Finding Hidden Patterns in Complex Multivariate Data

Ruben Zamar
Deapartment of Statistics
UBC

June 27, 2013

PART I

GROUPING ITEMS THAT SEEM ALIKE

BACKGROUND

- **Taxonomists** pioneered the grouping - or **clustering** - of plants and animals to form species.

BACKGROUND

- **Taxonomists** pioneered the grouping - or **clustering** - of plants and animals to form species.
- They needed consistent procedures (across scientists) to assign similar specimens to the same groups.

BACKGROUND

- **Taxonomists** pioneered the grouping - or **clustering** - of plants and animals to form species.
- They needed consistent procedures (across scientists) to assign similar specimens to the same groups.
- Initially, **clustering** was done **manually.**

BACKGROUND

- **Taxonomists** pioneered the grouping - or **clustering** - of plants and animals to form species.
- They needed consistent procedures (across scientists) to assign similar specimens to the same groups.
- Initially, **clustering** was done **manually.**
- Taxonomists used measurements (**grouping variables)** to help their task.

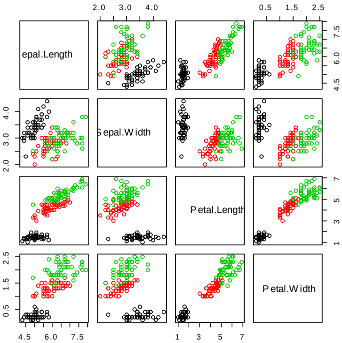
DIFFERENT SUBSPECIES OF IRIS PLANTS

MEASUREMENTS

IRIS DATA

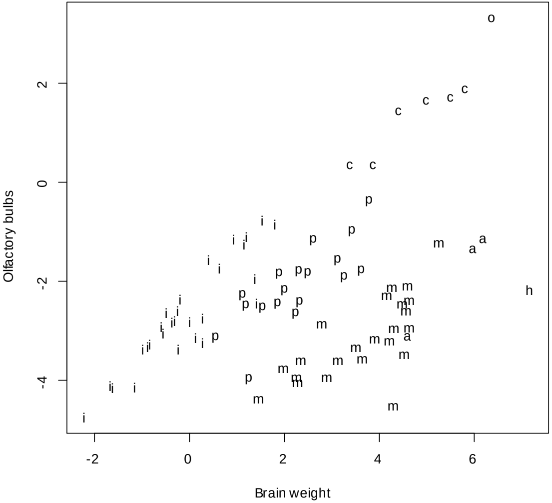
Item	sepal length	sepal width	petal length	petal width
plant 1	5.1	3.5	1.4	0.2
plant 2	4.9	3.0	1.4	0.2
plant 3	5.4	3.9	1.7	0.4
⋮	⋮	⋮	⋮	⋮
plant 150	5.9	3.0	5.1	1.8

FISHER - ANDERSON "IRIS DATA"

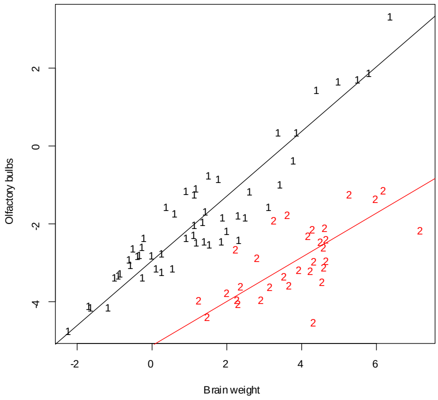


Black = Setosa, Green = Virginica, Red = Versicolor

JERISON (1973) “ALLOMETRY DATA”



JERISON (1973) ALLOMETRY DATA



(Insectivores, Carnivores, Horse, Prosimians), (Apes, Monkeys, Human)

FROM TAXONOMY TO MODERN CLUSTERING

- IN STATISTICS AND COMPUTER SCIENCE, CLUSTERING MEANS
  “AUTOMATIC, COMPUTER AIDED, GROUPING OF SIMILAR ITEMS BASED ON SOME SIMILARITY MEASURE”.

FROM TAXONOMY TO MODERN CLUSTERING

- IN STATISTICS AND COMPUTER SCIENCE, CLUSTERING MEANS
  "AUTOMATIC, COMPUTER AIDED, GROUPING OF SIMILAR ITEMS BASED ON SOME SIMILARITY MEASURE".
- THE **NUMBER** OF CLUSTERS (GROUPS) IS UNKNOWN

FROM TAXONOMY TO MODERN CLUSTERING

- IN STATISTICS AND COMPUTER SCIENCE, CLUSTERING MEANS
  "AUTOMATIC, COMPUTER AIDED, GROUPING OF SIMILAR ITEMS BASED ON SOME SIMILARITY MEASURE".
- THE **NUMBER** OF CLUSTERS (GROUPS) IS UNKNOWN
- THE **RELATIVE SIZE** OF THE CLUSTERS IS UNKNOWN

FROM TAXONOMY TO MODERN CLUSTERING

- IN STATISTICS AND COMPUTER SCIENCE, CLUSTERING MEANS
  "AUTOMATIC, COMPUTER AIDED, GROUPING OF SIMILAR ITEMS BASED ON SOME SIMILARITY MEASURE".
- THE **NUMBER** OF CLUSTERS (GROUPS) IS UNKNOWN
- THE **RELATIVE SIZE** OF THE CLUSTERS IS UNKNOWN
- FINDING ALL OF THAT FROM THE DATA IS A VERY **CHALLENGING STATISTICAL PROBLEM.**

MAIN GOALS OF CLUSTERING

- TO **FIND** AND **NAME** HIDDEN GROUPS OF SIMILAR ITEMS

MAIN GOALS OF CLUSTERING

- TO **FIND** AND **NAME** HIDDEN GROUPS OF SIMILAR ITEMS

- TO **EXPLAIN** AND **INTERPRET** THE GROUPS

MAIN GOALS OF CLUSTERING

- TO **FIND** AND **NAME** HIDDEN GROUPS OF SIMILAR ITEMS

- TO **EXPLAIN** AND **INTERPRET** THE GROUPS

- TO **SUMMARIZE** AND **DISPLAY** THE GROUPS

SOME EXAMPLES OF CLUSTERING APPLICATIONS

- GROUPING DIFFERENT CANCER TUMORS BASED ON GENE EXPRESSION DATA

SOME EXAMPLES OF CLUSTERING APPLICATIONS

- GROUPING DIFFERENT CANCER TUMORS BASED ON GENE EXPRESSION DATA

- FORMING SOCIAL CLASSES BASED ON SOCIO-ECONOMICAL FEATURES

SOME EXAMPLES OF CLUSTERING APPLICATIONS

- GROUPING DIFFERENT CANCER TUMORS BASED ON GENE EXPRESSION DATA

- FORMING SOCIAL CLASSES BASED ON SOCIO-ECONOMICAL FEATURES

- FINDING SIMILAR TYPES OF CUSTOMERS BASED ON PURCHASING PATTERNS

The Problem (In Math Notation)

- **d VARIABLES (FEATURES) ARE MEASURED IN n ITEMS**

The Problem (In Math Notation)

- **d VARIABLES (FEATURES) ARE MEASURED IN n ITEMS**
- **DATA TABLE**

Item	X_1	X_2	$\cdots$	X_d
1	x_{11}	x_{12}	$\cdots$	x_{1d}
2	x_{21}	x_{22}	$\cdots$	x_{2d}
3	x_{31}	x_{32}	$\cdots$	x_{3d}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nd}

The Problem (In Math Notation)

- **d VARIABLES (FEATURES) ARE MEASURED IN n ITEMS**
- **DATA TABLE**

Item	X_1	X_2	$\cdots$	X_d
1	x_{11}	x_{12}	$\cdots$	x_{1d}
2	x_{21}	x_{22}	$\cdots$	x_{2d}
3	x_{31}	x_{32}	$\cdots$	x_{3d}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nd}

- **FIND PATTERNS IN THE NUMBERS TO IDENTIFY THE GROUPS**

- **DEVELOP/IMPLEMENT ALGORITHMS TO FIND PATTERNS IN THE OBSERVATIONS**

CLUSTER ALGORITHMS

- **DEVELOP/IMPLEMENT ALGORITHMS TO FIND PATTERNS IN THE OBSERVATIONS**

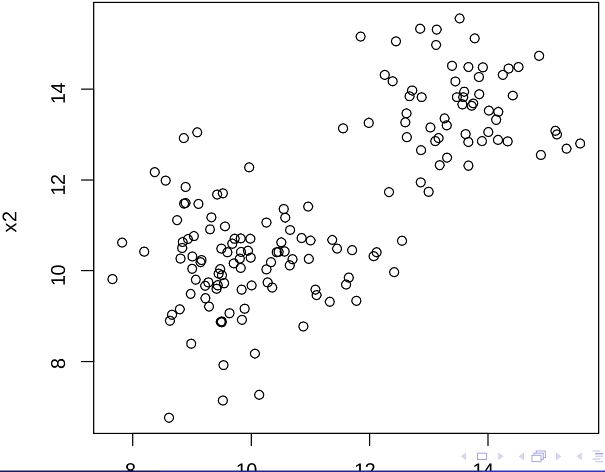
- **IDENTIFY GROUPS OF ITEMS THAT EXHIBIT SIMILAR PATTERNS**

