Введение в кластеризацию

Кластеризация

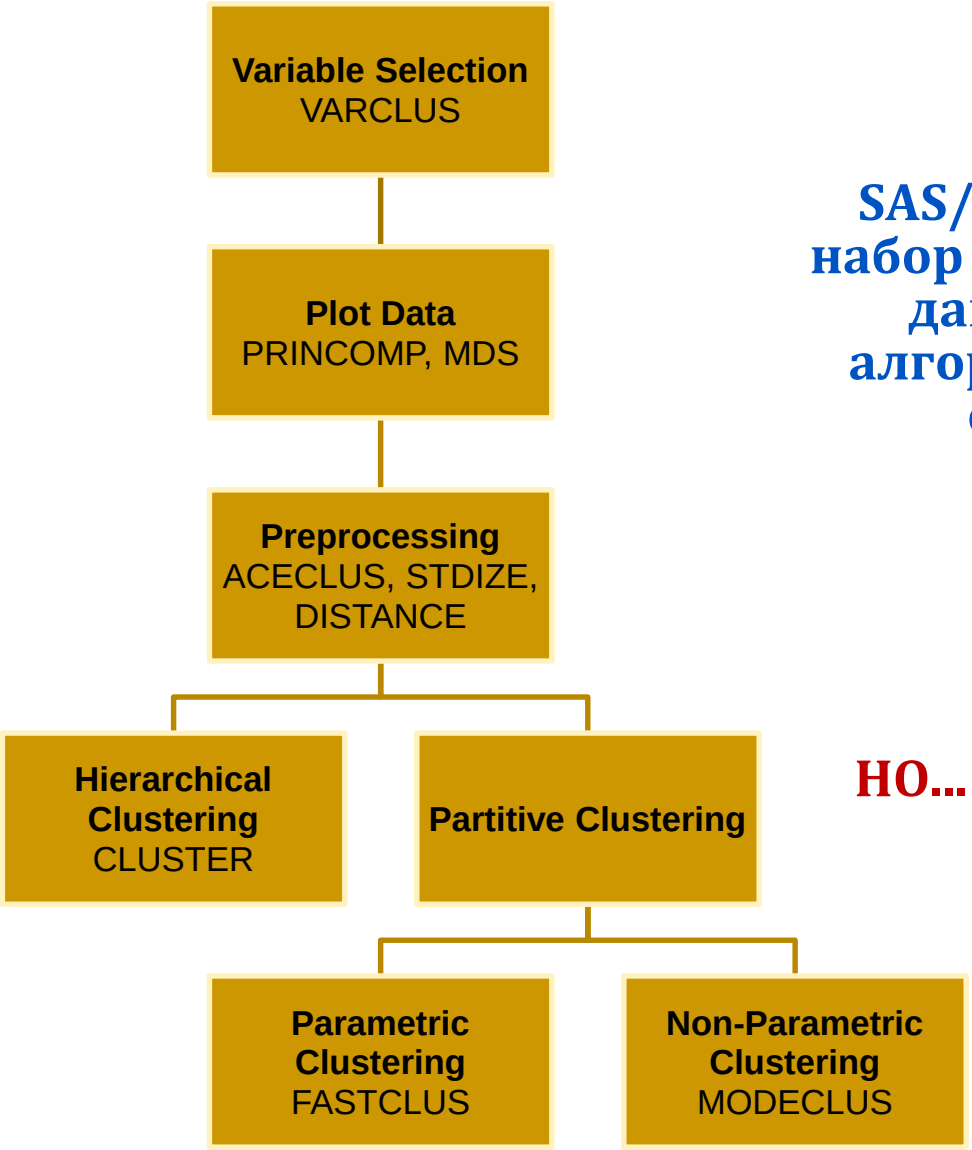
Кластеризация – процедура автоматического разбиения некоторого множества объектов на группы (кластеры) на основе степени их схожести

Кластеризация – **обучение без учителя** (целевая переменная не требуется, нужны лишь характеристики объектов)

Признаки «хорошего» кластера:

- близость объектов внутри кластера
- удаленность от остальных кластеров

Процедуры кластеризации в SAS/STAT



SAS/STAT содержит богатый набор процедур для подготовки данных, широкий выбор алгоритмов кластеризации и оценки результатов моделирования

НО… ЗАЧЕМ ВООБЩЕ НУЖНА КЛАСТЕРИЗАЦИЯ?

Example: Clustering for Customer Types

While you have thousands of customers, there are really only a handful of major types into which most of your customers can be grouped.

- Bargain hunter
- Man/woman on a mission
- Impulse shopper
- Weary parent
- DINK (dual income, no kids)

Example: Clustering for Store Location

You want to open new grocery stores in the U.S. based on demographics. Where should you locate the following types of new stores?

- low-end budget grocery stores
- small boutique grocery stores
- large full-service supermarkets

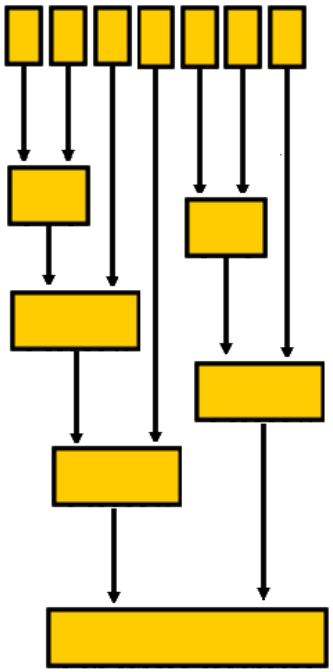
Профилирование кластеров

- *Профилирование* – это попытка вывести «бытовой» смысл группировки конкретных объектов в кластер
- Цель – определить уникальные черты (или их комбинации), характеризующие объекты в кластере

