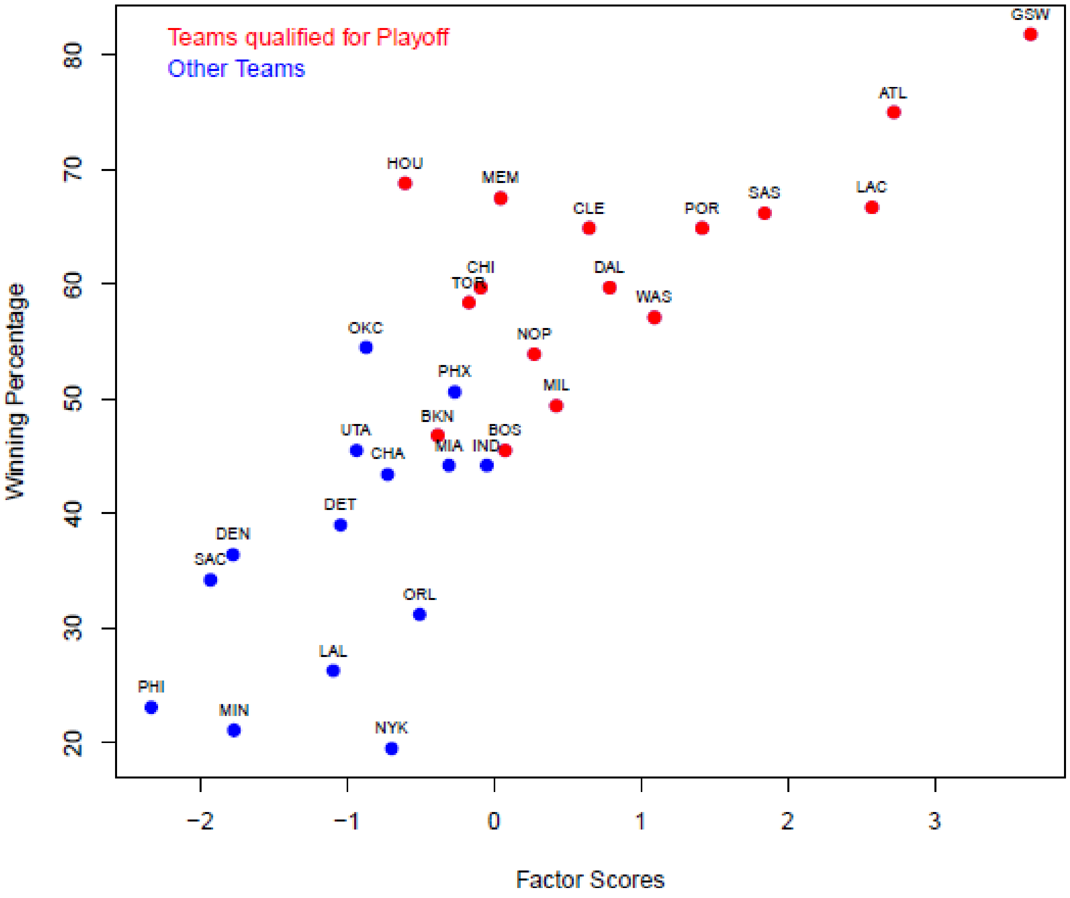
- Traditional Stats with high loadings
- Some new indicators were discovered

MFA: factor scores (teams)



- Similar set of teams with high factor scores

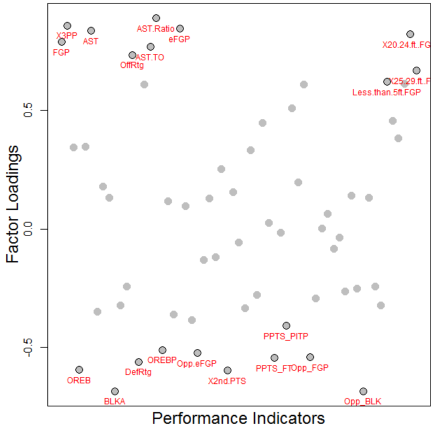
Multivariate performance score



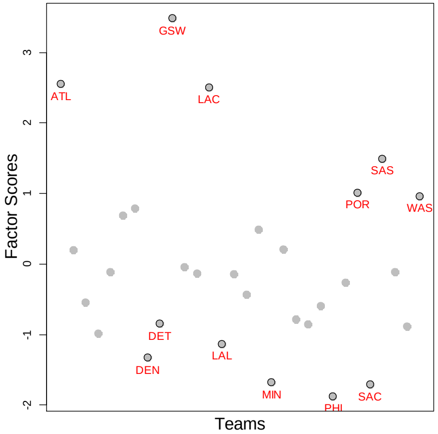
$$MPS(team_k) = \sum \ell_j X_j$$

Biclustering using FABIA

loadings (variables)

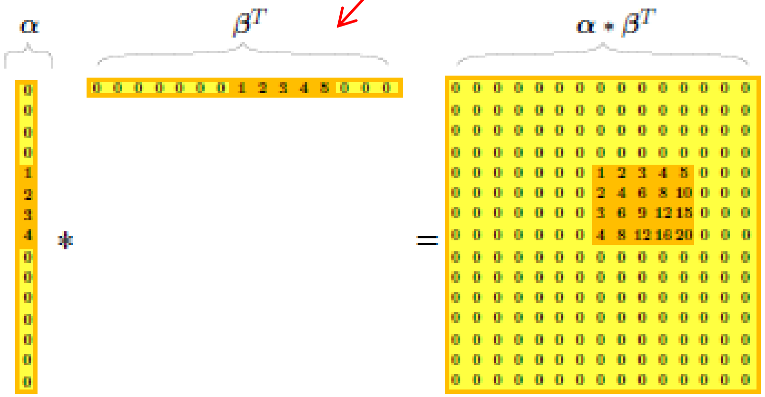


scores (teams)



MFA:
After normalization: FA
with one factor

$$X = [X_1 \mid X_2 X_3 X_4 X_5 X_6 X_7]$$

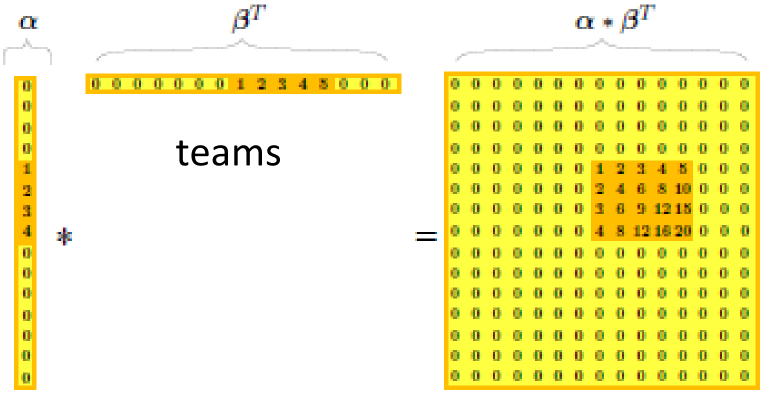


$$\overbrace{\begin{matrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{matrix}}^{\alpha} * \overbrace{\begin{matrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{matrix}}^{\beta^T} = \overbrace{\begin{matrix}0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\end{matrix}}^{\alpha * \beta^T}$$

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{G1} & X_{G2} & \dots & X_{Gn} \end{pmatrix} .$$

FABIA with one factor (BC)

Applying FABIA: data structure + model



$$\overbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}^{\alpha} * \overbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}^{\beta^T} = \overbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}^{\alpha * \beta^T}$$

teams

Performance indicators

Data

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{G1} & X_{G2} & \ldots & X_{Gn} \end{pmatrix} .$$

teams

Performance indicators

The model

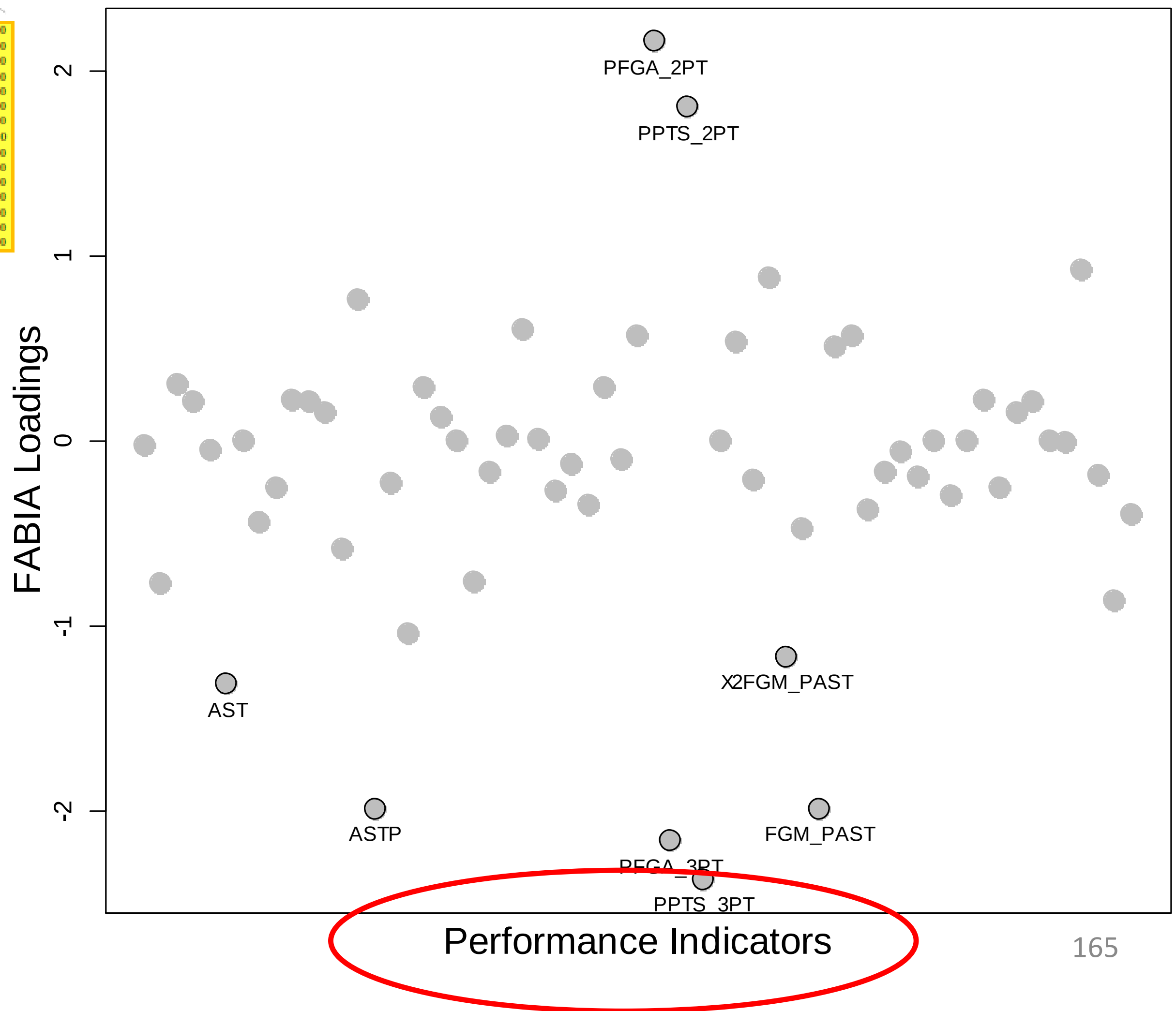
$$\mathbf{X} = \sum_{i=1}^{p} \lambda_i {\gamma_i}^T + \Upsilon,$$

Aim:

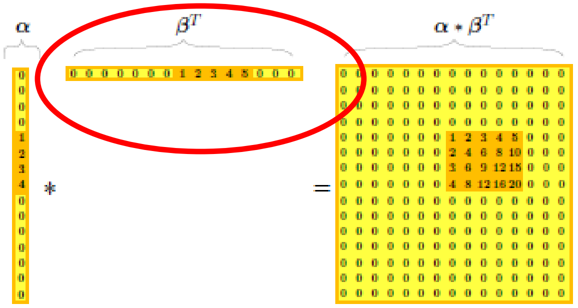
1. Find a group of teams that share patterns in performance indicators.
2. Correlation to overall performance.