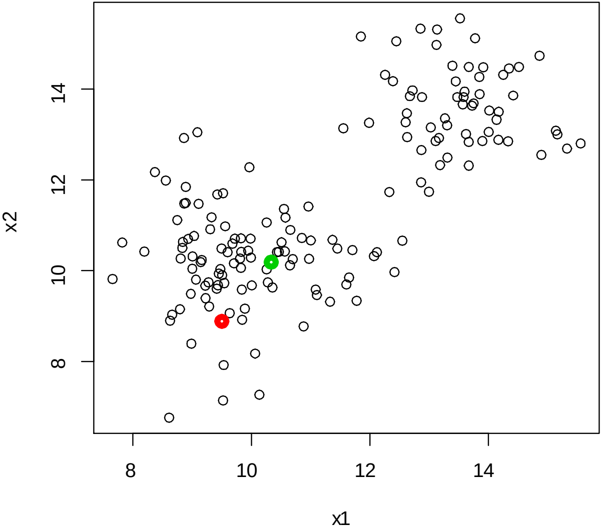
MEASUREMENTS

IRIS DATA

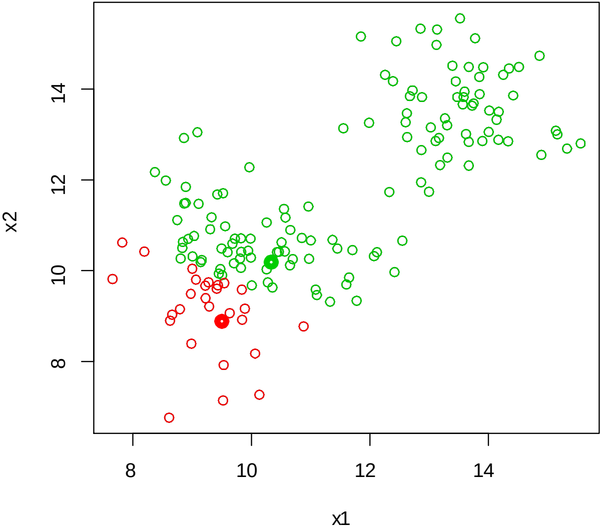
Item	sepal length	sepal width	petal length	petal width
plant 1	5.1	3.5	1.4	0.2
plant 2	4.9	3.0	1.4	0.2
plant 3	5.4	3.9	1.7	0.4
⋮	⋮	⋮	⋮	⋮
plant 150	5.9	3.0	5.1	1.8