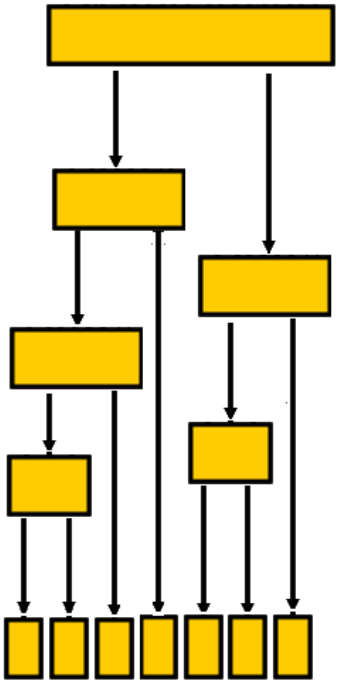
Виды кластеризации

Иерархическая кластеризация

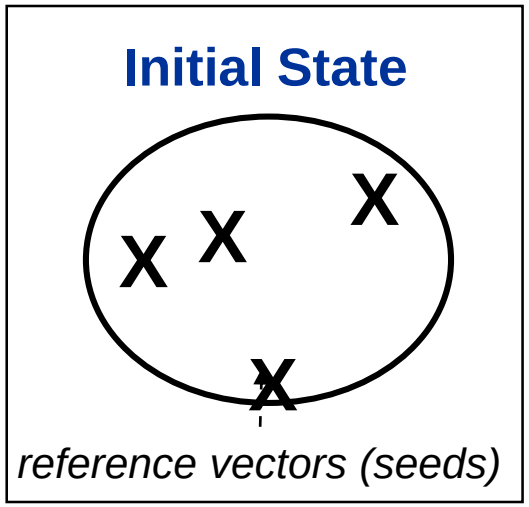
Агломеративная

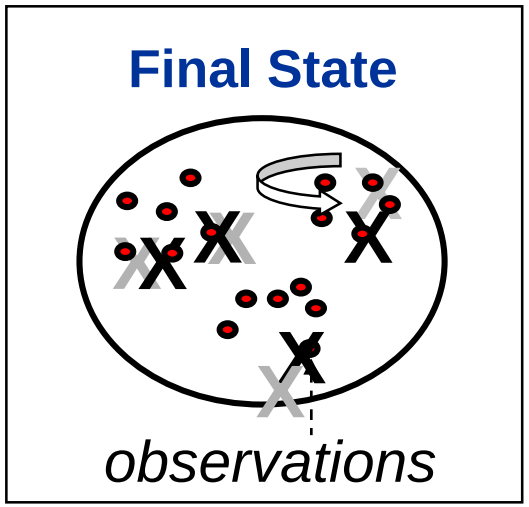


Дробящая



Partitive clustering





- **PROBLEMS!**
  - make you guess the number of clusters present
  - make assumptions about the shape of the clusters
  - influenced by seed location, outliers, and order of reading observations
  - impossible to determine the optimal grouping, due to the combinatorial explosion of potential solutions.

Partitive clustering: Heuristic Search

1. Generate an initial partitioning (based on the seeds) of the observations into clusters.
2. Calculate the change in error produced by moving each observation from its own cluster to each of the other clusters.
3. Make the move that produces the greatest reduction.
4. Repeat steps 2 and 3 until no move reduces error.

Меры сходства объектов

Principles of a Good Similarity Metric

Properties of a **good** similarity metrics:

1. symmetry:
   $d(x,y) = d(y,x)$
2. non-identical distinguishability:
   $d(x,y) \neq 0 \rightarrow x \neq y$
3. identical non-distinguishability:
   $d(x,y) = 0 \rightarrow x = y$
4. triangular inequality:
   $d(x,y) \leq d(x,z) + d(y,z)$

The DISTANCE Procedure

General form of the DISTANCE procedure:

```
PROC DISTANCE DATA=SAS-data-set
        METHOD=similarity-metric <options>;
    VAR level (variables < / option-list >);
RUN;
```

- Provides different distance measures for interval and nominal variables

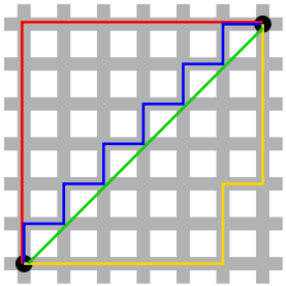
Simple popular Distance Metrics (Interval Vars)

- Euclidean distance

$$D_E = \|\mathbf{x} - \mathbf{w}\| = \sqrt{\sum_{i=1}^{k} (x_i - w_i)^2}$$

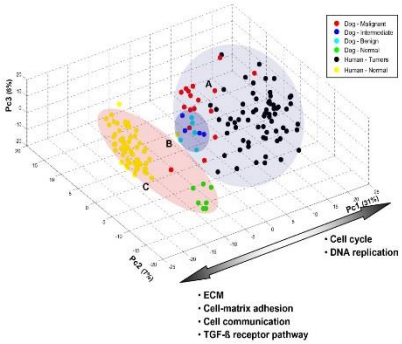
- City Block Distance

$$D_{M_1} = \sum_{i=1}^{d} |x_i - w_i|$$

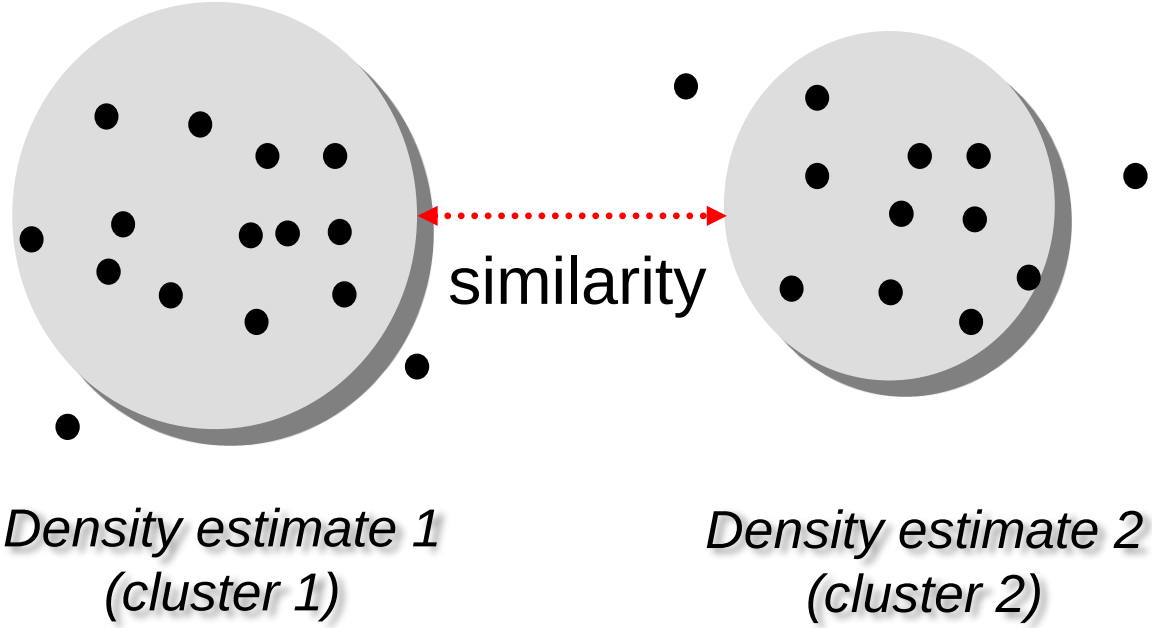


- Correlation

$$r = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})(y_i - \bar{y}) \right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$



Distance between Clusters: density-based



During more complex clustering processes one must not only calculate distances between *objects*, but also calculate distances between *sub-clusters*

Density-based methods define similarity as the distance between derived density “bubbles” (hyper-spheres).