BC1 FABIA: factor loadings

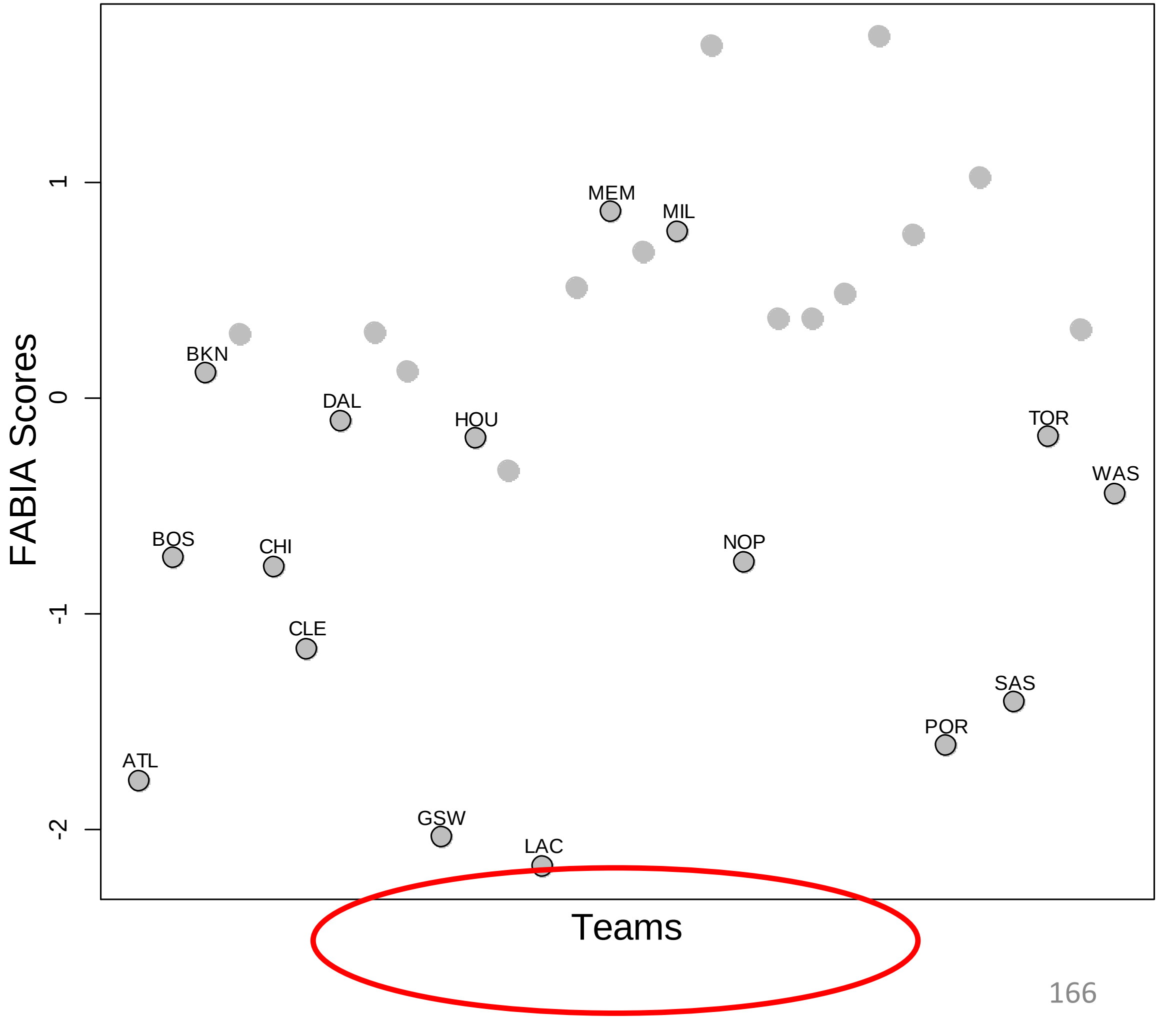
$$\underbrace{\begin{pmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}}_{\alpha} * \underbrace{\begin{pmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{pmatrix}}_{\beta^T} = \underbrace{\begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}}_{\alpha * \beta^T}$$



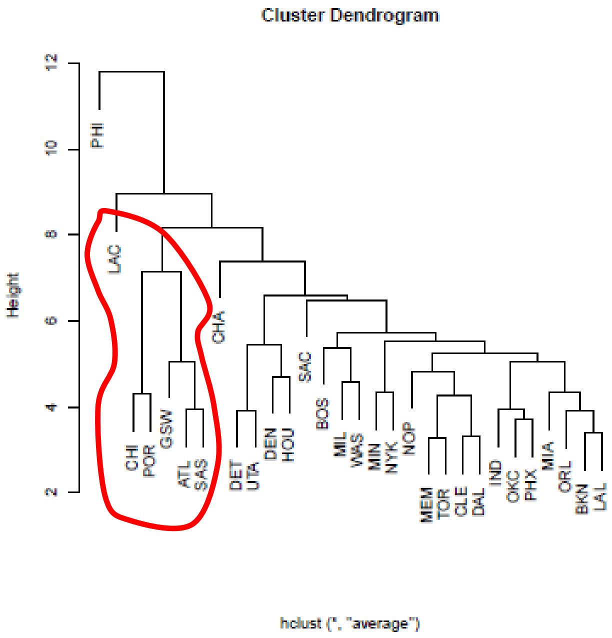
BC1 FABIA: factor scores

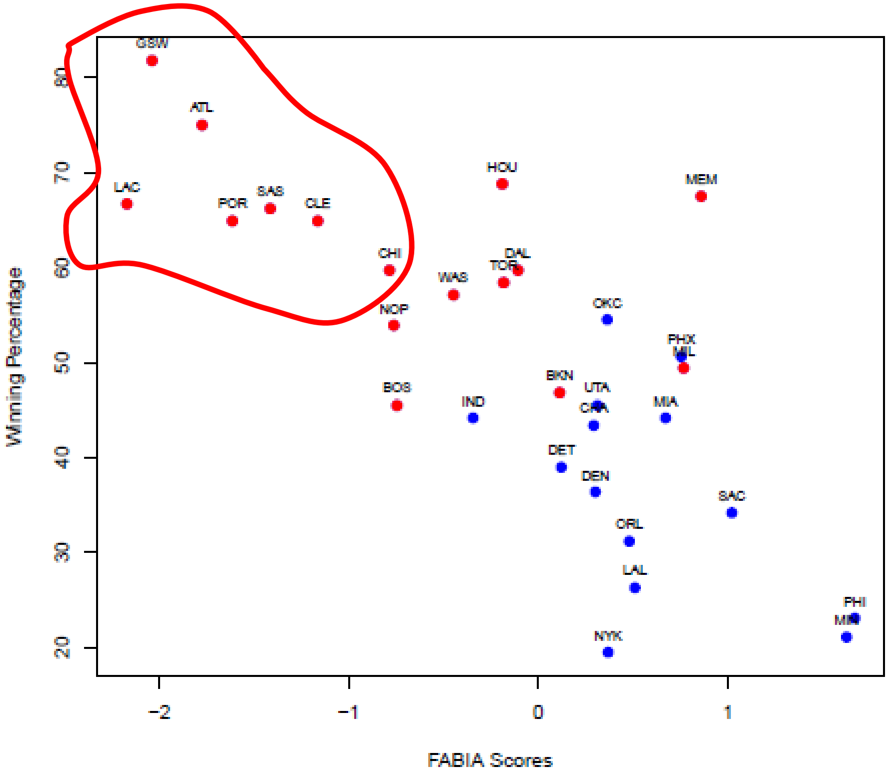


$$\overbrace{\begin{matrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{matrix}}^{\alpha} * \overbrace{\begin{matrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{matrix}}^{\beta^T} = \overbrace{\begin{matrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\
0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\
0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\
0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{matrix}}^{\alpha * \beta^T}$$

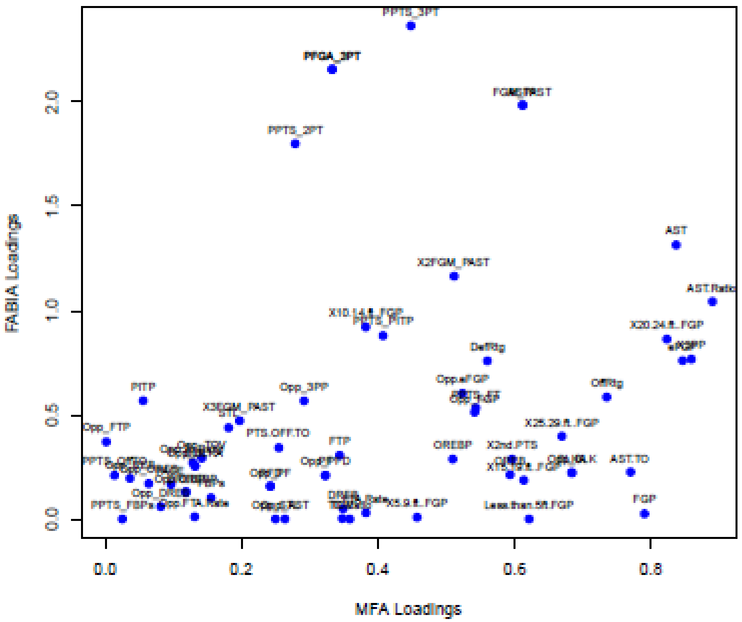


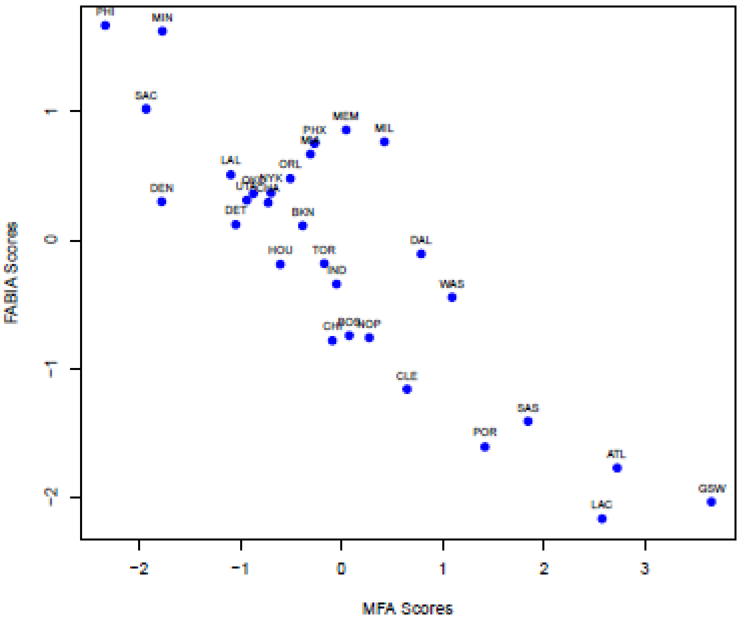
Overall performance score





FABIA and MFA





Software

Part 4.4

Enrichment of Gene Expression Modules using Multiple Factor Analysis and Biclustering

Motivation

- A bicluster contains

  genes that are

  coordinately regulated

  under a subset of conditions

Gene module

- Summarized expression profiles of these genes
- Many biclusters -> many possible gene modules

Motivation

- A bicluster contains

  genes that are not only

  coordinately regulated

  under a subset of conditions

  but are also mostly functionally coherent.

Gene module

- Summarized expression profiles of these genes that act in concert to carry out a specific function

Motivation

- A bicluster contains

  genes that are not only

  coordinately regulated

  under a subset of conditions

  but are also mostly functionally coherent.

- Availability of a subset of “lead” genes/compounds
  - Genes related to a phenotype of interest
  - Genes that are known to be part of a biological pathway
  - Some hypothesis generated from previous experiments

Aim

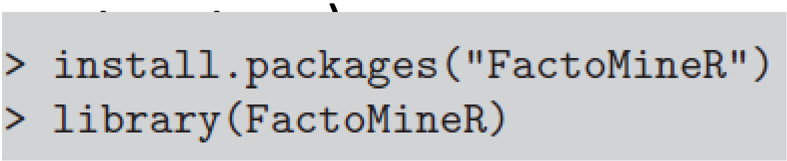
- To enrich this set of *M* “lead” genes

$$\mathbf{X}_M = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{M1} & X_{M2} & \ldots & X_{Mn} \end{pmatrix}$$

Idea

- Run biclustering algorithm
  - Search for the bicluster that contains most of the genes in the list of “lead” genes
  - Not necessarily the first (ranking) bicluster
  - Dependent on the sparsity parameter, etc.