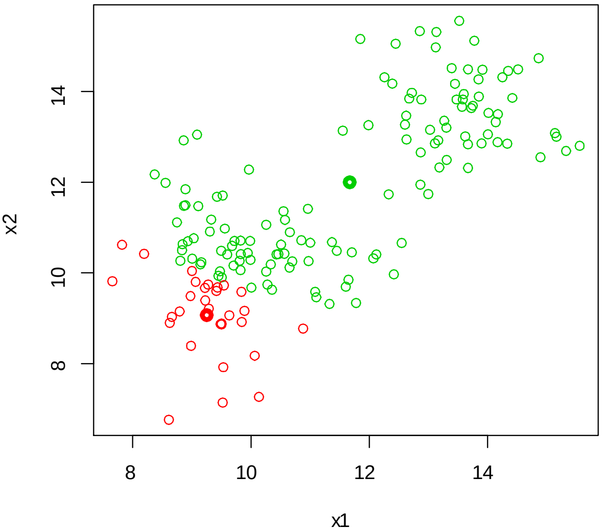
SIMPLE NUMERICAL ILLUSTRATION

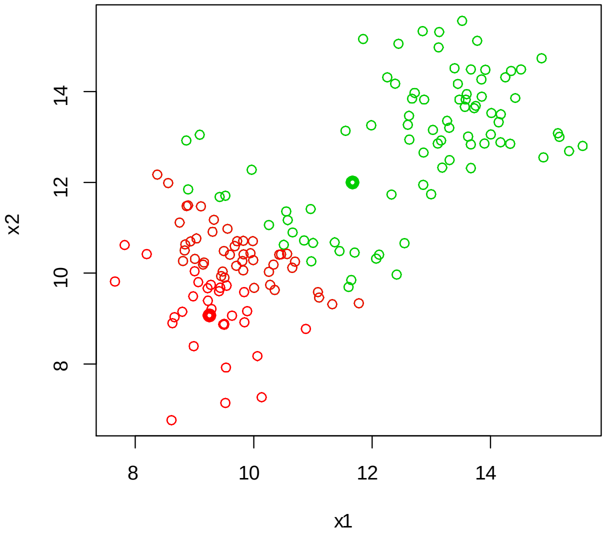




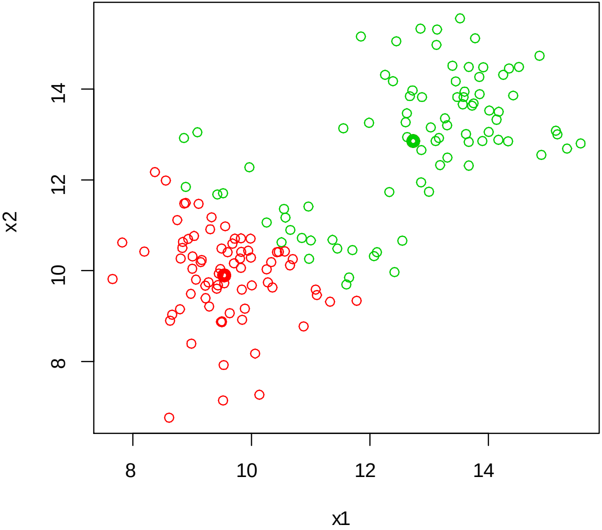
x2
x1
8
10
12
14
8
10
12
14

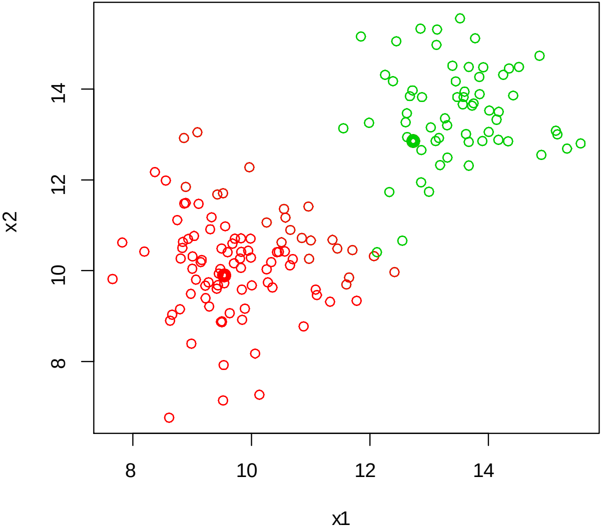

8
10
12
14
x1
8
10
12
14
x2


x2
x1
8
10
12
14
8
10
12
14

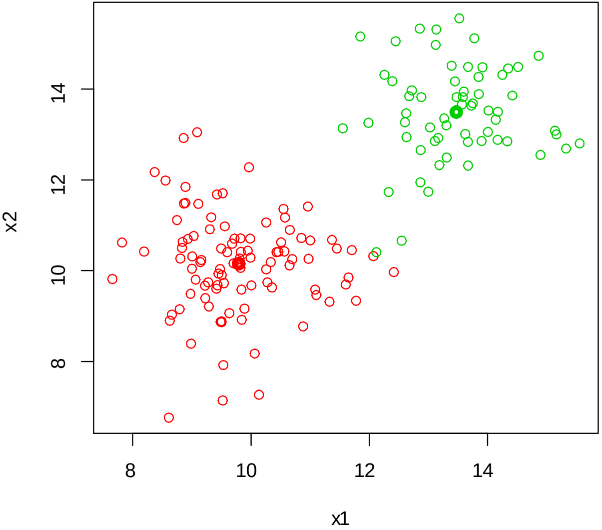


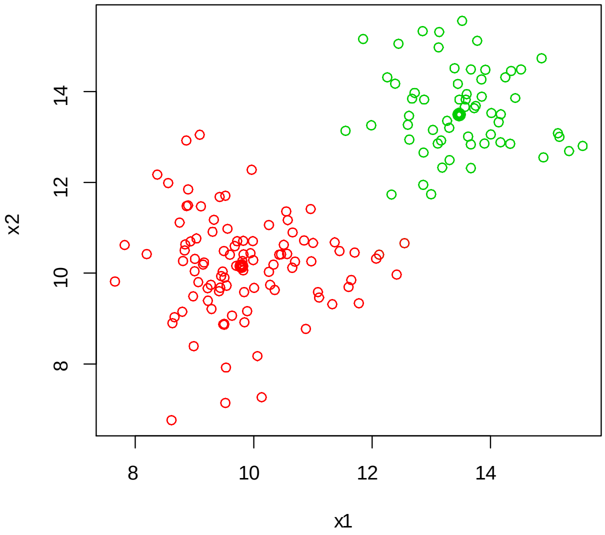
x2
8
10
12
14
x1


x2
x1
8
10
12
14
8
10
12
14

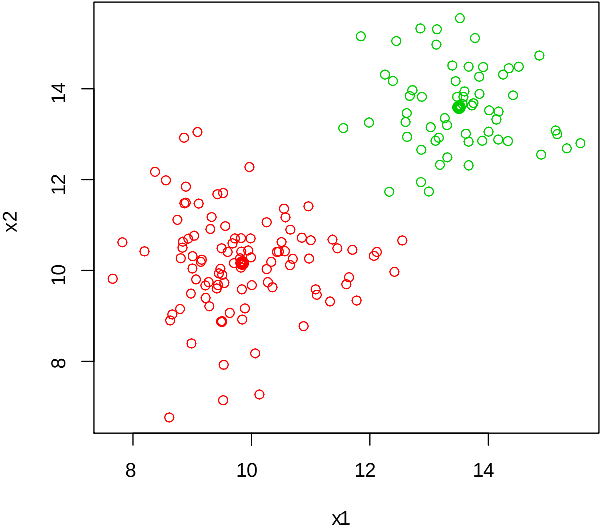


x2
14
12
10
8
8
10
12
14
x1

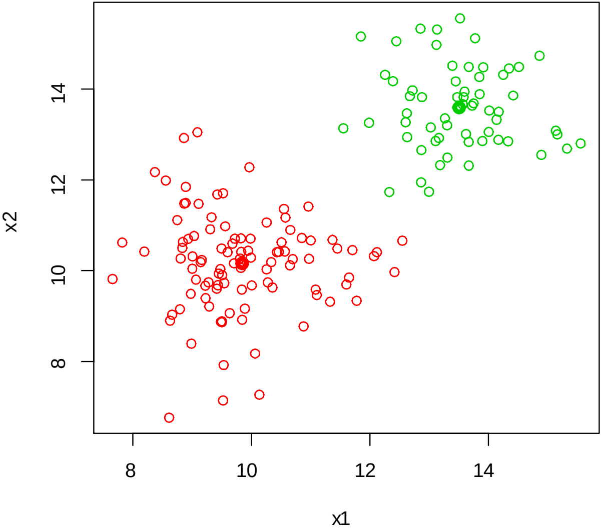

x2
14
12
10
8
8
10
12
14
x1



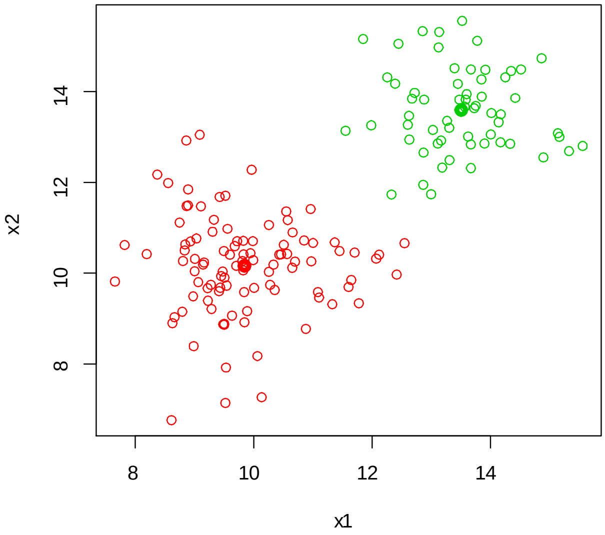
14
12
10
8
x2
8
10
12
14
x1

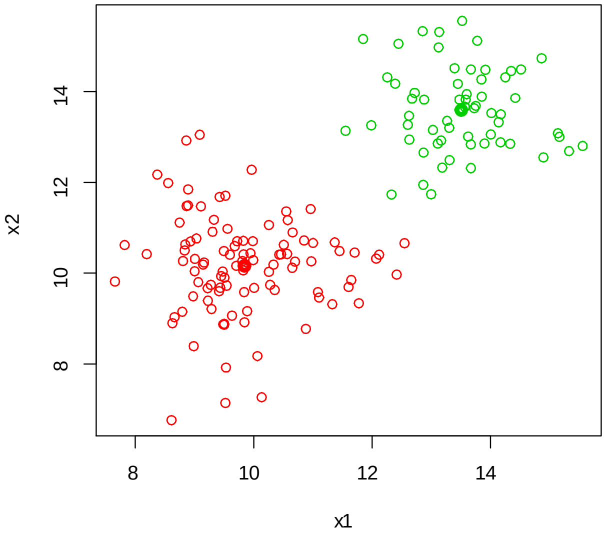


x2
x1
8
10
12
14



x2
14
12
10
8
8
10
12
14
x1


x2
8
10
12
14
x1
8
10
12
14


x2
x1
8
10
12
14
8
10
12
14

- **CENTROID BASED CLUSTER**

DIFFERENT APPROACHES TO CLUSTERING

- **CENTROID BASED CLUSTER**
- **PROBABILITY MODEL BASED CLUSTER**

DIFFERENT APPROACHES TO CLUSTERING

- **CENTROID BASED CLUSTER**
- **PROBABILITY MODEL BASED CLUSTER**
- **DISTANCE BASED CLUSTER**

DIFFERENT APPROACHES TO CLUSTERING

- **CENTROID BASED CLUSTER**
- **PROBABILITY MODEL BASED CLUSTER**
- **DISTANCE BASED CLUSTER**
- **POINT MIGRATING CLUSTER (PEAK HUNTING)**

DIFFERENT APPROACHES TO CLUSTERING

- **CENTROID BASED CLUSTER**
- **PROBABILITY MODEL BASED CLUSTER**
- **DISTANCE BASED CLUSTER**
- **POINT MIGRATING CLUSTER (PEAK HUNTING)**
- **SPARSE CLUSTER**

CENTROID BASED CLUSTER

- MINIMIZE A LOSS FUNCTION

$$
\begin{aligned}
J\left(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\right) &= \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \mathbf{x}_i - \mathbf{t}_j \right\|^2, \quad \mathbf{t}_j = \frac{1}{n_k} \sum_{i \in \mathcal{C}_j} \mathbf{x}_i \\
n_k &= \text{number of items in } \mathcal{C}_j = \#\mathcal{C}_j
\end{aligned}
$$

CENTROID BASED CLUSTER

- MINIMIZE A LOSS FUNCTION

$$
\begin{aligned}
J\left(\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\right) &= \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \mathbf{x}_i - \mathbf{t}_j \right\|^2, \quad \mathbf{t}_j = \frac{1}{n_k} \sum_{i \in \mathcal{C}_j} \mathbf{x}_i \\
n_k &= \text{number of items in } \mathcal{C}_j = \#\mathcal{C}_j
\end{aligned}
$$

- SIMILAR (IN SPIRIT) TO LS-REGRESSION

CENTROID BASED CLUSTER

- MINIMIZE A LOSS FUNCTION

$$J\left(\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\right) = \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \mathbf{x}_i - \mathbf{t}_j \right\|^2, \quad \mathbf{t}_j = \frac{1}{n_k} \sum_{i \in \mathcal{C}_j} \mathbf{x}_i$$

$$n_k = \text{number of items in } \mathcal{C}_j = \#\mathcal{C}_j$$

- SIMILAR (IN SPIRIT) TO LS-REGRESSION
- EXAMPLE: **PACKAGE** *kmeans* **IN R**

CENTROID BASED CLUSTER

- **Van Aelst, Wang, Zamar and Zhu (2006) (CSD)**

CENTROID BASED CLUSTER

- **Van Aelst, Wang, Zamar and Zhu (2006) (CSD)**
  - LINEAR GROUPING USING ORTHOGONAL REGRESSION