Оценка качества кластеризации

От кластеров к классам

Если часть объектов выборки относится к разным классам, то это можно использовать для оценки качества кластеризации

Class	Cluster 1	Cluster 2	Cluster 3	Total
A	50	0	0	50
B	0	50	0	50
C	0	0	50	50
Total	50	50	50	150

Идеальные кластеры

Class	Cluster 1	Cluster 2	Cluster 3	Total
A	50	0	0	50
B	0	40	10	50
C	10	5	35	50
Total	60	45	45	150

Типичные кластеры

Class	Cluster 1	Cluster 2	Cluster 3	Total
A	50	0	0	50
B	50	0	0	50
C	50	0	0	50
Total	150	0	0	150

Это – не кластеризация ☺

От кластеров к вероятностям классов

The probability that a cluster represents a given class is given by the cluster's proportion of the row total.

Class	Cluster 1	Cluster 2	Cluster 3	Total
A	50	0	0	50
B	0	40	10	50
C	10	5	35	50
Total	60	45	45	150

Frequency

Class	Cluster 1	Cluster 2	Cluster 3	Total
A	1	0	0	1
B	0	0.8	0.2	1
C	0.2	0.1	0.7	1

Probability

Меры качества кластеризации

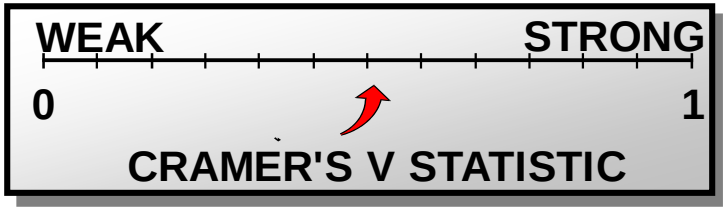
- The chi-square statistic is used to determine whether an association exists.

$$\chi^2 = \sum_i \sum_j \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

- Because the chi-square value grows with sample size, it does **not measure** the strength of the association.

- Normally, **Cramer's V ranges from 0 to 1**
  For 2x2 tables only, it ranges between -1 and 1

$$\phi_c = \sqrt{\frac{\varphi^2}{(k-1)}} = \sqrt{\frac{\chi^2}{N(k-1)}}$$



Подготовка и разведочный анализ данных

The Challenge of Opportunistic Data

Getting anything useful out of tons of data

Подготовка и анализ данных

1. Выбор данных и создание подвыборок
   (*Что я разбиваю на кластеры?*)

2. Отбор переменных
   (*Какие характеристики объектов важны?*)

3. Визуальный анализ данных
   (*Какой формы кластеры и сколько их?*)

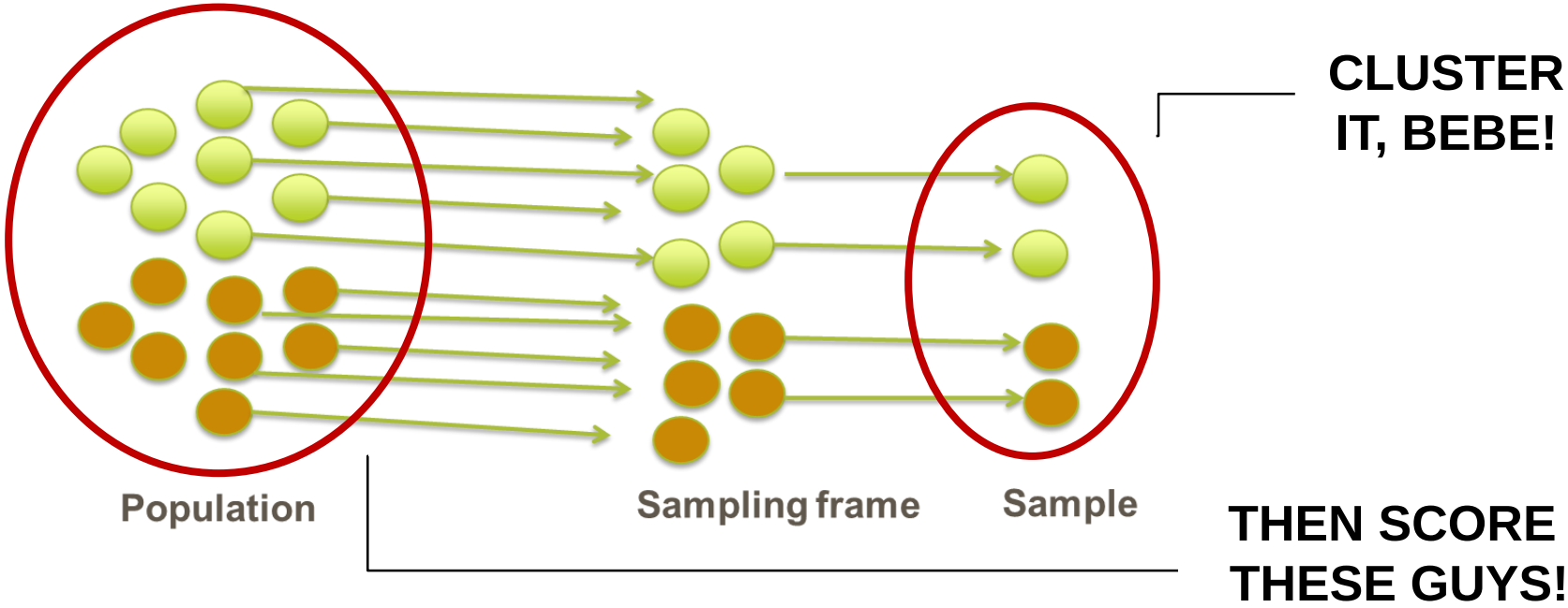
4. Стандартизация переменных
   (*Сравнимы ли масштабы переменных?*)