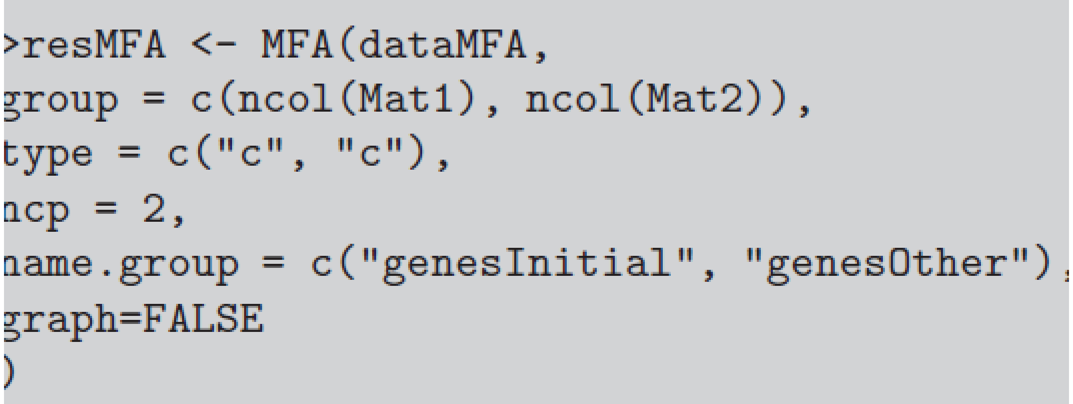
- MFA
  - find links between datasets (presence of common structure)



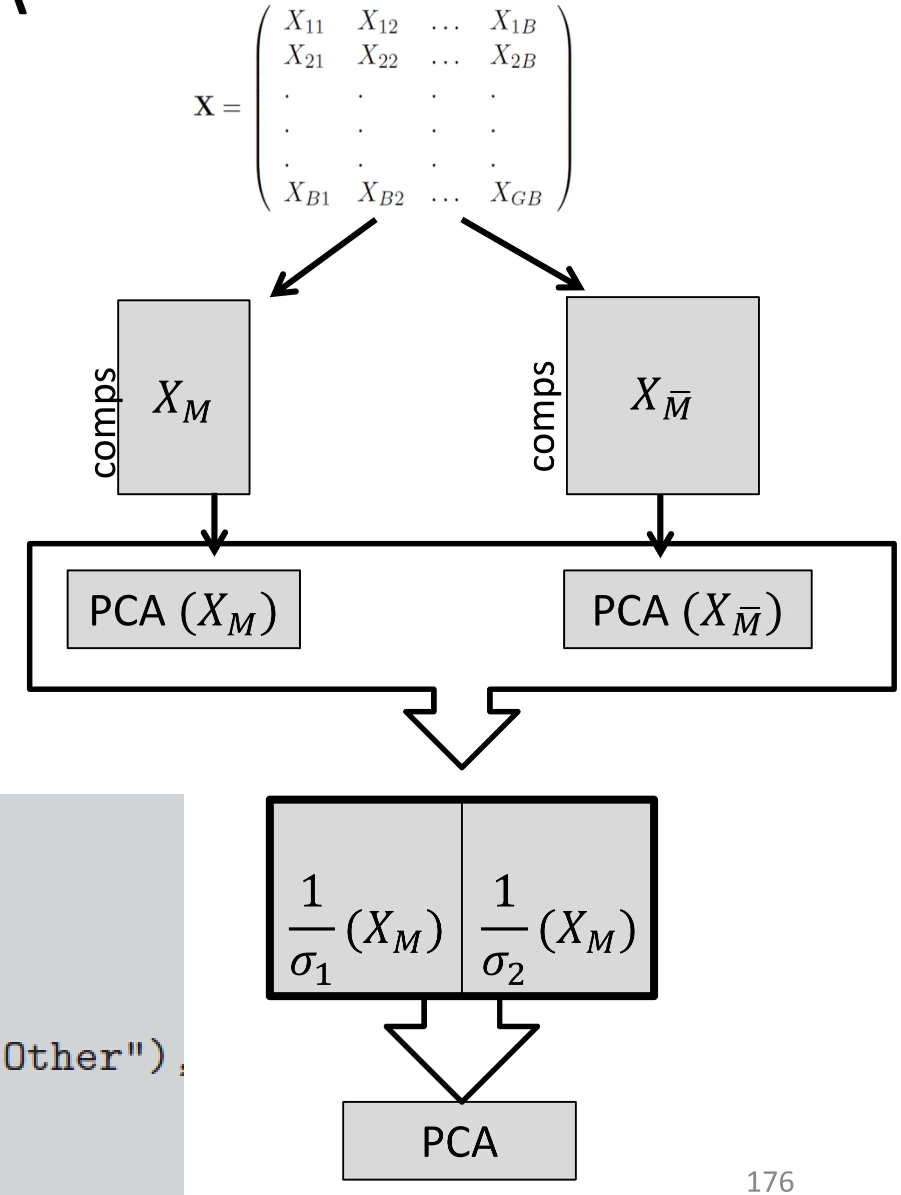
```
> install.packages("FactoMineR")
> library(FactoMineR)
```

MFA

- All variables standardized
- Normalize each data matrix
- Concatenate all normalized datasets and perform PCA on the combined weighted data
  - Factor scores describe compounds
  - Factor loadings describe variables



```
>resMFA <- MFA(dataMFA,
group = c(ncol(Mat1), ncol(Mat2)),
type = c("c", "c"),
ncp = 2,
name.group = c("genesInitial", "genesOther"),
graph=FALSE
)
```

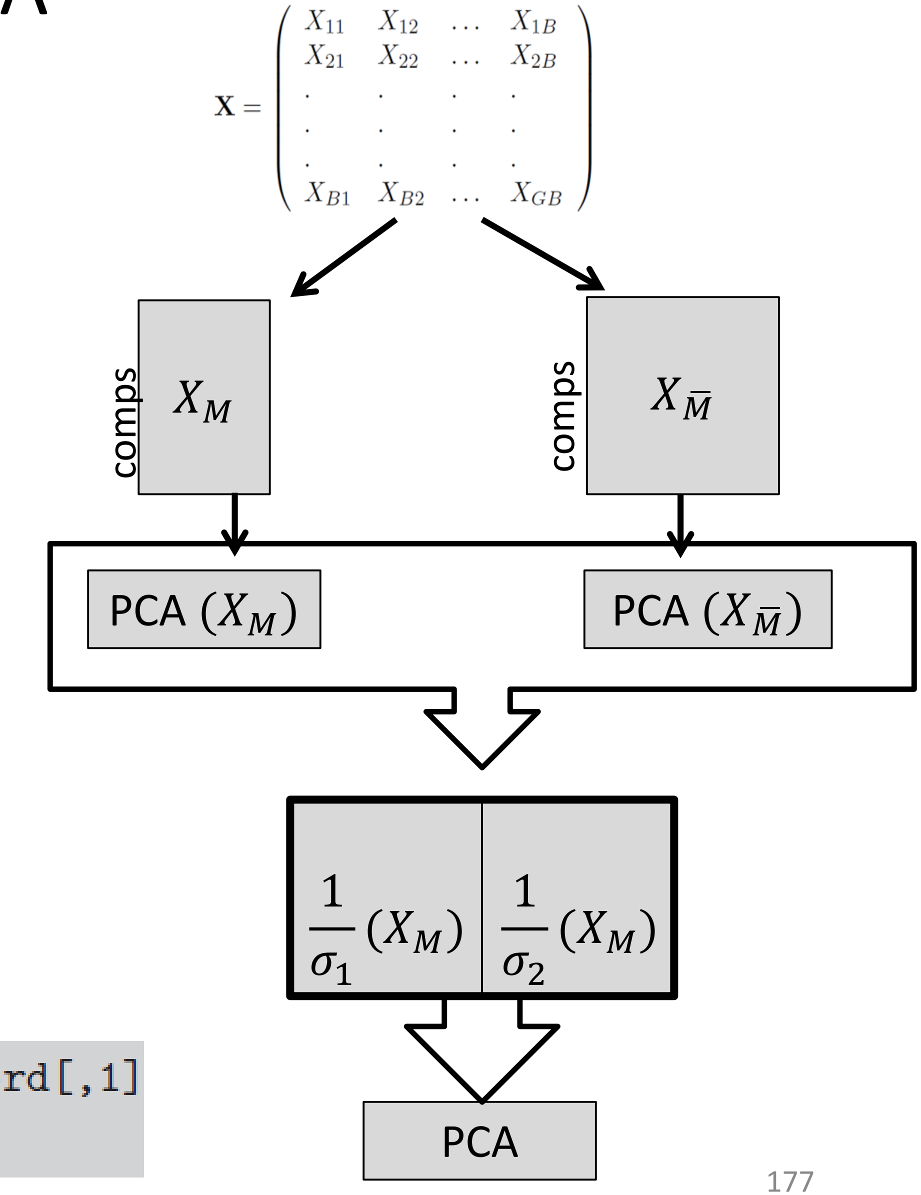


$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1B} \\ X_{21} & X_{22} & \dots & X_{2B} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{B1} & X_{B2} & \dots & X_{GB} \end{pmatrix}$
comps
X_M
comps
$X_{\bar{M}}$
PCA (X_M)
PCA $(X_{\bar{M}})$
$\frac{1}{\sigma_1}(X_M)$
$\frac{1}{\sigma_2}(X_M)$
PCA

MFA

- All variables standardized
- Normalize each data matrix
- Concatenate all normalized datasets and perform PCA on the combined weighted data
  - – Compound scores
  - – Gene loadings

```
> loadings1 <- resMFA$quanti.var$coord[,1]
> scores1 <- resMFA$ind$coord[,1]
```

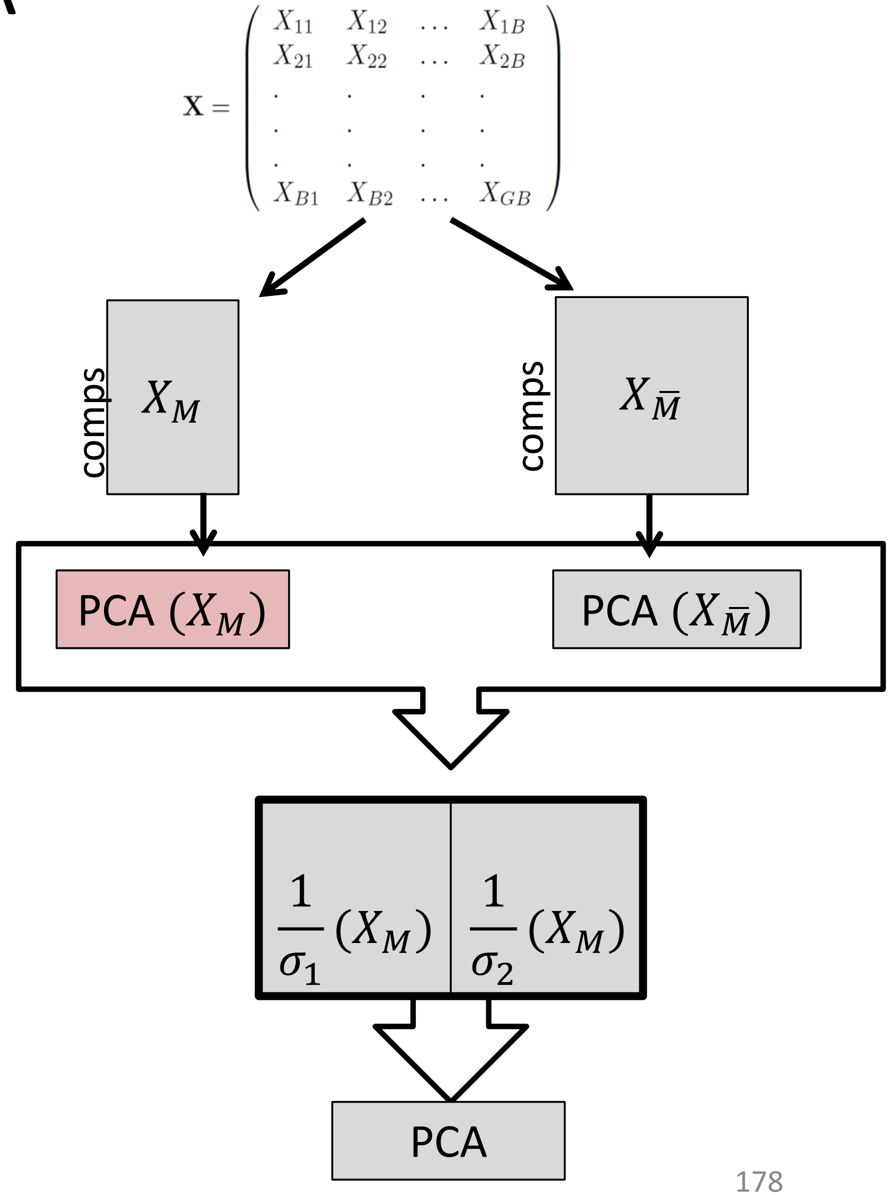


$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1B} \\ X_{21} & X_{22} & \dots & X_{2B} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{B1} & X_{B2} & \dots & X_{GB} \end{pmatrix}$
comps
X_M
comps
$X_{\bar{M}}$
PCA (X_M)
PCA $(X_{\bar{M}})$
$\frac{1}{\sigma_1}(X_M)$
$\frac{1}{\sigma_2}(X_M)$
PCA

MFA

PCA (X_M)

- a one factor solution in a factor analysis model will capture a substantial proportion from the total variability of the genes



$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1B} \\ X_{21} & X_{22} & \dots & X_{2B} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{B1} & X_{B2} & \dots & X_{GB} \end{pmatrix}$$

comps X_M

comps $X_{\bar{M}}$

PCA (X_M)