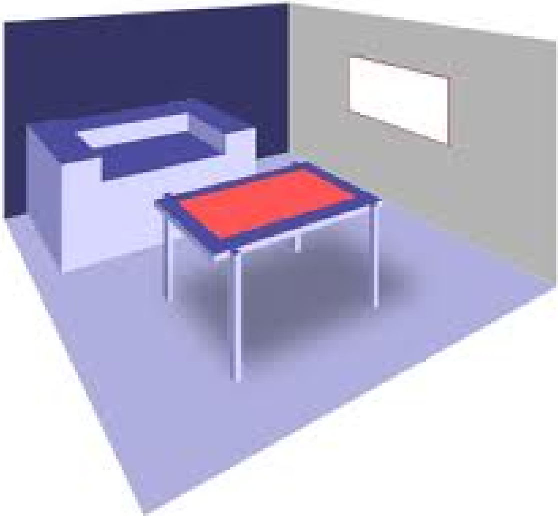
CENTROID BASED CLUSTER

- **Van Aelst, Wang, Zamar and Zhu (2006) (CSD)**
  - LINEAR GROUPING USING ORTHOGONAL REGRESSION
  - FIND GROUPS OF POINTS **CLUSTERED AROUND LINEAR VARIATIES**

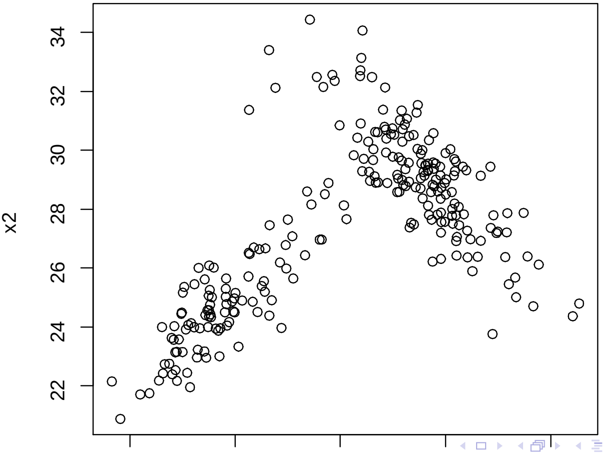
CENTROID BASED CLUSTER

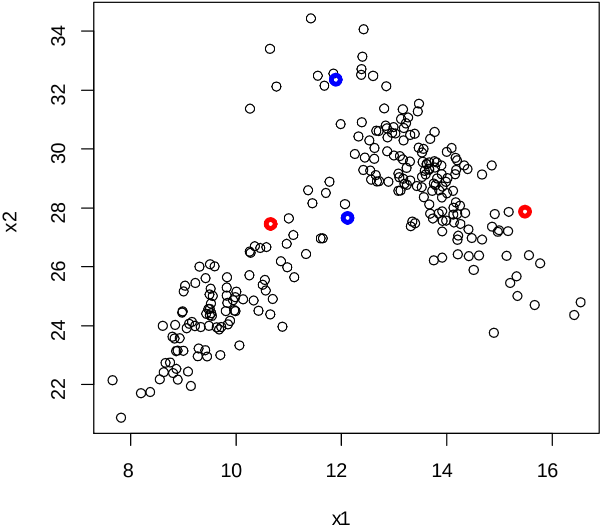
- **Van Aelst, Wang, Zamar and Zhu (2006) (CSD)**
  - LINEAR GROUPING USING ORTHOGONAL REGRESSION
  - FIND GROUPS OF POINTS **CLUSTERED AROUND LINEAR VARIATIES**
  - **EXAMPLE:** POINTS CLUSTERED AROUND **CENTROIDS, LINES** AND **PLANES** IN HIGHER DIMENSIONAL SPACES

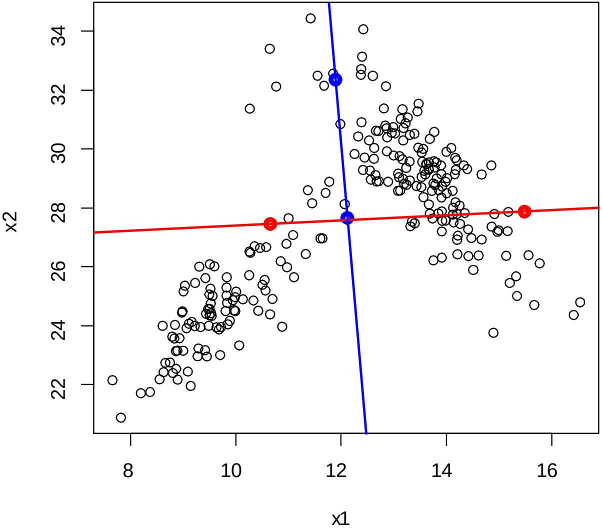
LINES AND PLANES IN 3 DIMENSIONAL SPACES (COMPUTER VISION)

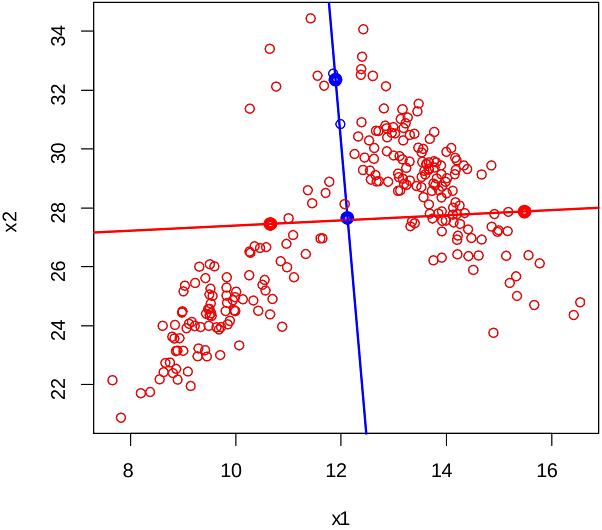


EXAMPLE: CLUSTER OF POINTS AROUND TWO LINES

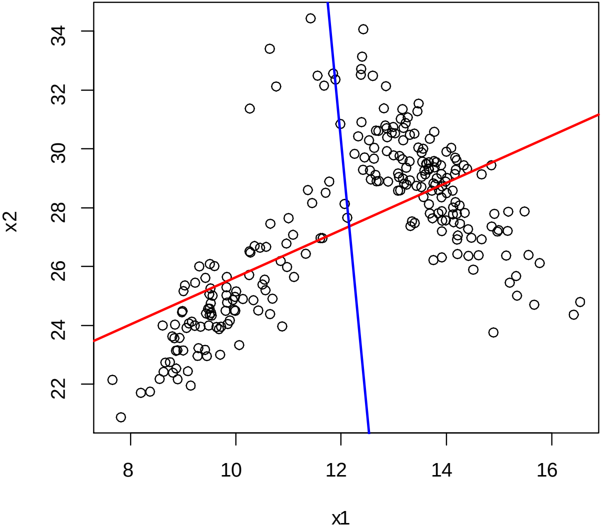



x2
x1
8
10
12
14
16
22
24
26
28
30
32
34

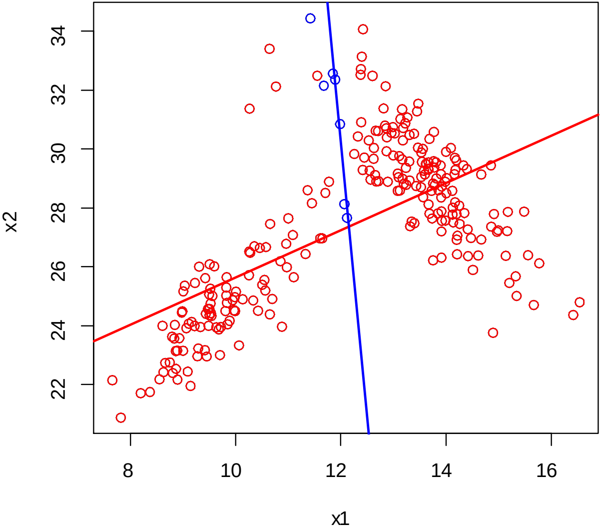

34
32
30
28
26
24
22
x2
8
10
12
14
16
x1

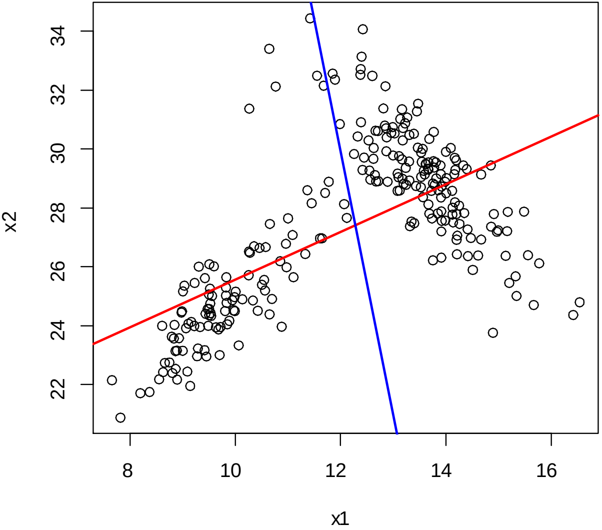


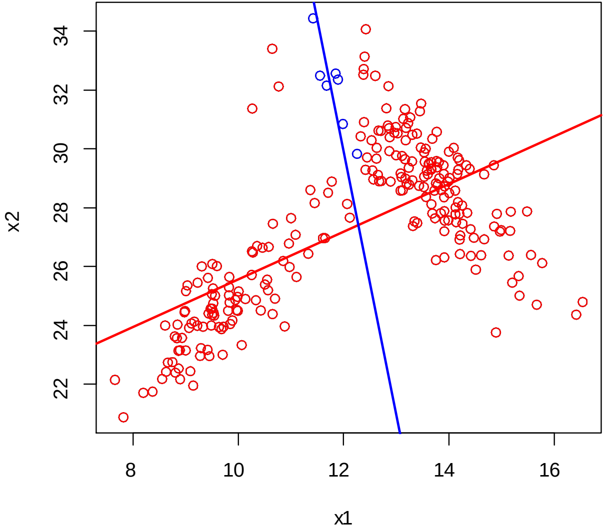
34
32
30
28
26
24
22
x2
8
10
12
14
16
x1