5. Трансформация переменных
   (*Переменные коррелируют? Кластеры не сферичны?*)

Data and Sample Selection

- Not necessary to cluster a large population if you use clustering techniques that lend themselves to scoring (for example: *Ward's, k-means*)

- It is useful to take a random sample for clustering and score the remainder of the larger population



Подготовка и разведочный анализ данных

Отбор переменных

Подготовка и анализ данных

1. Выбор данных и создание подвыборок
   (*Что я разбиваю на кластеры?*)

2. Отбор переменных
   (*Какие характеристики объектов важны?*)

3. Визуальный анализ данных
   (*Какой формы кластеры и сколько их?*)

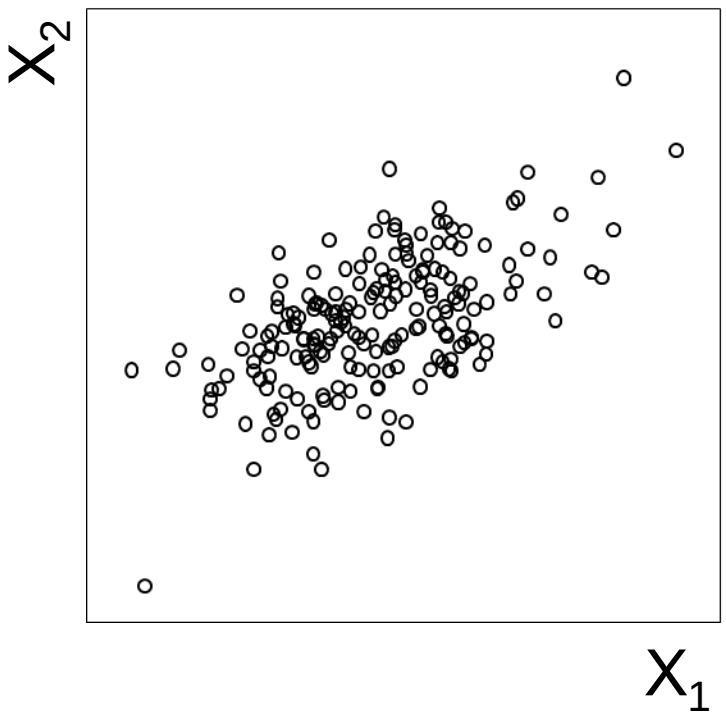
4. Стандартизация переменных
   (*Сравнимы ли масштабы переменных?*)

5. Трансформация переменных
   (*Переменные коррелируют? Кластеры не сферичны?*)

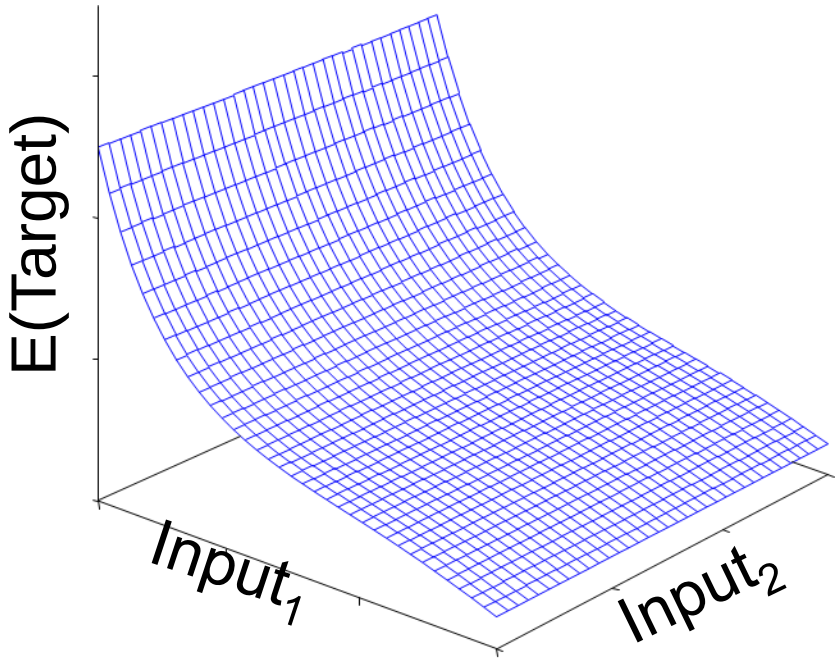
Снижение размерности

Необходимо ли анализировать ***все*** данные?

Избыточность



Незначимость

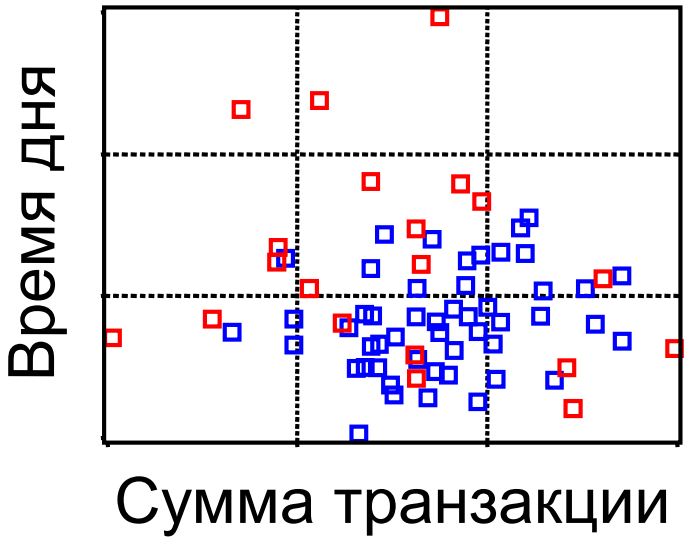


Отбор значимых переменных

- Регрессионные модели автоматически определяют значимость переменных на основе их влияния на целевую переменную
- Но в кластерном анализе целевой переменной НЕТ
- Поэтому все незначимые переменные должны быть удалены перед проведением кластеризации путем:
  - *Анализа важности переменных на специально подготовленной выборке с целевой*
  - *Подключения априорных соображений*