PCA $(X_{\bar{M}})$

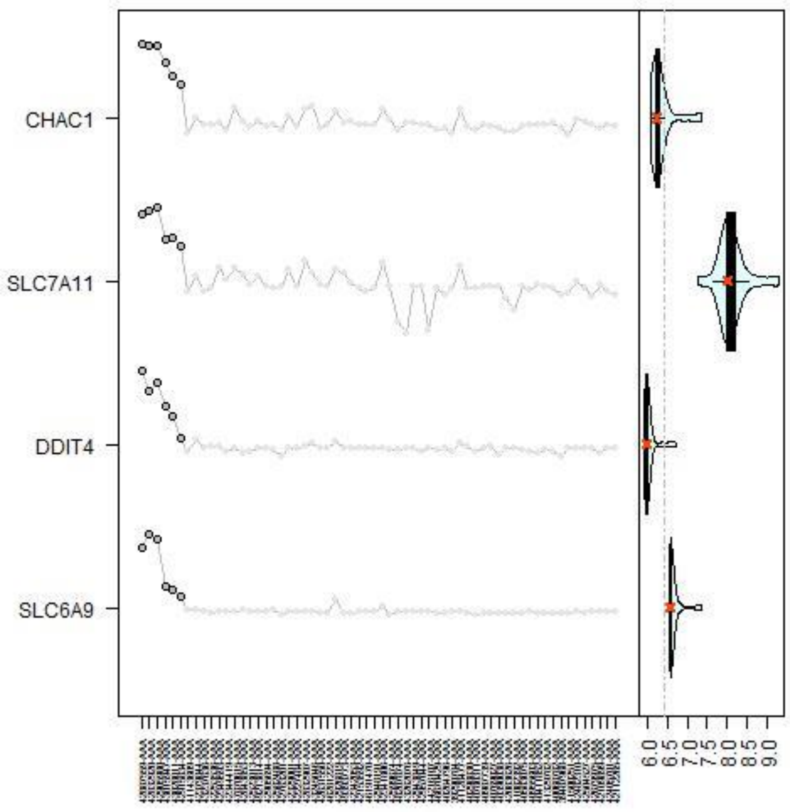
$\frac{1}{\sigma_1}(X_M)$

$\frac{1}{\sigma_2}(X_M)$

PCA

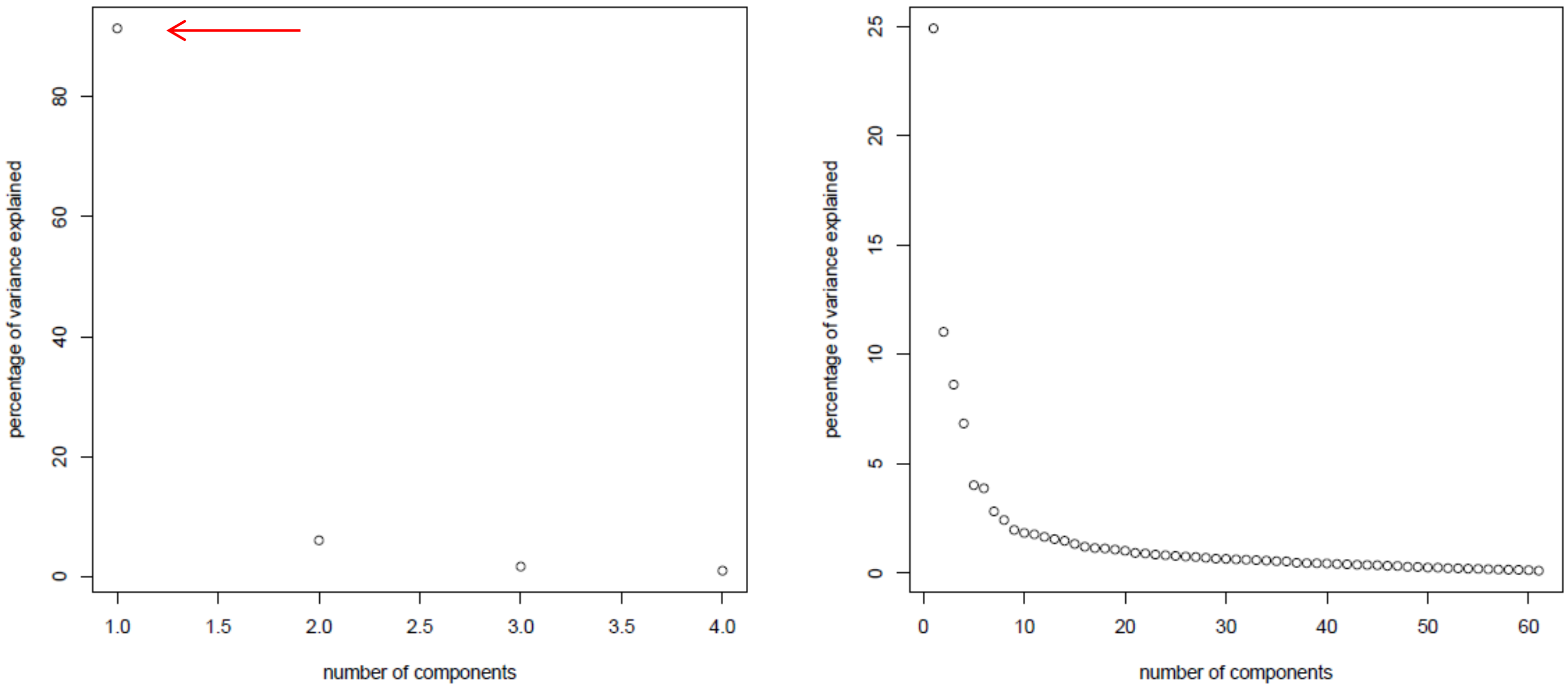
Motivating Data: mGlu2 project

- n= 62 compounds
- G = 566 genes
- M=4 genes that are known to be biologically related and are linked to the phenotype of interest.



Scree plots

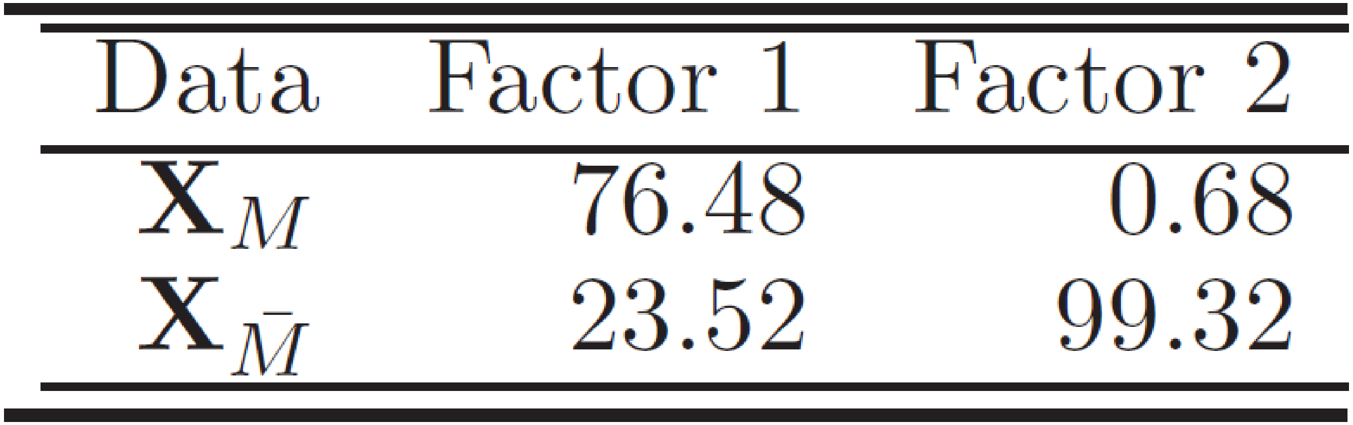
- One known structure in X_M
- Escoufier’s Rv coefficient = 26%



(a) scree plot for $\mathbf{X}_M$

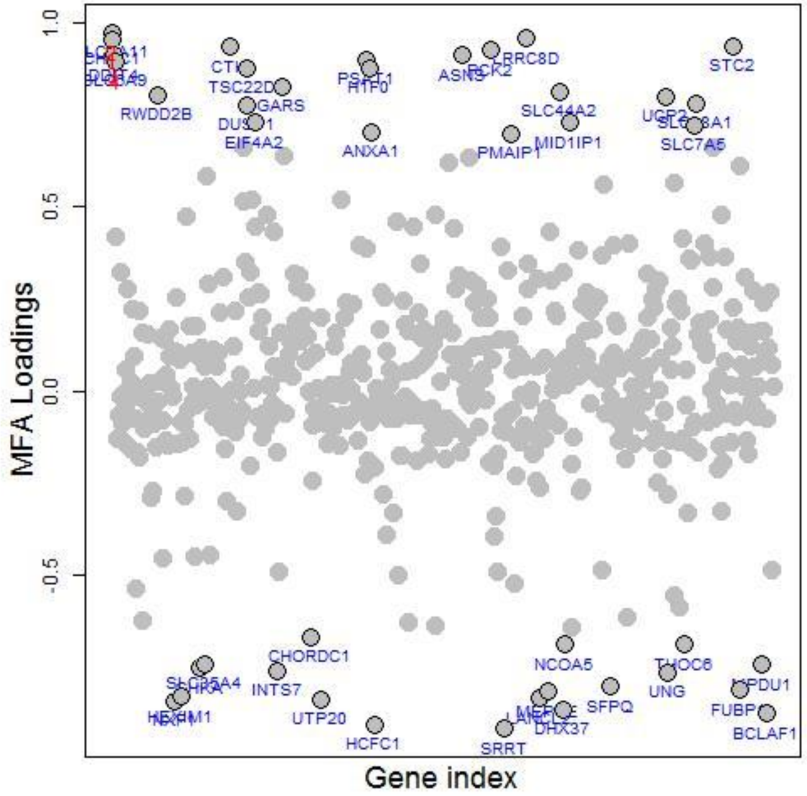
(b) scree plot for $\mathbf{X}_{\bar{M}}$

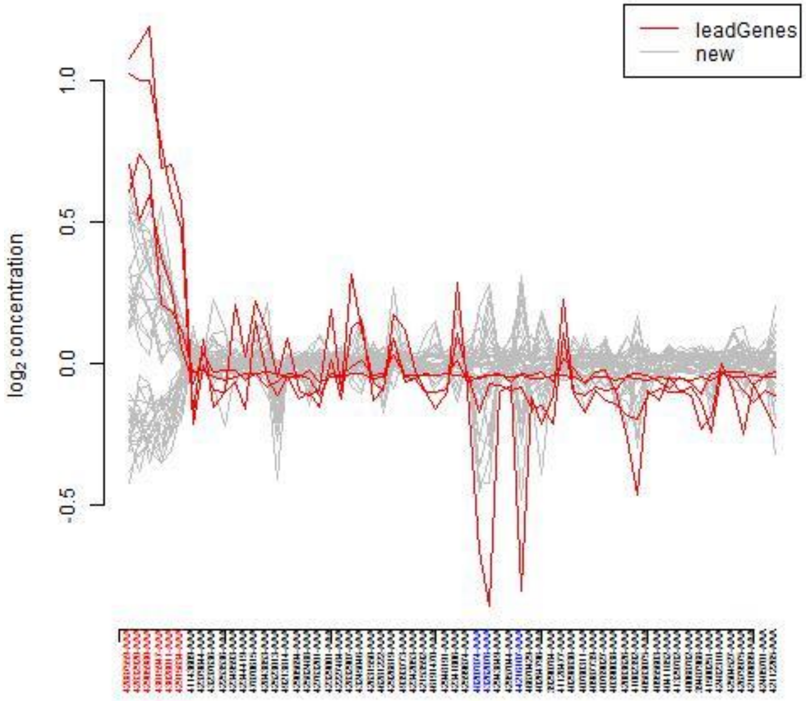
Data Contribution to the main factors



Data	Factor 1	Factor 2
$\mathbf{X}_M$	76.48	0.68
$\mathbf{X}_{\bar{M}}$	23.52	99.32