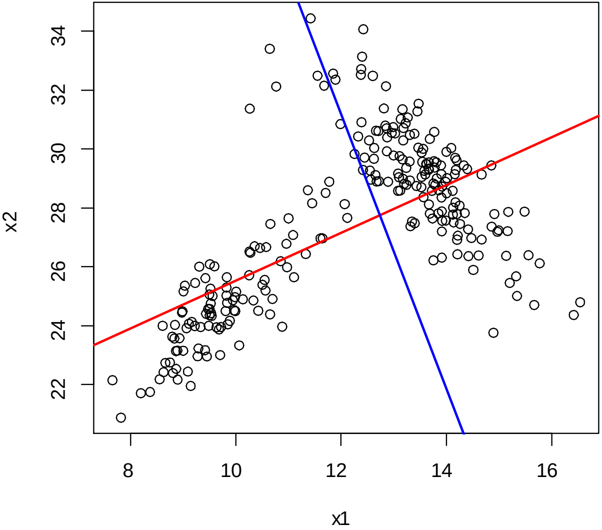


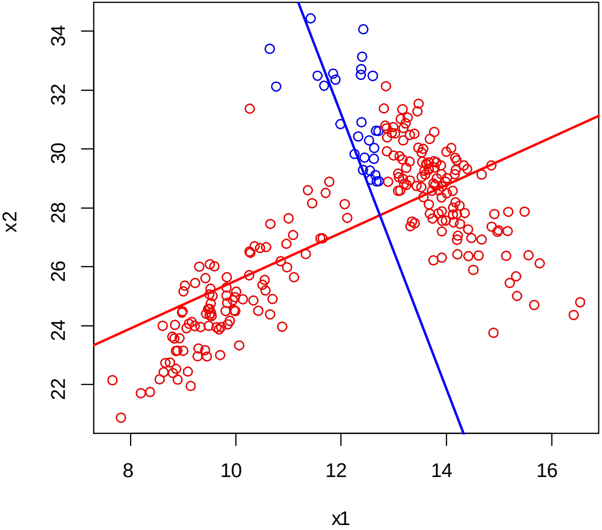
x2
x1
22
24
26
28
30
32
34
8
10
12
14
16