Секрет качественной кластеризации

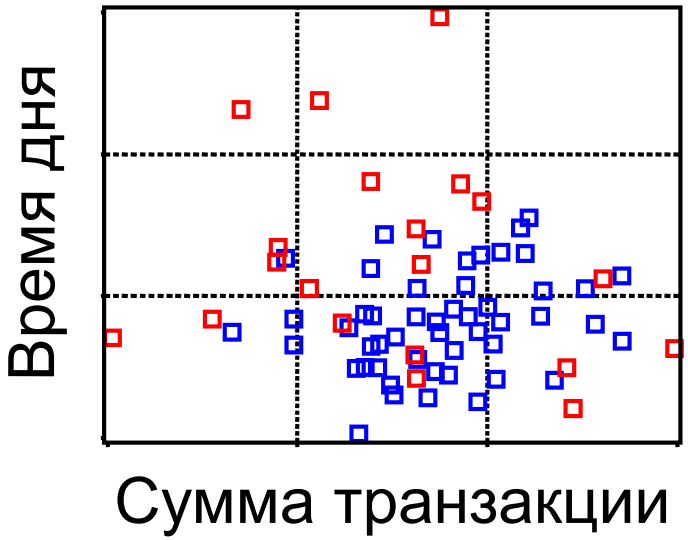
- Мошенник
- OK

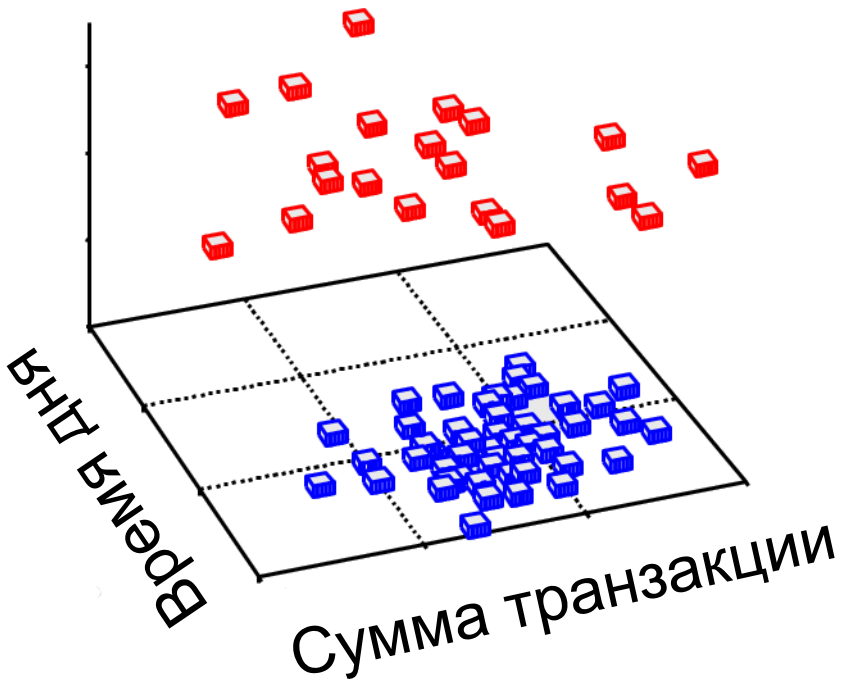


...

Секрет качественной кластеризации

- Мошенник
- OK



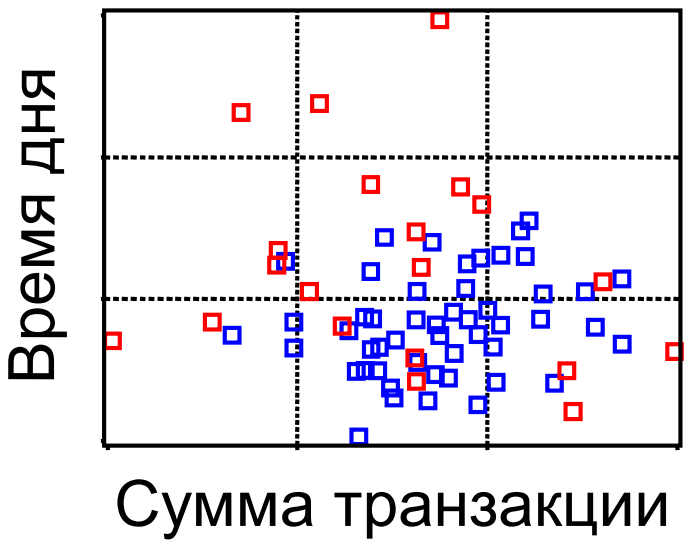


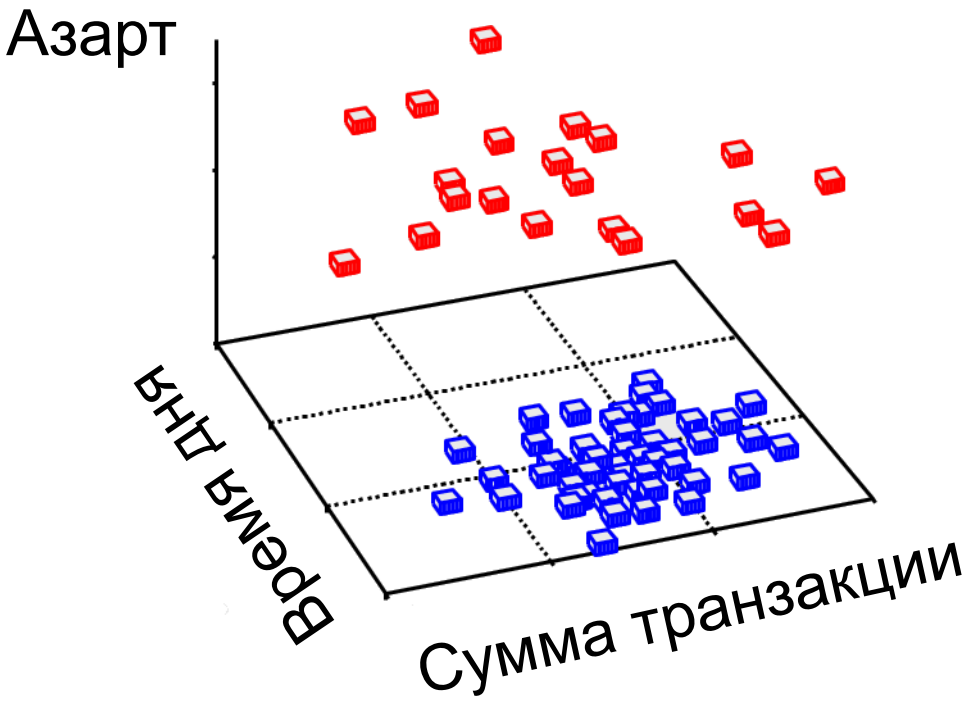
...