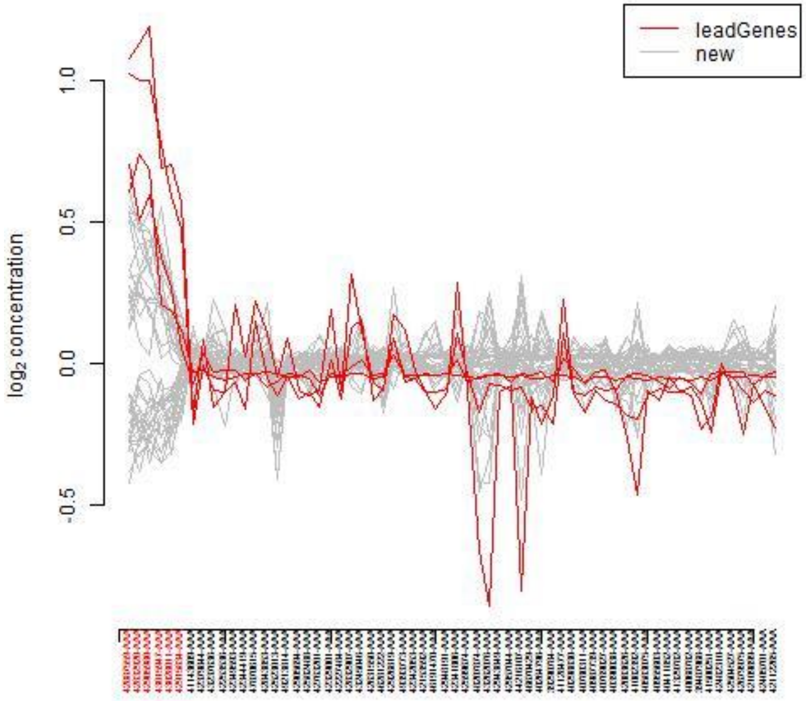
MFA 1 Gene loadings



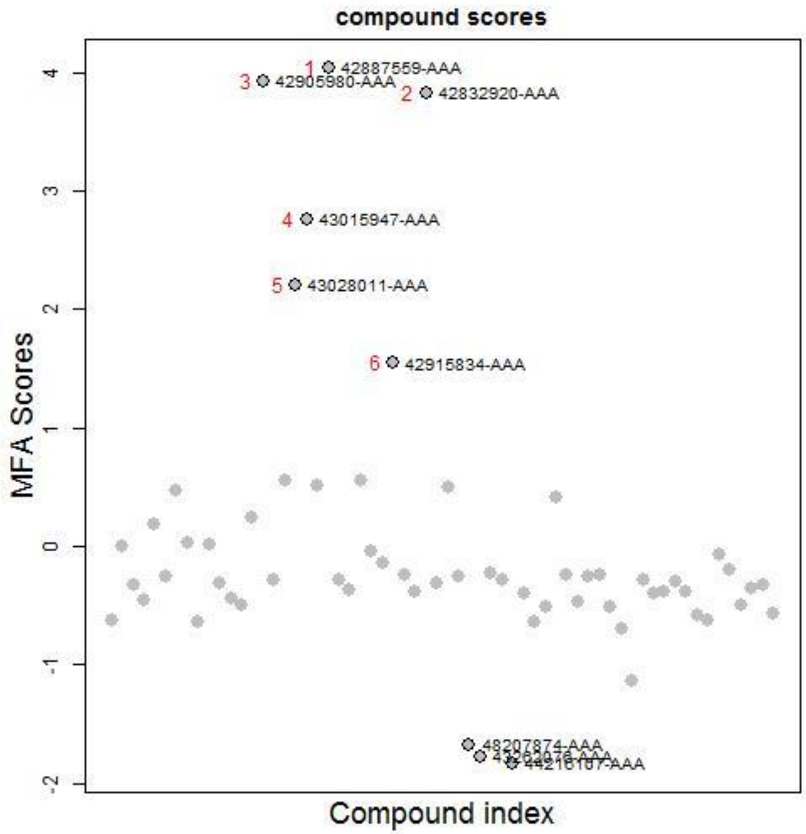


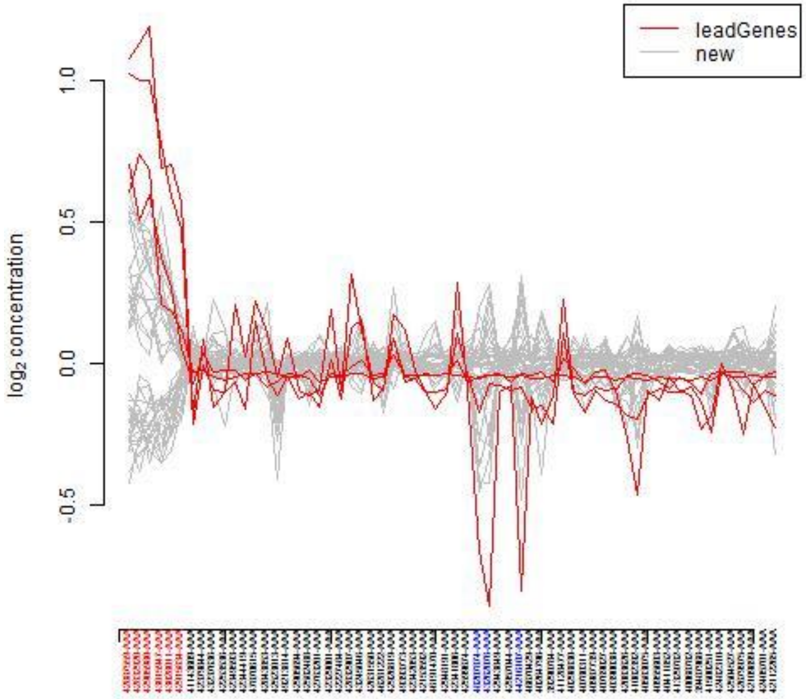
Fabia Bicluster 2

- Absence of lead genes
- Fabia searches only for correlated profiles across a subset of samples
- MFA uses the similarity of gene profiles across all compounds.
- As a result, some genes discovered by MFA are not part of the fabia bicluster and vice versa

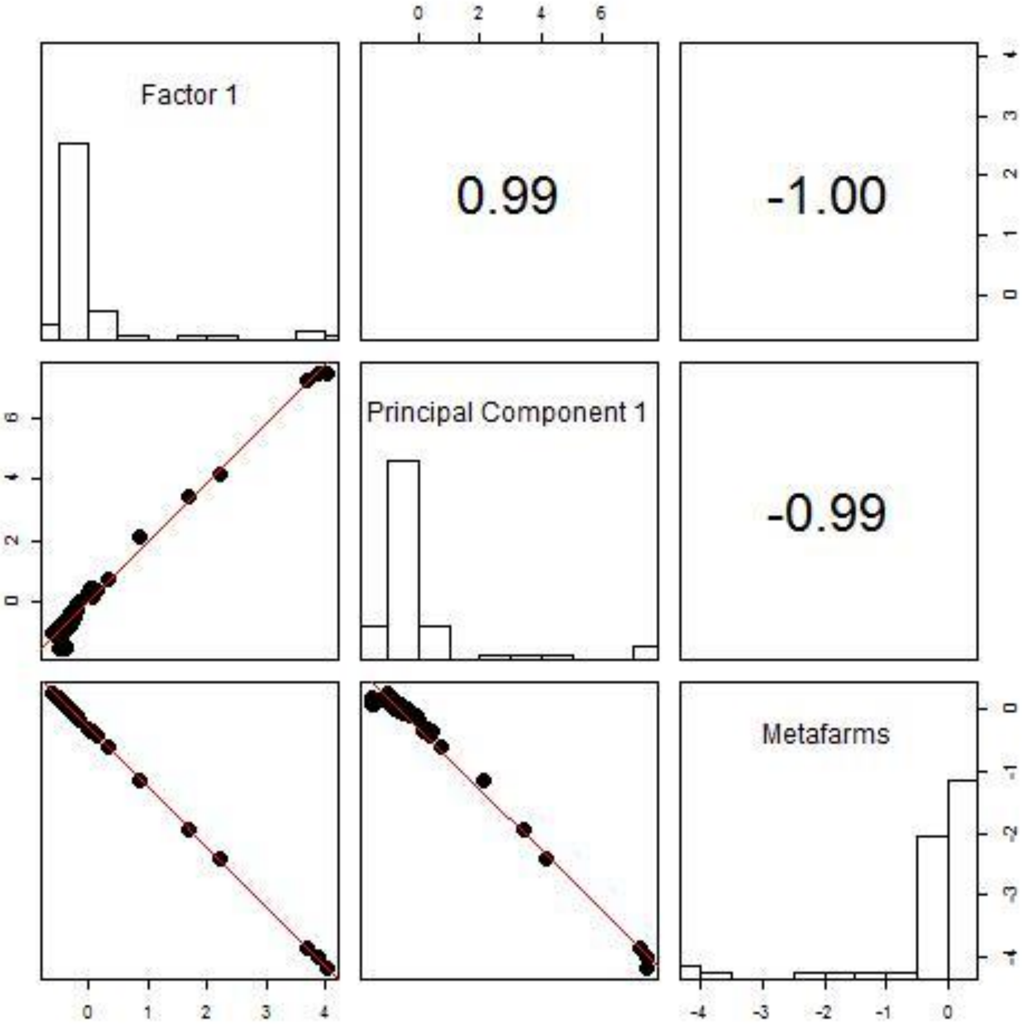


MFA 1 Compound Scores





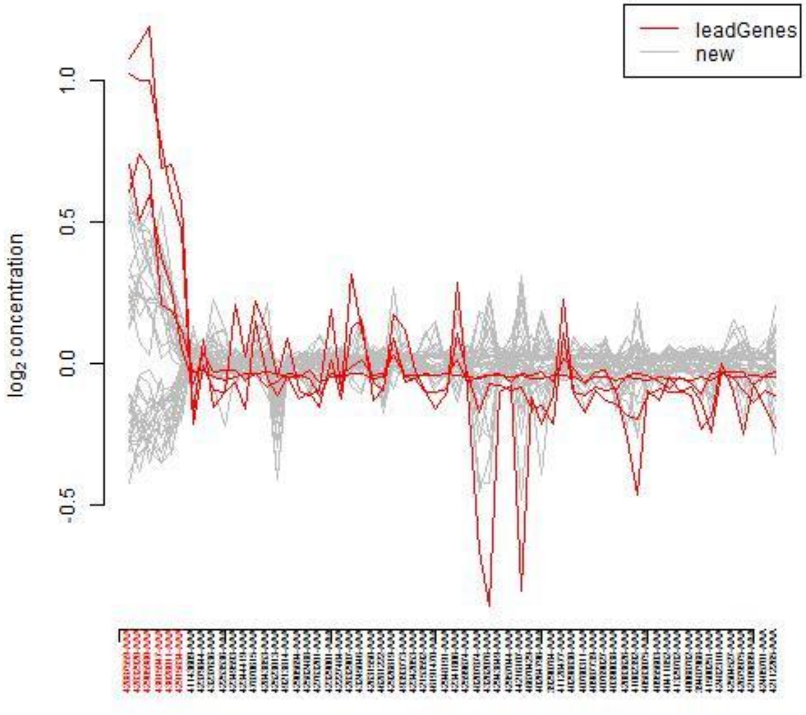
Gene Module Summarization – one



Fabia Bicluster 2

- Absence of lead genes
- Run biclustering
- explore interesting biclusters
- Lead genes are in Fabia bicluster 2

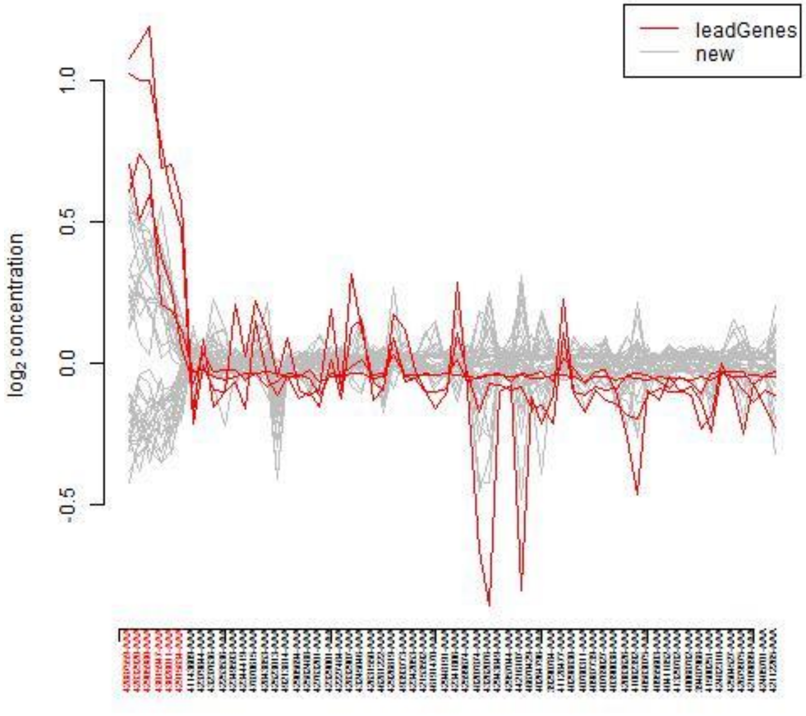
FABIA BC 2 genes



Fabia Bicluster 2

- Absence of lead genes
- Fabia → correlated profiles across a subset of samples
- MFA → similarity of gene profiles across all compounds
  - some genes discovered by MFA are not part of the fabia bicluster and vice versa