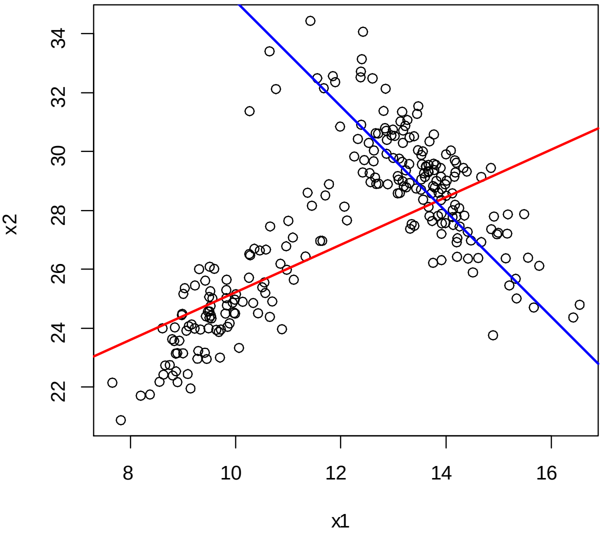

x2
x1
22
24
26
28
30
32
34
8
10
12
14
16


x2
x1
8
10
12
14
16
22
24
26
28
30
32
34

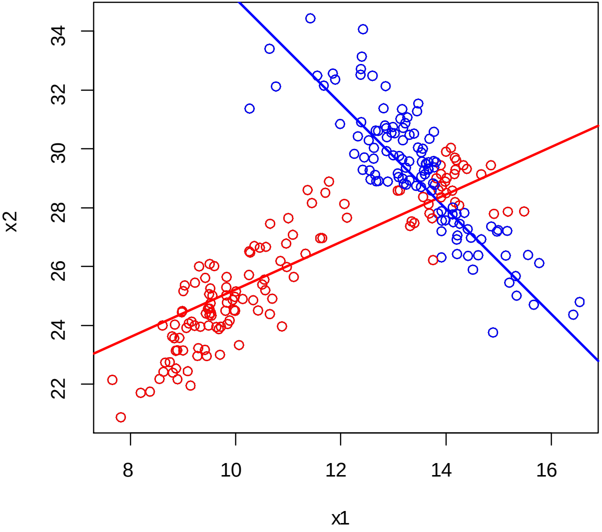

x2
x1
22
24
26
28
30
32
34
8
10
12
14
16

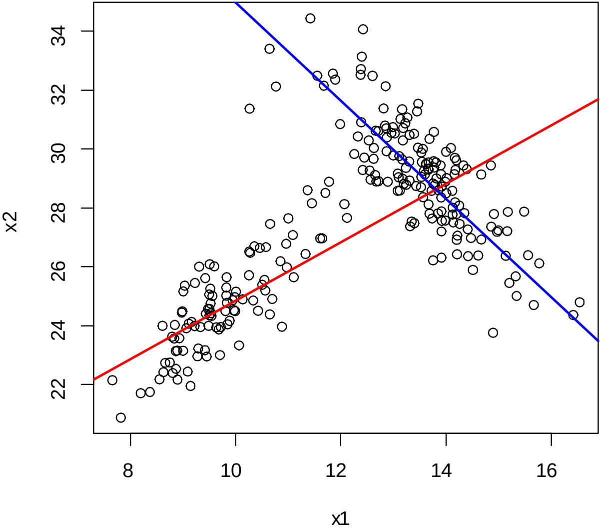

x2
34
32
30
28
26
24
22
8
10
12
14
16
x1


34
32
30
28
26
24
22
x2
8
10
12
14
16
x1

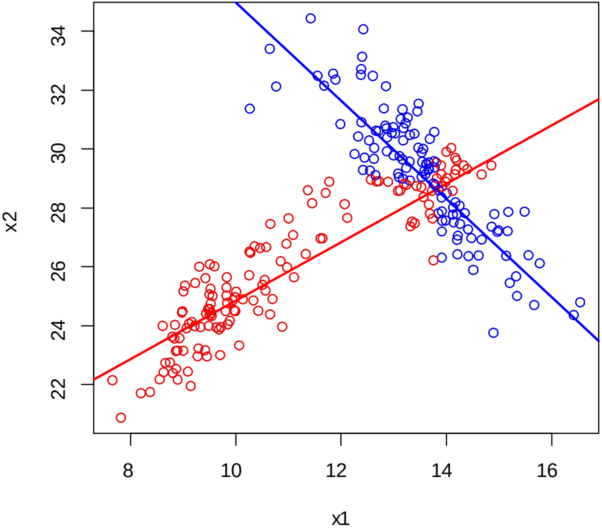


x2
x1
8
10
12
14
16
22
24
26
28
30
32
34

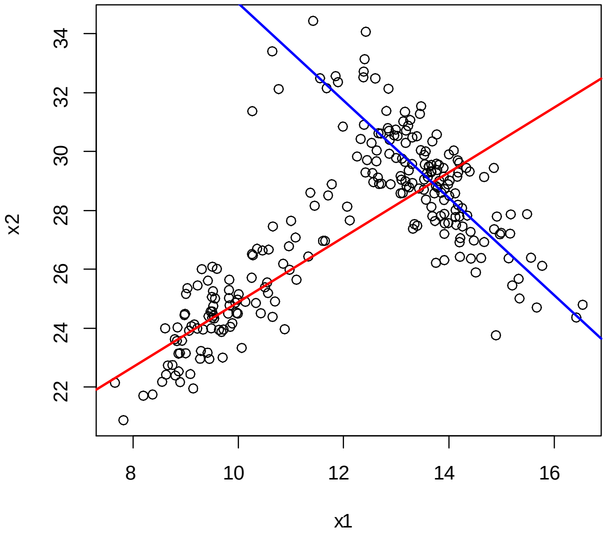

x2
22
24
26
28
30
32
34
x1
8
10
12
14
16

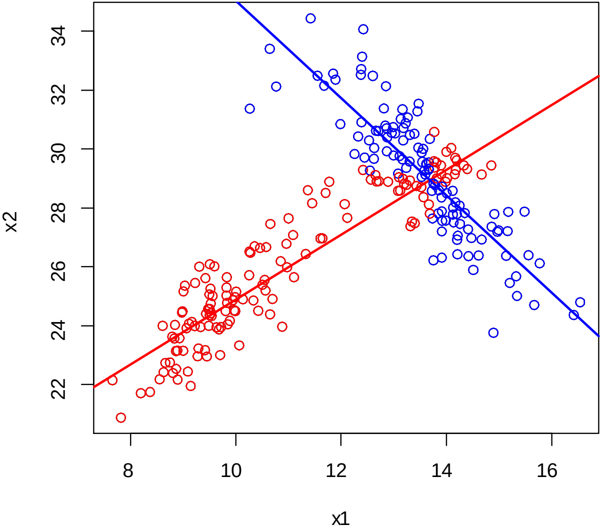


x2
x1
22
24
26
28
30
32
34
8
10
12
14
16

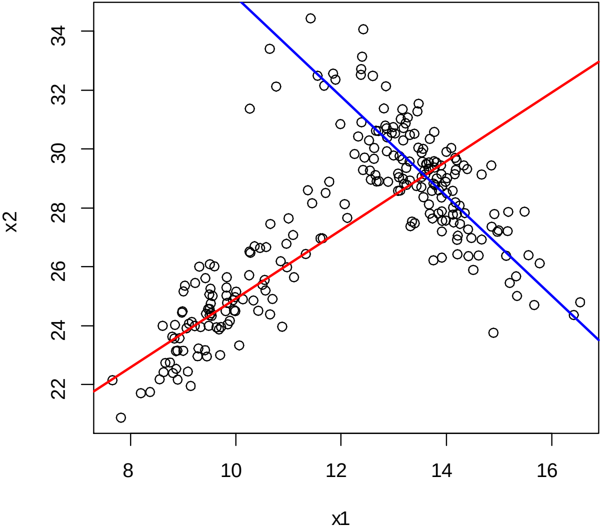


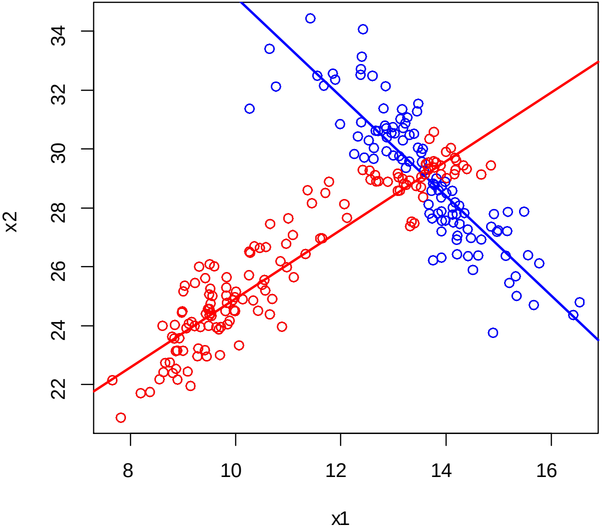
x2
x1
8
10
12
14
16
22
24
26
28
30
32
34

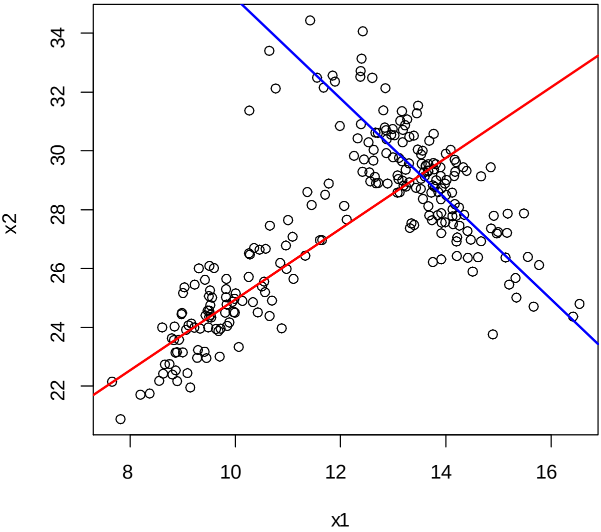

x2
22
24
26
28
30
32
34
8
10
12
14
16
x1



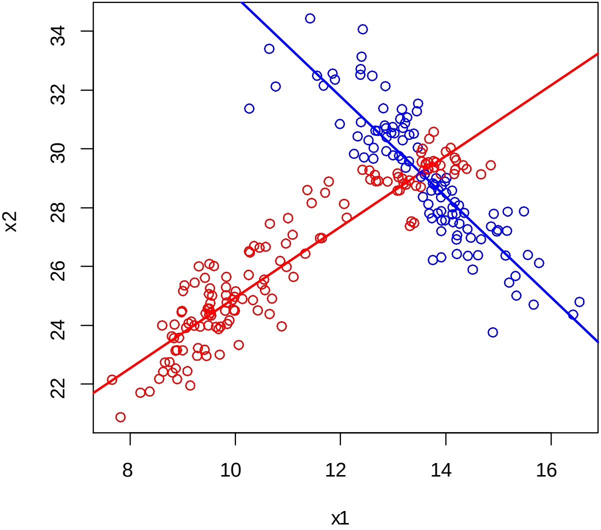
x2
x1
22
24
26
28
30
32
34
8
10
12
14
16


34
32
30
28
26
24
22
x2
8
10
12
14
16
x1


x2
34
32
30
28
26
24
22
x1
8
10
12
14
16



x2
34
32
30
28
26
24
22
8
10
12
14
16
x1


x2
34
32
30
28
26
24
22
8
10
12
14
16
x1

CENTROID BASED CLUSTER

- **Garcia-Escudero, Gordaliza, San Martin, Van Aelst, and Zamar(2009) (JRSS)**

CENTROID BASED CLUSTER

- **Garcia-Escudero, Gordaliza, San Martin, Van Aelst, and Zamar(2009) (JRSS)**
    - ROBUST EXTENSION OF LINEAR CLUSTERING USING "IMPARTIAL TRIMMING"

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{k} \left[\alpha_i f_i(\mathbf{x})\right]^{\delta_i}, \quad \delta_i = 0, 1, \quad 0 < \alpha_i < 1 \\
\sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1
\end{aligned}
$$

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^{k} \left[\alpha_i f_i(\mathbf{x})\right]^{\delta_i}, \quad \delta_i = 0, 1, \quad 0 < \alpha_i < 1 \\ \sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1 \end{aligned}$$

  - MAXIMIZE THE LIKELIHOOD FUNCTION

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^{k} [\alpha_i f_i(\mathbf{x})]^{\delta_i}, \quad \delta_i = 0,1, \quad 0 < \alpha_i < 1 \\ \sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1 \end{aligned}$$

  - MAXIMIZE THE LIKELIHOOD FUNCTION
    - EXPECTATION-MINIMIZATION (EM) ALGORITHMS

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{k} \left[\alpha_i f_i(\mathbf{x})\right]^{\delta_i}, \quad \delta_i = 0, 1, \quad 0 < \alpha_i < 1 \\
\sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1
\end{aligned}
$$

  - MAXIMIZE THE LIKELIHOOD FUNCTION
    - EXPECTATION-MINIMIZATION (EM) ALGORITHMS
    - **EXAMPLE:** PACKAGE *mclust* IN R

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{k} [\alpha_i f_i(\mathbf{x})]^{\delta_i}, \quad \delta_i = 0, 1, \quad 0 < \alpha_i < 1 \\
\sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1
\end{aligned}
$$

  - MAXIMIZE THE LIKELIHOOD FUNCTION
    - EXPECTATION-MINIMIZATION (EM) ALGORITHMS
    - **EXAMPLE:** PACKAGE *mclust* IN R
- Yan, Welch, and Zamar (2010) **(CJS)**

DIFFERENT APPROACHES TO CLUSTERING

- MODEL BASED CLUSTERING
  - MODEL THE CLUSTERS USING A "MIXTURE" PROBABILITY DENSITY

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{k} \left[\alpha_i f_i(\mathbf{x})\right]^{\delta_i}, \quad \delta_i = 0, 1, \quad 0 < \alpha_i < 1 \\
\sum_{i=1}^{k} \alpha_i &= \sum_{i=1}^{k} \delta_i = 1
\end{aligned}
$$