Секрет качественной кластеризации

Мошенник

OK





Больше нескоррелированных переменных = кластеры лучше!

...

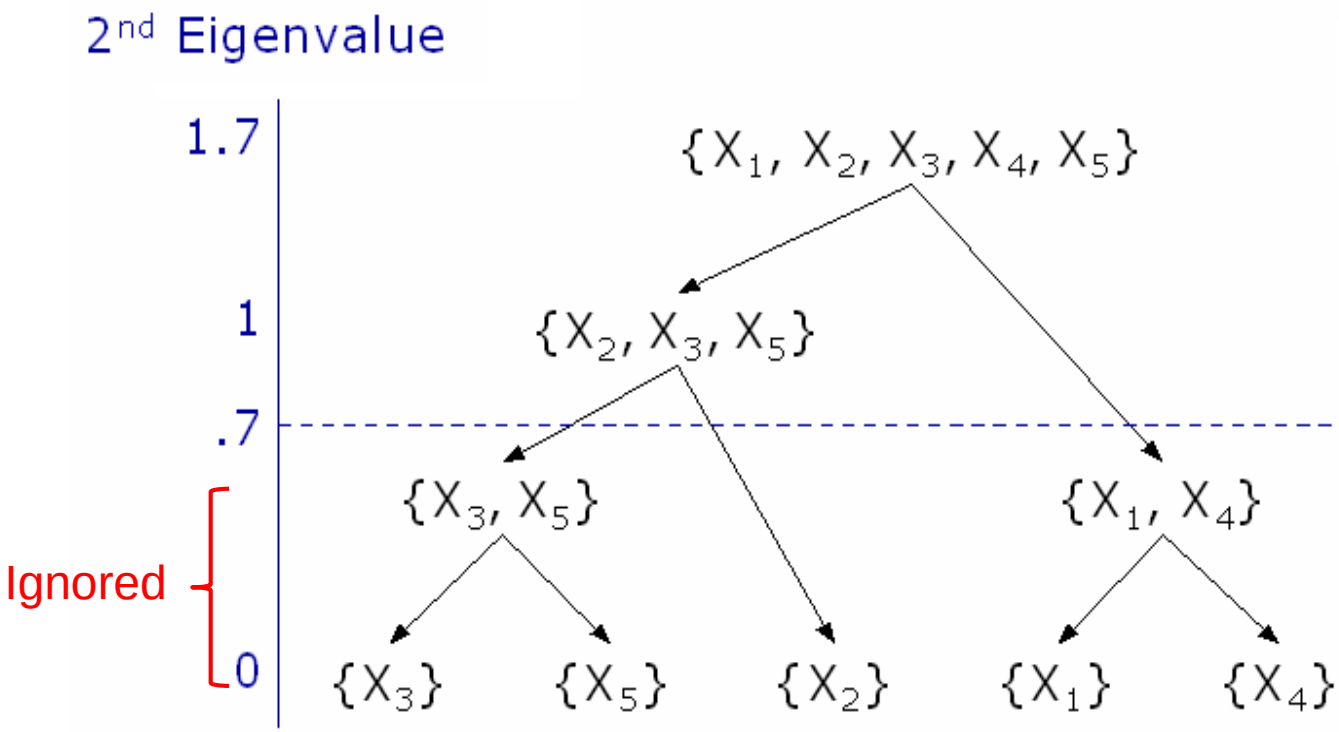
Удаление избыточных переменных

```
PROC VARCLUS DATA=SAS-data-set <options>;
      BY variables;
      VAR variables;
RUN;
```

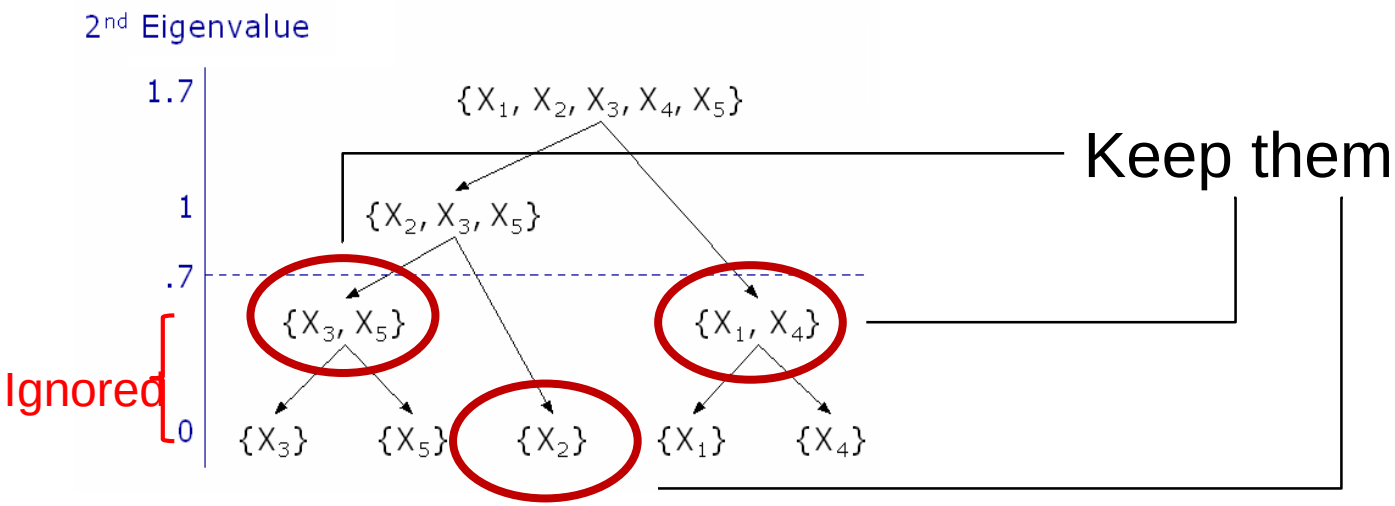
- **PROC VARCLUS** группирует избыточные переменные
- Из каждой группы выбирается по одному представителю, а остальные переменные удаляются, тем самым снижая коллинеарность и число переменных

Divisive Clustering

PROC VARCLUS uses *divisive clustering* to create variable subgroups that are as dissimilar as possible.