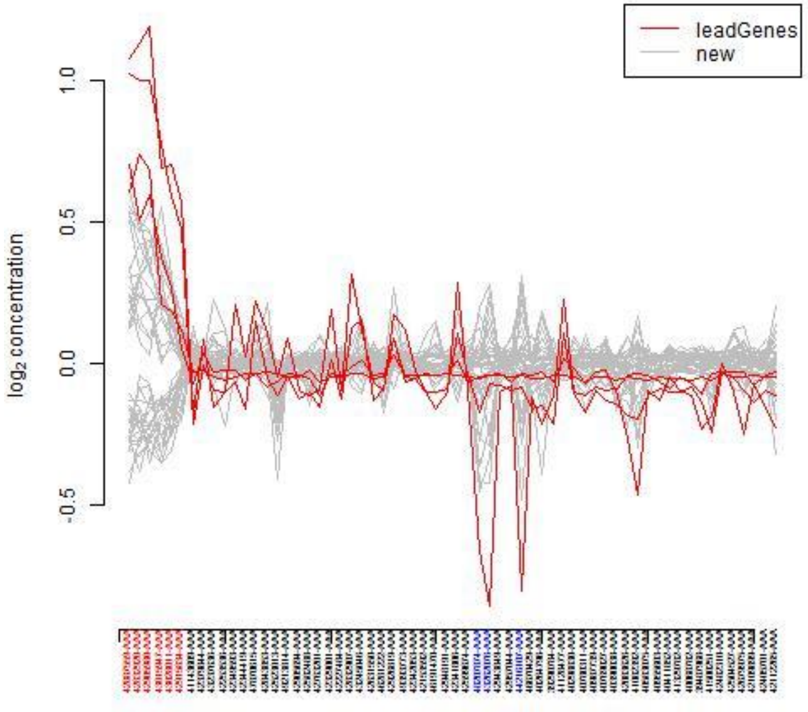
FABIA BC 2 genes



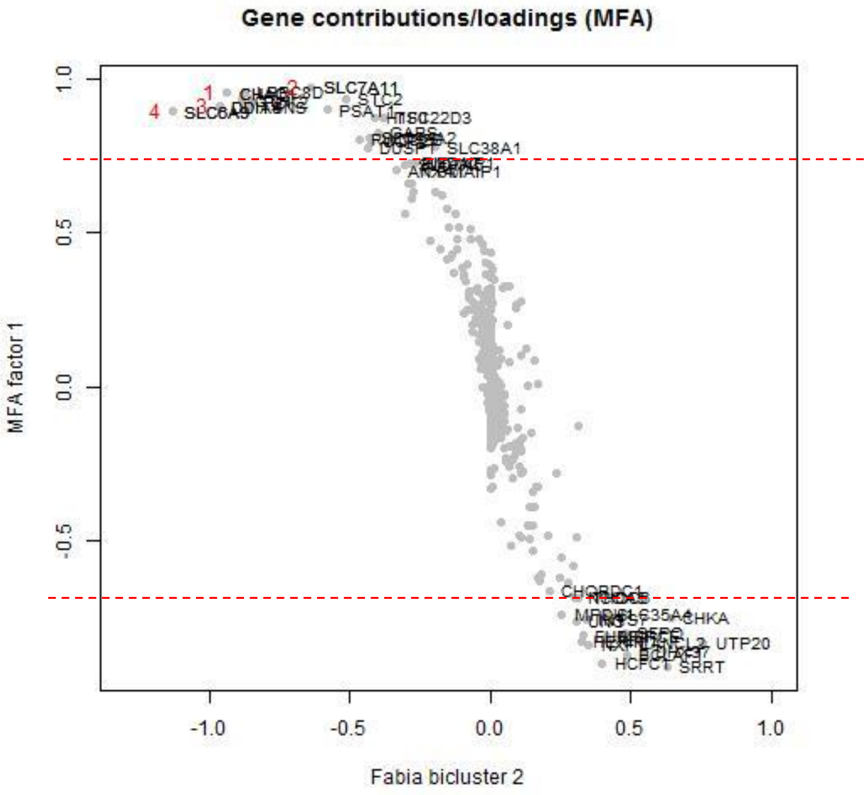
MFA 1

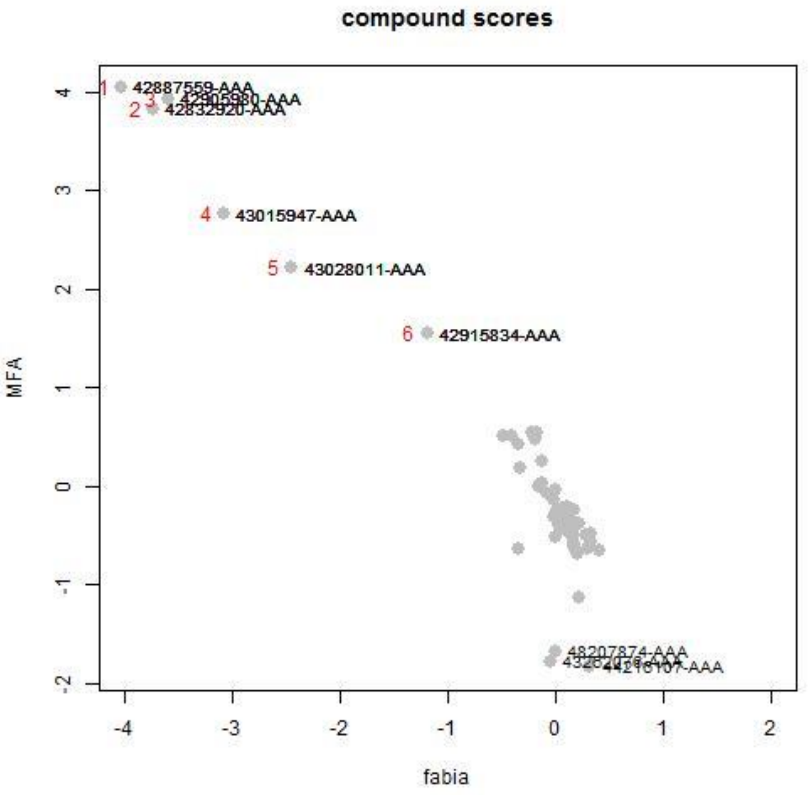
- Absence of lead genes
- Fabia → correlated profiles across a subset of samples
- MFA → similarity of gene profiles across all compounds
  - some genes discovered by MFA are not part of the fabia bicluster and vice versa

MFA 1 genes



Gene loadings and compound scores



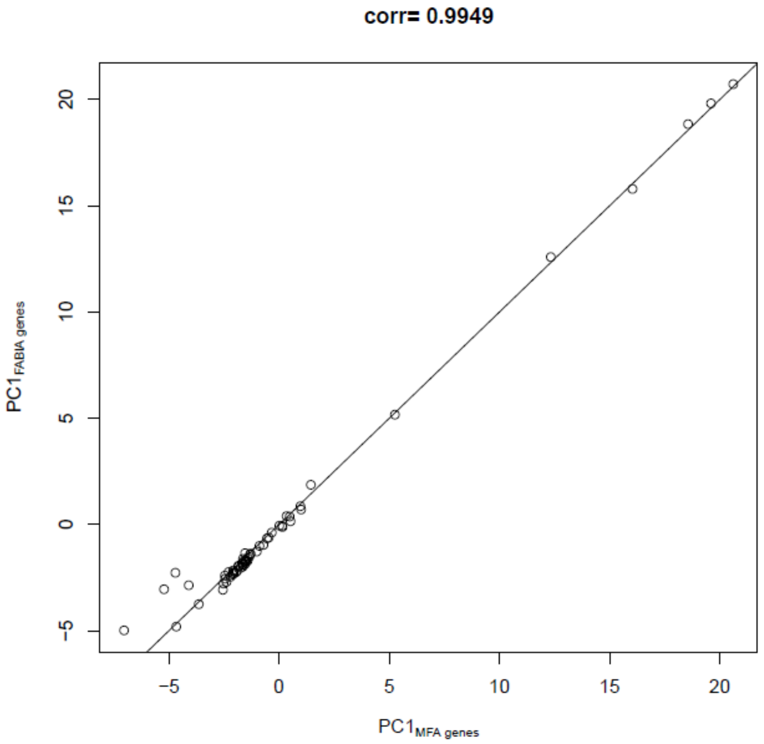


Gene Module

- Absence of lead genes
- Fabia → correlated profiles across a subset of samples
- MFA → similarity of gene profiles across all compounds
  - some genes discovered by MFA are not part of the fabia bicluster and vice versa

- The underlying latent structure is almost identical

MFA 1 genes



Part 4.5

Drug Discovery (II)

Ranking of BCs

Motivation

- how to determine which biclusters are most informative and rank them on the basis of their importance?
  - Data-driven, statistical measure (information content (FABIA))
  - biological context - gene ontology annotations or other literature-based enrichment analysis

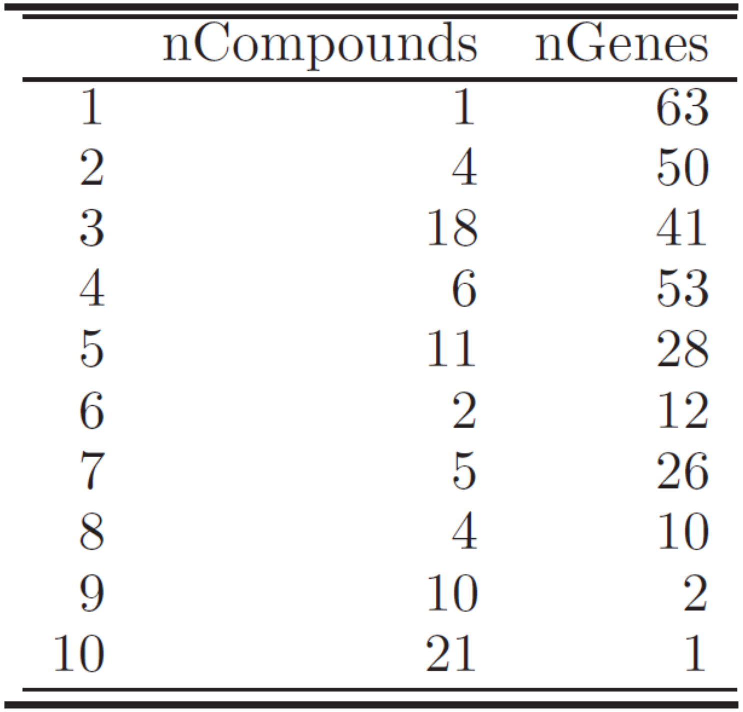
Idea for early drug discovery

- rank based on another source of information, (e.g. the chemical structure, target predictions, HCS, etc)

- investigate whether compounds in a bicluster are also structurally similar

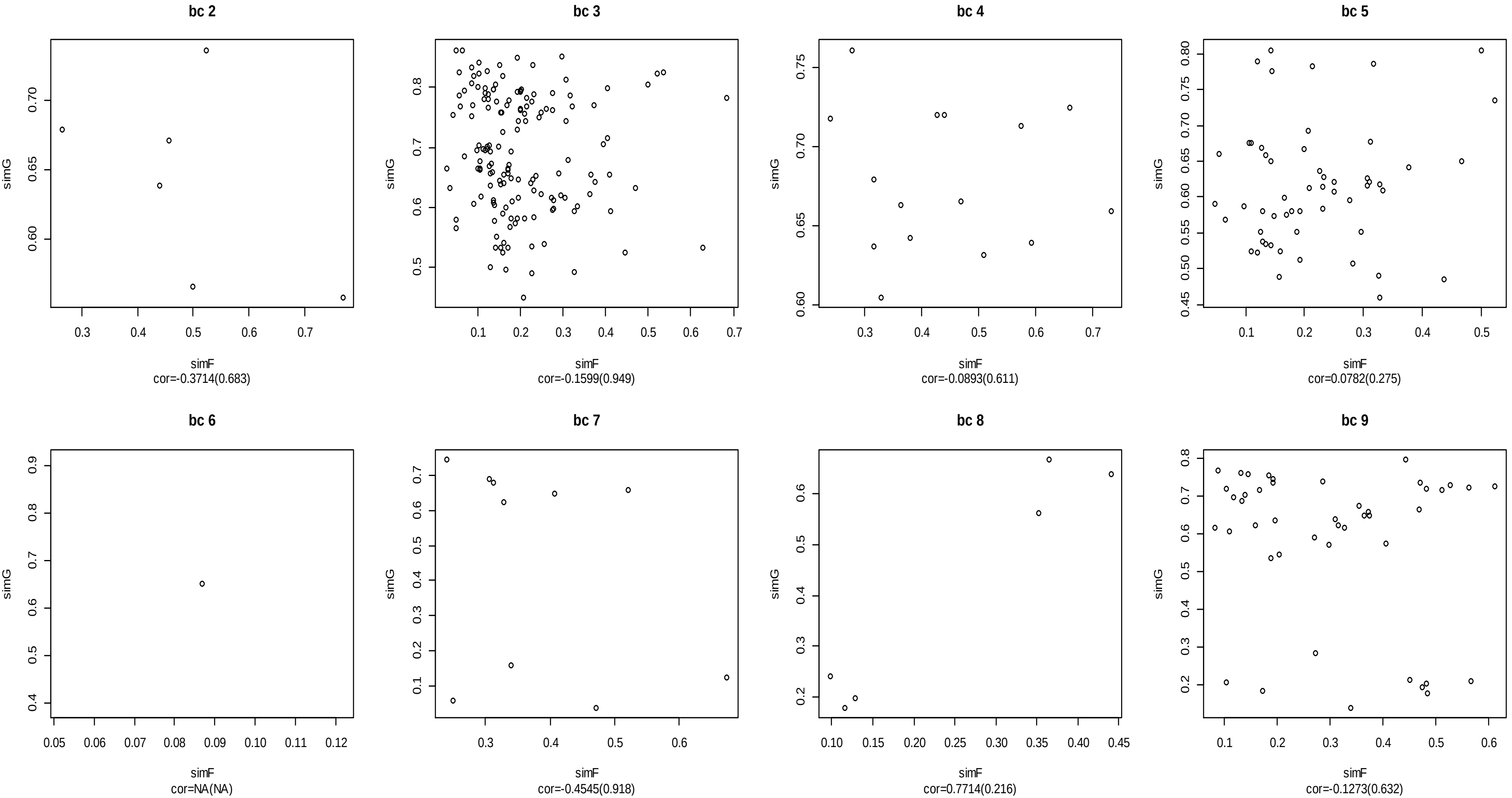
- Similar activity and similar structure → desirable compound set!