  - MAXIMIZE THE LIKELIHOOD FUNCTION
    - EXPECTATION-MINIMIZATION (EM) ALGORITHMS
    - **EXAMPLE:** PACKAGE *mclust* IN R
- Yan, Welch, and Zamar (2010) **(CJS)**
  - MODEL–BASED **LINEAR CLUSTERING**

DISTANCE BASED CLUSTERING

- USE THE NOTION OF "DISTANCE" BETWEEN TWO GROUPS OF OBJECTS

- USE THE NOTION OF “DISTANCE” BETWEEN TWO GROUPS OF OBJECTS
    - MINIMUM, MAXIMUM OR AVERAGE DISTANCE

DISTANCE BASED CLUSTERING

- USE THE NOTION OF "DISTANCE" BETWEEN TWO GROUPS OF OBJECTS
  - MINIMUM, MAXIMUM OR AVERAGE DISTANCE
  - AGLOMERATIVE OR DIVISIVE

DISTANCE BASED CLUSTERING

- USE THE NOTION OF "DISTANCE" BETWEEN TWO GROUPS OF OBJECTS
  - MINIMUM, MAXIMUM OR AVERAGE DISTANCE
  - AGLOMERATIVE OR DIVISIVE
- EXAMPLE PACKAGE *hclust* IN R

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
  - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
    - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS
    - PACKAGE *clues* IN R

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
  - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS
  - PACKAGE *clues* IN R
- **Pena, Viladomat, and Zamar (2012). (SADM).**

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
  - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS
  - PACKAGE *clues* IN R
- **Pena, Viladomat, and Zamar (2012). (SADM)**.
  - NEAREST-NEIGHBORS MEDIAN CLUSTER ALGORITHM

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
  - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS
  - PACKAGE *clues* IN R
- **Pena, Viladomat, and Zamar (2012). (SADM)**.
  - NEAREST-NEIGHBORS MEDIAN CLUSTER ALGORITHM
  - IMPROVEMENT OVER *clues*

MIGRATING POINTS (BUMP HUNTING)

- ITERATIVELY, COMPUTE LOCAL AVERAGES AND MIGRATE POINTS TOWARD THEM
- **Wang, Qiu and Zamar (2007) (CSDA)**
  - MIGRATES POINTS TOWARD THEIR LOCAL MEDIANS
  - PACKAGE *clues* IN R
- **Pena, Viladomat, and Zamar (2012). (SADM)**.
  - NEAREST-NEIGHBORS MEDIAN CLUSTER ALGORITHM
  - IMPROVEMENT OVER *clues*
  - ALGORITHM *"ATTACTORS"* AVAILABLE FOR MATHLAB

PART II

PART II

THE NEEDLE

IN THE HAYSTACK

DRUG DISCOVERY

- **BIOLOGICAL TARGET:** TO CURE OR PALLIATE A MEDICAL CONDITION

DRUG DISCOVERY

- **BIOLOGICAL TARGET:** TO CURE OR PALLIATE A MEDICAL CONDITION
- EXAMPLES:

GAUCHER'S DISEASE

CHRONIC IMFLAMATION

HIV

LUNG CANCER CELLS

THE HAYSTACK

- SOME STUDIES BEGIN WITH 3000 TO 5000 **"CANDIDATE COMPOUNDS"**

THE HAYSTACK

- SOME STUDIES BEGIN WITH 3000 TO 5000 **"CANDIDATE COMPOUNDS"**
- THESE COMPOUNDS ARE EXAMINED IN BIOLOGICAL ASSAYS

THE HAYSTACK

- SOME STUDIES BEGIN WITH 3000 TO 5000 **"CANDIDATE COMPOUNDS"**
- THESE COMPOUNDS ARE EXAMINED IN BIOLOGICAL ASSAYS
- BIOLOGICAL ASSAYS ARE **EXPENSIVE** AND **TIME CONSUMING**

THE NEEDLES

- A SMALL FRACTION OF THE CONSIDERED COMPOUNDS ARE ACTIVE (AND DESERVE FURTHER INVESTIGATION)

THE NEEDLES

- A SMALL FRACTION OF THE CONSIDERED COMPOUNDS ARE ACTIVE (AND DESERVE FURTHER INVESTIGATION)
- **SEARCHING FOR THE GOLDEN NEEDLE**

THE NEEDLES

- A SMALL FRACTION OF THE CONSIDERED COMPOUNDS ARE ACTIVE (AND DESERVE FURTHER INVESTIGATION)
- **SEARCHING FOR THE GOLDEN NEEDLE**
- SOME OR EVEN ALL THE ACTIVE COMPOUNDS MAY BE ULTIMATELY DISCARDED FOR OTHER REASONS SUCH AS UNDEDESIRABLE SIDE EFFECTS.

EXAMPLES OF BIOLOGICAL ASSAYS

	ASSAY			
	AID348	AID362	AID364	AID371
NUMBER OF COMPOUNDS	4946	4279	3311	3312
NUMBER OF ACTIVES	48	60	50	278
FRACTION OF ACTIVES	0.0097	0.0140	0.0151	0.0839

ENLARGING THE HAYSTACK

- NEED TO EXAMINE A MUCH LARGER LIST OF COMPOUNDS

ENLARGING THE HAYSTACK

- NEED TO EXAMINE A MUCH LARGER LIST OF COMPOUNDS
- **IDEA:** SORT THE COMPOUNDS SO THAT THE ACTIVE ONES ARE CLOSER TO THE TOP

ENLARGING THE HAYSTACK

- NEED TO EXAMINE A MUCH LARGER LIST OF COMPOUNDS
- **IDEA:** SORT THE COMPOUNDS SO THAT THE ACTIVE ONES ARE CLOSER TO THE TOP
- **BRING THE NEEDLES TO THE TOP OF THE LIST!**

DESCRIPTOR SETS

-

DESCRIPTOR SET	ASSAY AID348	AID362	AID364	AID371
ATOM PAIRS	367	360	380	382
BURDEN NUMBERS	24	24	24	24
CARHART ATOM PAIRS	1795	1319	1585	1498
FRAGMENT PAIRS	570	563	580	580
PHARMACOPHORES	122	112	120	119
NUMBER OF VARIABLES				

DESCRIPTOR SETS

DESCRIPTOR SET	ASSAY AID348	AID362	AID364	AID371
ATOM PAIRS	367	360	380	382
BURDEN NUMBERS	24	24	24	24
CARHART ATOM PAIRS	1795	1319	1585	1498
FRAGMENT PAIRS	570	563	580	580
PHARMACOPHORES	122	112	120	119
NUMBER OF VARIABLES				

- The descriptor sets are generated by the software PowerMV (Liu, Feng, and Young, 2005).

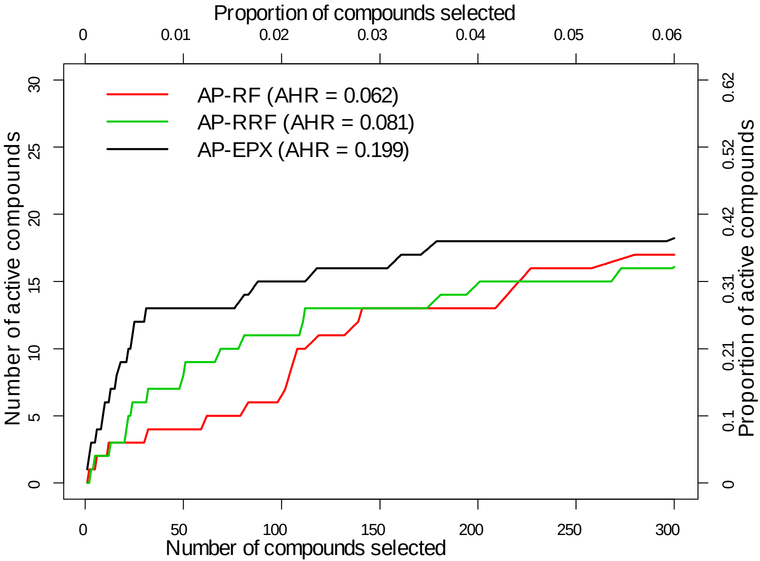
EVALUATING COMPETING SORTING PROCEDURES

- APPROPRIATE MEASURES FOR THIS EVALUATION WERE DEVELOPED TO THIS END

EVALUATING COMPETING SORTING PROCEDURES

- APPROPRIATE MEASURES FOR THIS EVALUATION WERE DEVELOPED TO THIS END
- I'LL DESCRIBE TWO OF THEM (THE MOST POPULAR ONES)

HIT CURVE


Proportion of compounds selected
0
0.01
0.02
0.03
0.04
0.05
0.06
AP-RF (AHR = 0.062)
AP-RRF (AHR = 0.081)
AP-EPX (AHR = 0.199)
Number of active compounds
0
5
10
15
20
25
30
Proportion of active compounds
0
0.1
0.21
0.31
0.42
0.52
0.62
0
50
100
150
200
250
300
Number of compounds selected