В основе метода –
Principal Component Analysis



Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.171782	0.5960	0.5960	2.290457	0.0175	
2	5.698305	0.8140	0.7868	0.828600	0.6592	0.3491
3	6.415722	0.9165	0.8822	0.235559	0.8774	0.2660

clus02d01.sas

Подготовка и разведочный анализ данных

Визуальный анализ

Подготовка и анализ данных

1. Выбор данных и создание подвыборок
   (*Что я разбиваю на кластеры?*)

2. Отбор переменных
   (*Какие характеристики объектов важны?*)

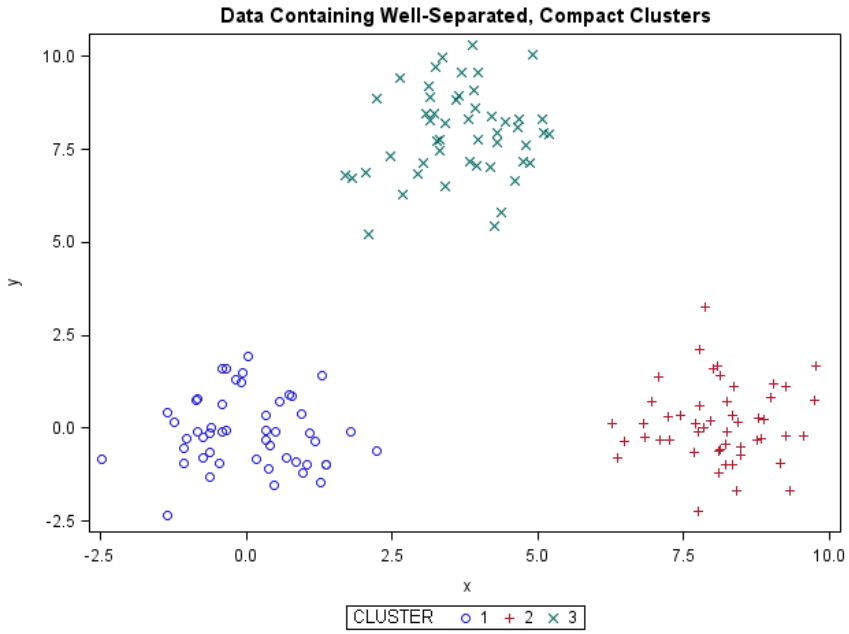
3. Визуальный анализ данных
   (*Какой формы кластеры и сколько их?*)

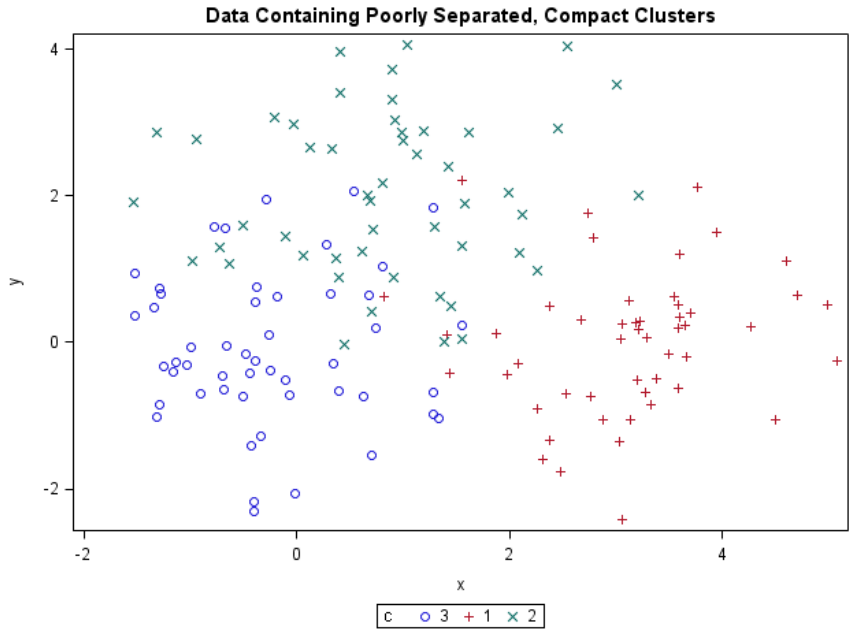
4. Стандартизация переменных
   (*Сравнимы ли масштабы переменных?*)

5. Трансформация переменных
   (*Переменные коррелируют? Кластеры не сферичны?*)

Визуальный анализ данных

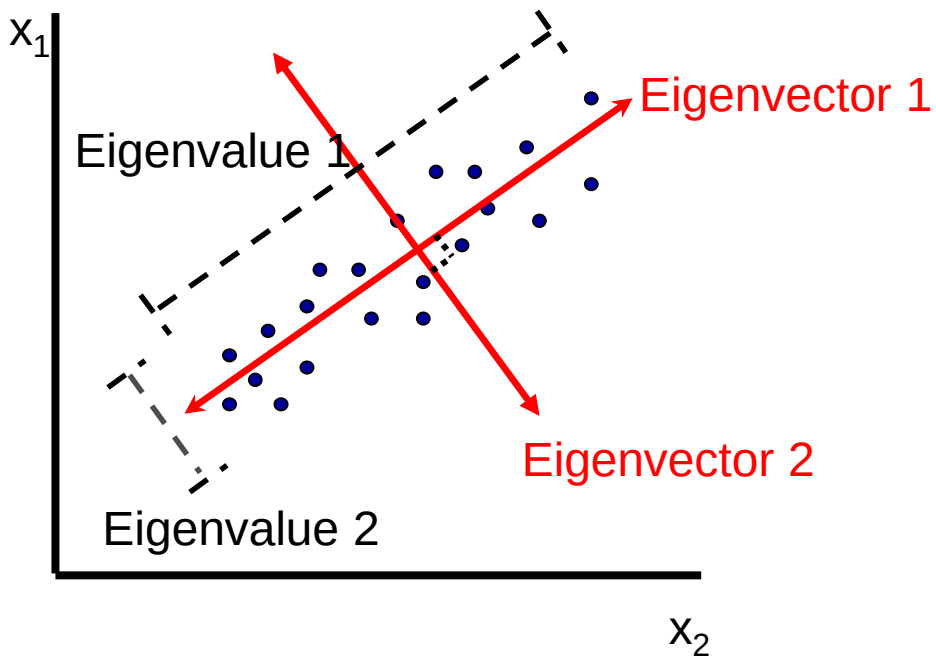
- Визуализация помогает установить такие ключевые параметры задачи, как
  - форму кластеров
  - дисперсию кластеров
  - примерное количество кластеров