Biclustering Results

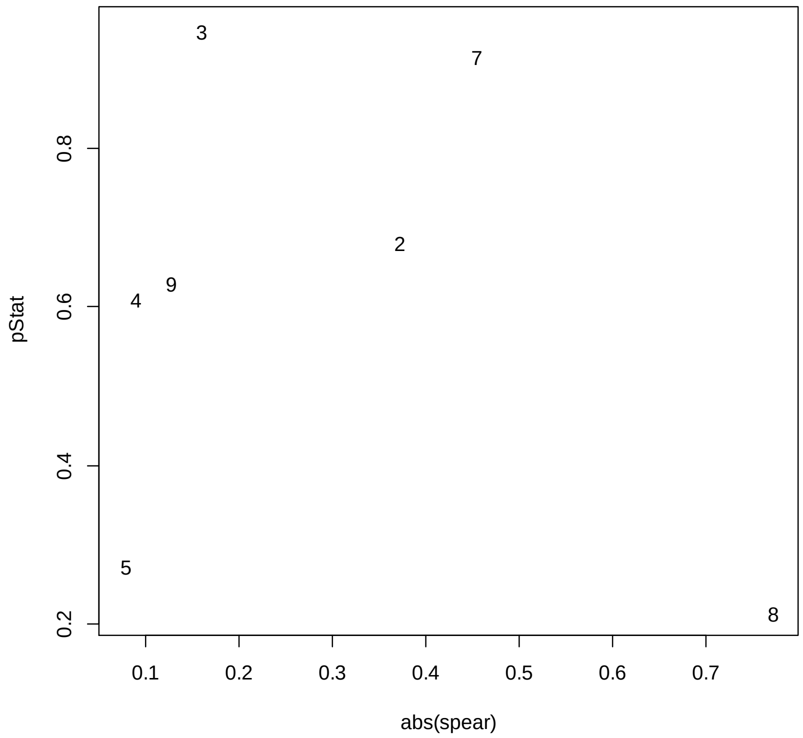


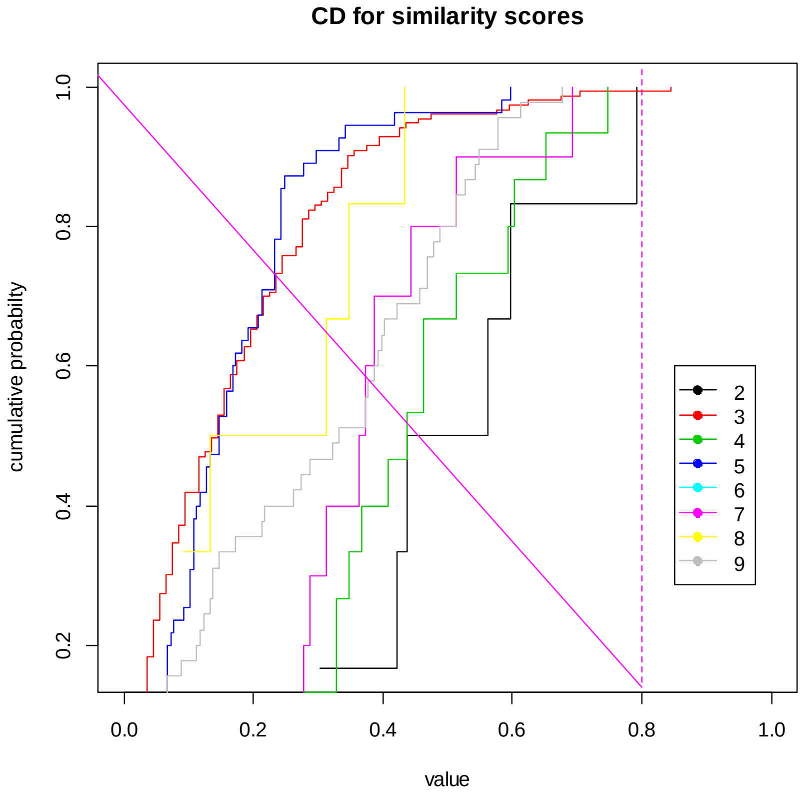
	nCompounds	nGenes
1	1	63
2	4	50
3	18	41
4	6	53
5	11	28
6	2	12
7	5	26
8	4	10
9	10	2
10	21	1

Spearman correlation of similarity scores

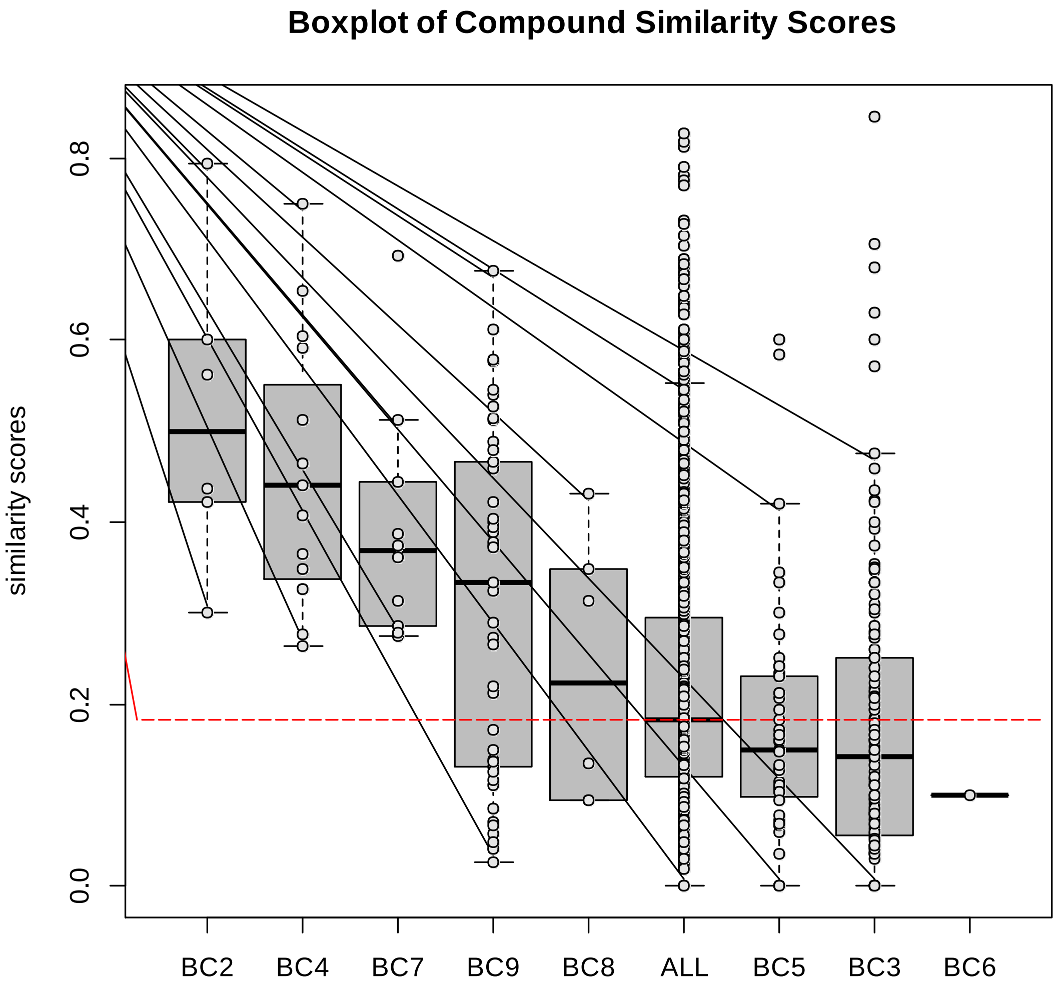


Ranking statistics



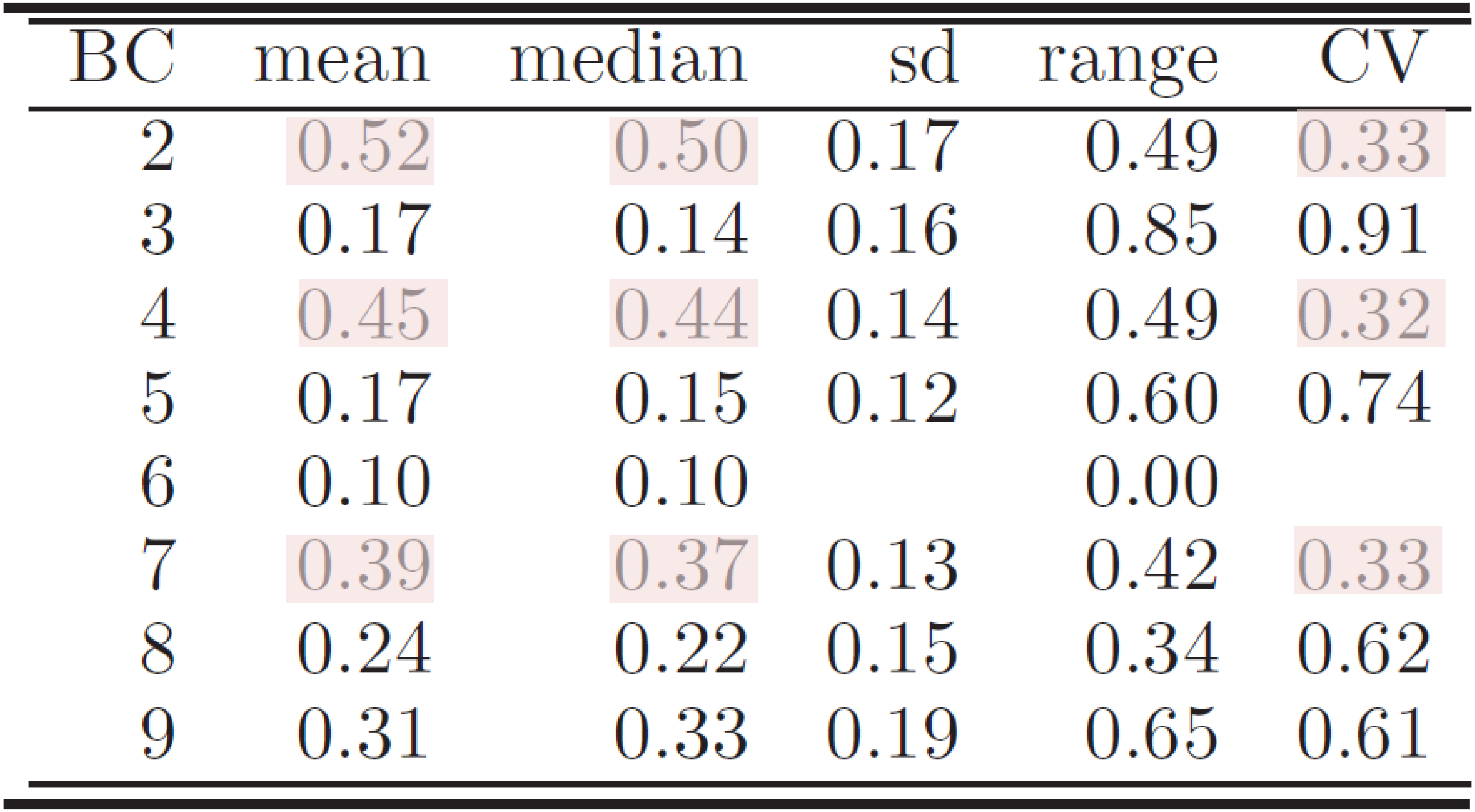

CD for similarity scores

BC Ranking based on median similarity scores (C)