AVERAGE HIT RATE

SYMBOL	MEANING
N	NUMBER OF COMPOUNDS IN THE ASSAY
A	NUMBER OF **ACTIVE COMPOUNDS**
A(t)	NUMBER OF **ACTIVES** AMONG THE FIRST **t** COMPOUNDS

AVERAGE HIT RATE (CONTINUED)

POSITION OF THE ACTIVE COMPOUNDS IN THE SORTED LIST:

$$t_1 < t_2 < t_3 < \cdots < t_A$$

HIT RATES:

$$H(t_j) = \frac{A(t_j)}{t_j}$$

AVERAGE HIT RATE

$$\overline{H} = \frac{H(t_1) + H(t_2) + \cdots + H(t_A)}{A}$$

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)
  - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)
  - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)
- CLASSICAL SOLUTION TO THIS PROBLEM: VARIABLE SELECTION (REGULARIZATION)

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)
  - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)
- CLASSICAL SOLUTION TO THIS PROBLEM: VARIABLE SELECTION (REGULARIZATION)
  - RIDGE REGRESSION

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)
  - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)
- CLASSICAL SOLUTION TO THIS PROBLEM: VARIABLE SELECTION (REGULARIZATION)
  - RIDGE REGRESSION
  - LASSO

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
    - SOME OF THEM ARE USELESS (PURE NOISE)
    - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)
- CLASSICAL SOLUTION TO THIS PROBLEM: VARIABLE SELECTION (REGULARIZATION)
    - RIDGE REGRESSION
    - LASSO
    - LARS

DEALING WITH VERY LARGE DESCRIPTOR SETS

- PROBLEM: DESCRIPTOR SETS HAVE A LARGE NUMBER OF VARIABLES
  - SOME OF THEM ARE USELESS (PURE NOISE)
  - SOME OF THEM ARE HIGHLY COLINEAR (REDUNDANT)
- CLASSICAL SOLUTION TO THIS PROBLEM: VARIABLE SELECTION (REGULARIZATION)
  - RIDGE REGRESSION
  - LASSO
  - LARS
  - RANDOM FOREST (IT HAS BUILT-IN VARIABLE SELECTION CAPABILITY)

PHALANX: A NEW REGULARIZING FRAMEWORK

- **IDEA:** INSTEAD OF SORTING THE COMPOUNDS WITH **A SINGLE** REGULARIZED MODEL, FORM **SEVERAL MODELS (CALLED PHALANXES)** AND COMBINE THEM (MODEL AVERAGING)

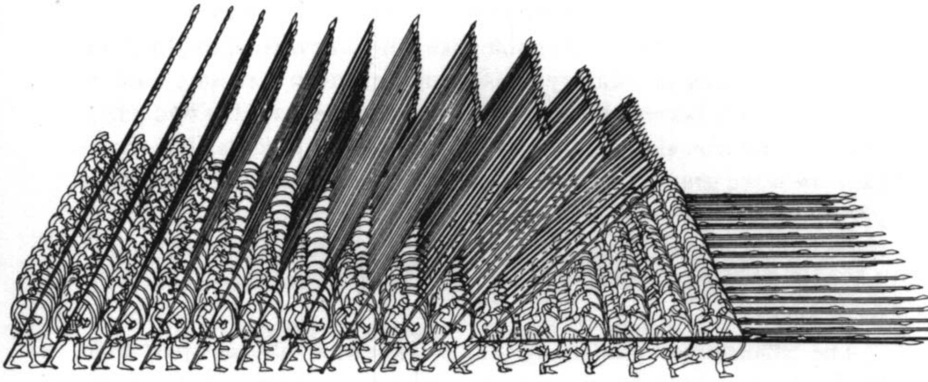
PHALANX: A NEW REGULARIZING FRAMEWORK

- **IDEA:** INSTEAD OF SORTING THE COMPOUNDS WITH **A SINGLE** REGULARIZED MODEL, FORM **SEVERAL MODELS (CALLED PHALANXES)** AND COMBINE THEM (MODEL AVERAGING)
- EACH MODEL MUST INCLUDE VARIABLES THAT **WORK WELL TOGETHER**

PHALANX: A NEW REGULARIZING FRAMEWORK

- **IDEA:** INSTEAD OF SORTING THE COMPOUNDS WITH **A SINGLE** REGULARIZED MODEL, FORM **SEVERAL MODELS (CALLED PHALANXES)** AND COMBINE THEM (MODEL AVERAGING)
- EACH MODEL MUST INCLUDE VARIABLES THAT **WORK WELL TOGETHER**
- THIS RESEMBLES THE ANCIENT MILITARY FORMATIONS USED BY **ALEXANDER THE GREAT** AND HIS FATHER **PHILIPPO II OF MACEDONIA**.

MACEDONIAN PHALANX



The Macedonian phalanx, here shown in its fighting formation of 256 men, the syntagma.

PHALANX: A NEW REGULARIZING FRAMEWORK

- WE CREATED AN ALGORITHM TO SELECT THE PHALANXES AND PRODUCE THE COMBINED SORTING

PHALANX: A NEW REGULARIZING FRAMEWORK

- WE CREATED AN ALGORITHM TO SELECT THE PHALANXES AND PRODUCE THE COMBINED SORTING
- UBC FILED A PRELIMINARY U.S.A. PATENT FOR THIS "INVENTION".

PHALANX: A NEW REGULARIZING FRAMEWORK

- WE CREATED AN ALGORITHM TO SELECT THE PHALANXES AND PRODUCE THE COMBINED SORTING
- UBC FILED A PRELIMINARY U.S.A. PATENT FOR THIS "INVENTION".
- OUR ALGORITHM IS A BIT INVOLVED AND WILL NOT BE DESCRIBED HERE

PHALANX: A NEW REGULARIZING FRAMEWORK

- WE CREATED AN ALGORITHM TO SELECT THE PHALANXES AND PRODUCE THE COMBINED SORTING
- UBC FILED A PRELIMINARY U.S.A. PATENT FOR THIS "INVENTION".
- OUR ALGORITHM IS A BIT INVOLVED AND WILL NOT BE DESCRIBED HERE
- PLEASE, REFER TO A FORTHCOMING PAPER (TOMAL, WELCH AND ZAMAR, 2013) AND TOMAL'S Ph.D. DISSERTATION (UBC)

- PHALANX PERFORMS BETTER (COMPARED TO STATE OF THE ART TECHNOLOGY) WHEN

- PHALANX PERFORMS BETTER (COMPARED TO STATE OF THE ART TECHNOLOGY) WHEN
  1. THE TRAINING DATA IS **“VARIABLES RICH”** AND **“OBSERVATIONS POOR”**

PHALANX PERFORMANCE

- PHALANX PERFORMS BETTER (COMPARED TO STATE OF THE ART TECHNOLOGY) WHEN
  1. THE TRAINING DATA IS **“VARIABLES RICH”** AND **“OBSERVATIONS POOR”**
  2. THERE ARE **FEW RARE CASES**

PHALANX PERFORMANCE

- PHALANX PERFORMS BETTER (COMPARED TO STATE OF THE ART TECHNOLOGY) WHEN
  1. THE TRAINING DATA IS **“VARIABLES RICH”** AND **“OBSERVATIONS POOR”**
  2. THERE ARE **FEW RARE CASES**
  3. IN SUMMARY: THE **HARDEST THE SORTING PROBLEM** IS, THE MOST PHALANX OUTPERFORMS AVAILABLE PROCEDURES

PHALANX DIVERSITY

- MODERN HIGH DIMENSIONAL PROBLEMS (E.G. GENOMICS, PROTEOMCS, FINANCE, ASTRONOMY) ARE COMPLEX AND MAY HAVE SEVERAL INTERNAL DRIVING FORCES

PHALANX DIVERSITY

- MODERN HIGH DIMENSIONAL PROBLEMS (E.G. GENOMICS, PROTEOMCS, FINANCE, ASTRONOMY) ARE COMPLEX AND MAY HAVE SEVERAL INTERNAL DRIVING FORCES
- PHALANXES ARE CAPABLE OF CAPTURING AND EXPLOITING THIS DIVERSITY

PHALANX DIVERSITY

- MODERN HIGH DIMENSIONAL PROBLEMS (E.G. GENOMICS, PROTEOMCS, FINANCE, ASTRONOMY) ARE COMPLEX AND MAY HAVE SEVERAL INTERNAL DRIVING FORCES
- PHALANXES ARE CAPABLE OF CAPTURING AND EXPLOITING THIS DIVERSITY
- INSTEAD OF **"CURSING DIMENSIONALITY"** PHALANX **"BLESSES DIMENSIONALITY"**.