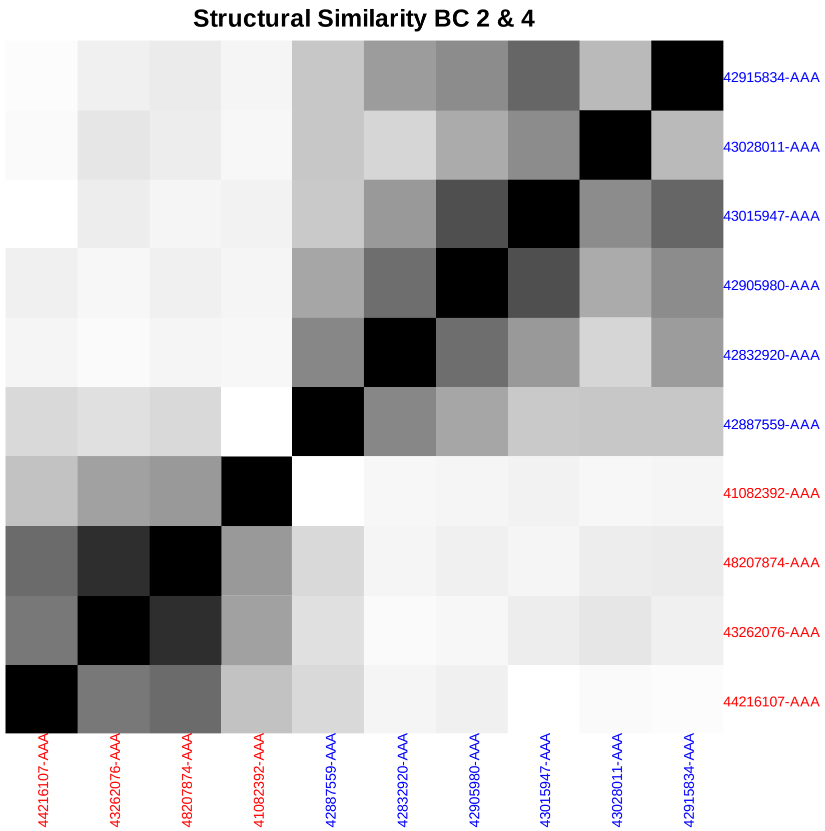
BC2	BC4	BC7	BC9	BC8	ALL	BC5	BC3	BC6
0.50	0.44	0.37	0.33	0.22	0.18	0.15	0.14	0.10

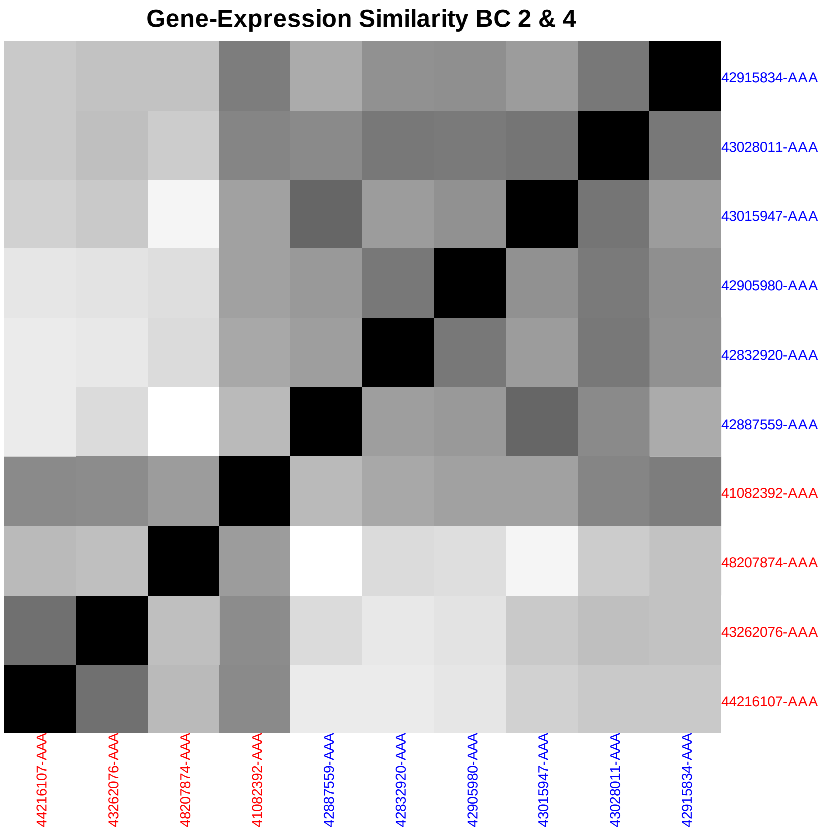
other statistics



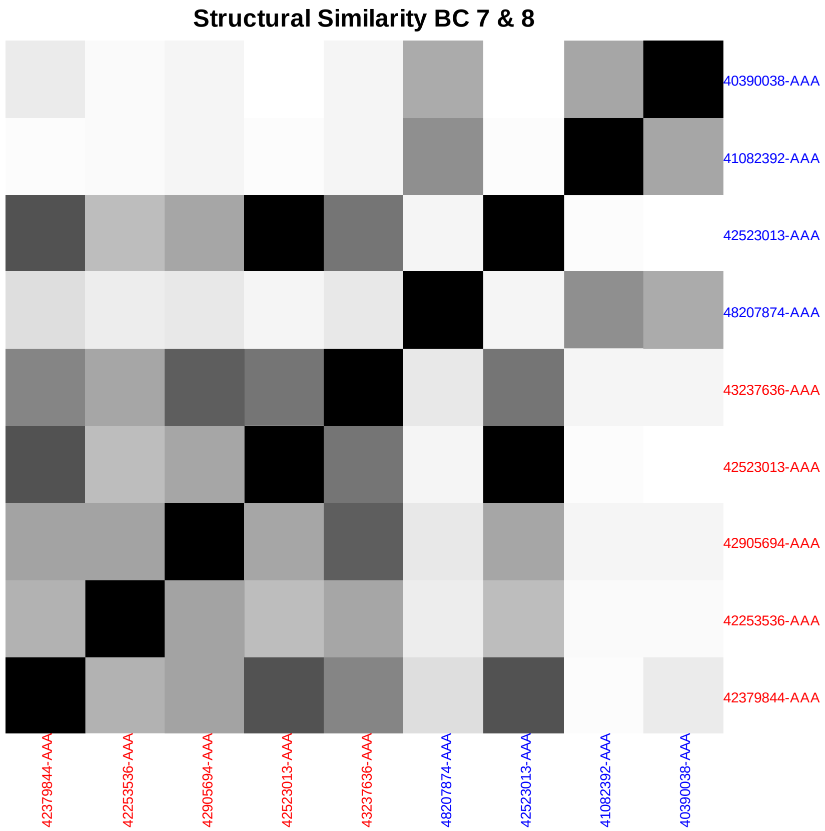
BC	mean	median	sd	range	CV
2	0.52	0.50	0.17	0.49	0.33
3	0.17	0.14	0.16	0.85	0.91
4	0.45	0.44	0.14	0.49	0.32
5	0.17	0.15	0.12	0.60	0.74
6	0.10	0.10		0.00	
7	0.39	0.37	0.13	0.42	0.33
8	0.24	0.22	0.15	0.34	0.62
9	0.31	0.33	0.19	0.65	0.61

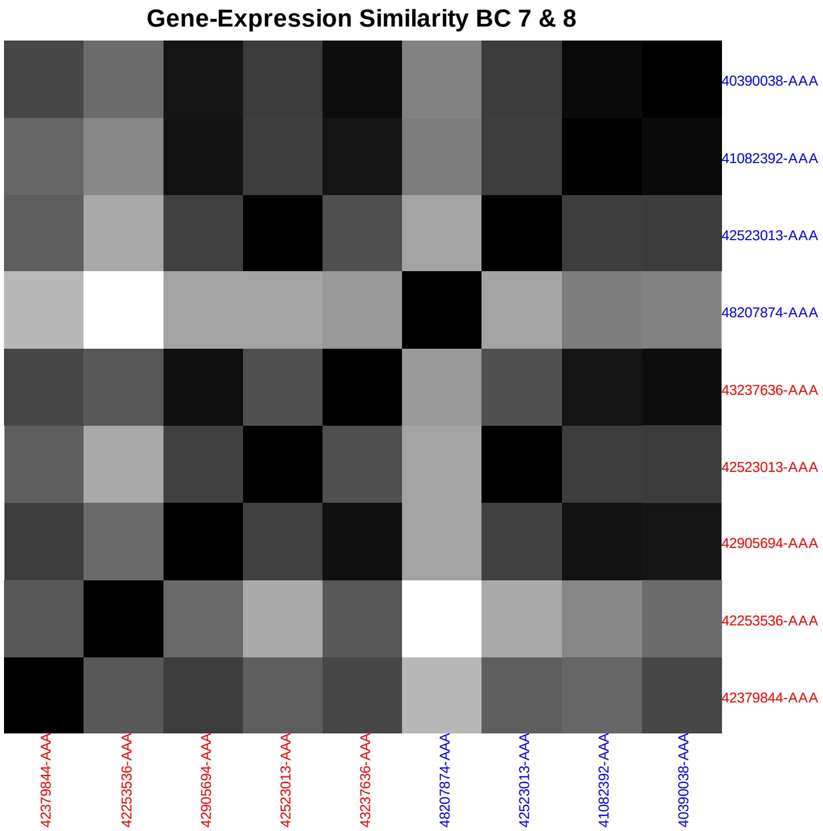
Similarity Scores of BC2 and BC4





Similarity Scores of BC7 and BC8



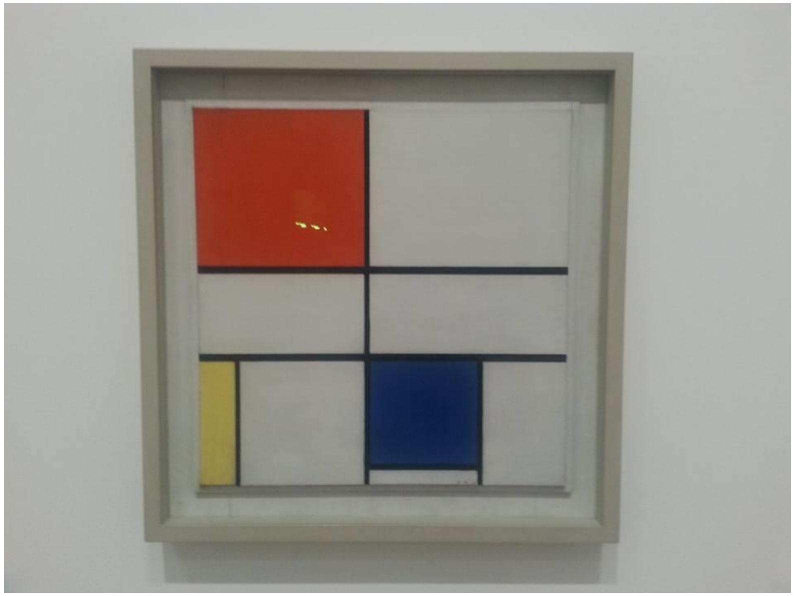


Discussion

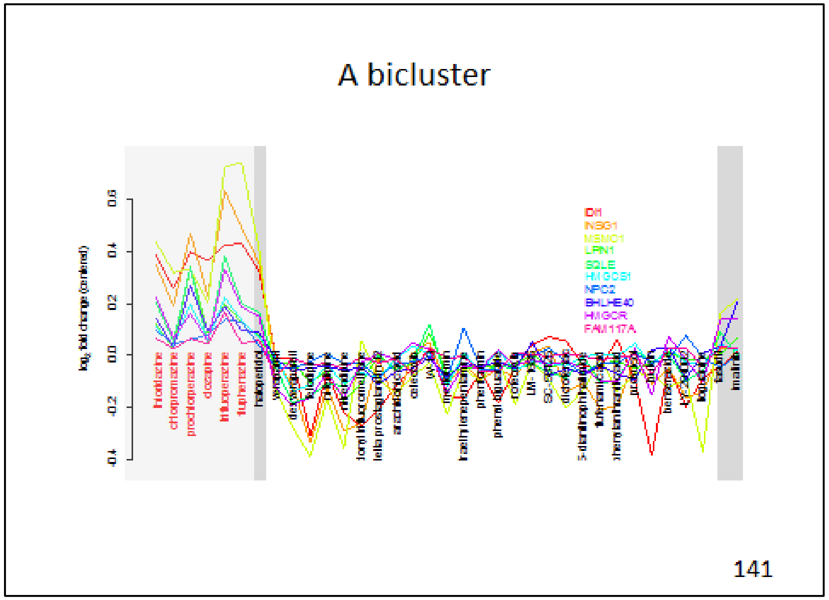
- Not ranking per se but to prioritize more interesting biclusters using extra information available

- Software: *bcRank*

Summary



Piet Mondrian, Tate modern.



A bicluster

141

Perualila et al, 2016

Biclustering: local patterns to understand the big picture.

Many areas of applications.

Many methods.

Software.