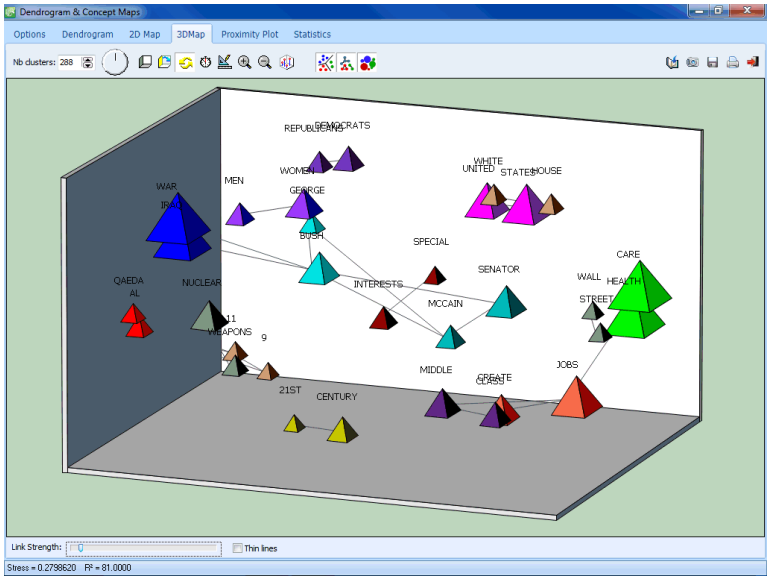
Principal Component Plots

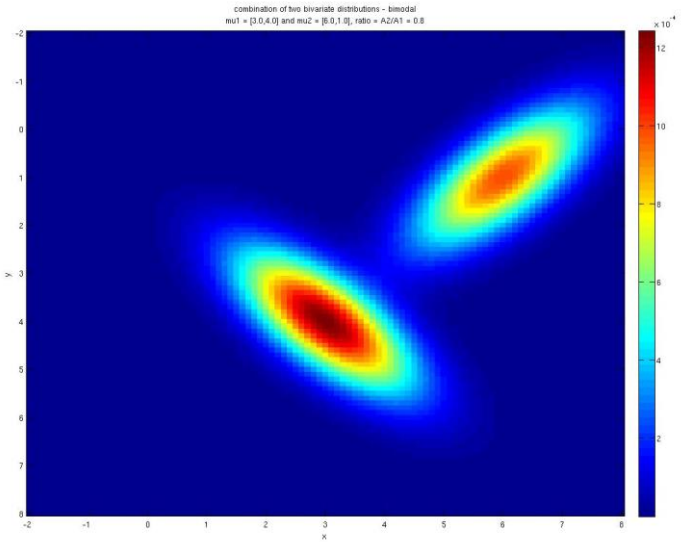
```
PROC PRINCOMP DATA=SAS-data-set <options>;
      BY variables;
      VAR variables;
RUN;
```



Multidimensional Scaling Plots

```
PROC MDS  DATA=distance_matrix <options>;
      VAR variables;
RUN;
```





Подготовка и разведочный анализ данных

Стандартизация переменных

Подготовка и анализ данных

1. Выбор данных и создание подвыборок
   (*Что я разбиваю на кластеры?*)

2. Отбор переменных
   (*Какие характеристики объектов важны?*)

3. Визуальный анализ данных
   (*Какой формы кластеры и сколько их?*)

4. Стандартизация переменных
   (*Сравнимы ли масштабы переменных?*)

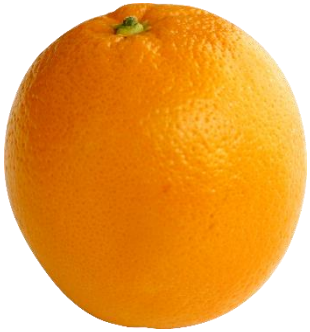
5. Трансформация переменных
   (*Переменные коррелируют? Кластеры не сферичны?*)

PROC STDIZE

Общий вид процедуры STDIZE:

```
PROC STDIZE DATA=SAS-data-set
        METHOD=method <options>;
      VAR variables;
RUN;
```

?



Опять сравнивают апельсины и слонов?

Хватит это терпеть!

Бесчисленные методы стандартизации

METHOD	LOCATION	SCALE
MEAN	mean	1
MEDIAN	median	1
SUM	0	sum
EUCLEN	0	Euclidean Length
USTD	0	standard deviation about origin
STD	mean	standard deviation
RANGE	minimum	range
MIDRANGE	midrange	range/2
MAXABS	0	maximum absolute value
IQR	median	interquartile range
MAD	median	median absolute deviation from median
ABW(c)	biweight 1-step M-estimate	biweight A-estimate
AHUBER(c)	Huber 1-step M-estimate	Huber A-estimate
AWAVE(c)	Wave 1-step M-estimate	Wave A-estimate
AGK(p)	mean	AGK estimate (ACECLUS)
SPACING(p)	mid minimum-spacing	minimum spacing
L(p)	L(p)	L(p) (Minkowski distances)
IN(ds)	read from data set	read settings from data set "ds"

Бесчисленные методы стандартизации

METHOD	LOCATION	SCALE
MEAN	mean	1
MEDIAN	median	1
SUM	0	sum
EUCLEN	0	Euclidean Length
USTD	0	standard deviation about origin
STD	mean	standard deviation
RANGE	minimum	range
MIDRA		
MAXA		
IQR	median	interquartile range
MAD	median	median absolute deviation from median
ABW(c)	biweight 1-step M-estimate	biweight A-estimate
AHUBER(c)	Huber 1-step M-estimate	Huber A-estimate
AWAVE(c)	Wave 1-step M-estimate	Wave A-estimate
AGK(p)	mean	AGK estimate (ACECLUS)
SPACING(p)	mid minimum-spacing	minimum spacing
L(p)	L(p)	L(p) (Minkowski distances)
IN(ds)	read from data set	read settings from data set "ds"

The best of the best of the best!

Подготовка и разведочный анализ данных

Трансформация переменных

Подготовка и анализ данных

1. Выбор данных и создание подвыборок
   (*Что я разбиваю на кластеры?*)

2. Отбор переменных
   (*Какие характеристики объектов важны?*)

3. Визуальный анализ данных
   (*Какой формы кластеры и сколько их?*)

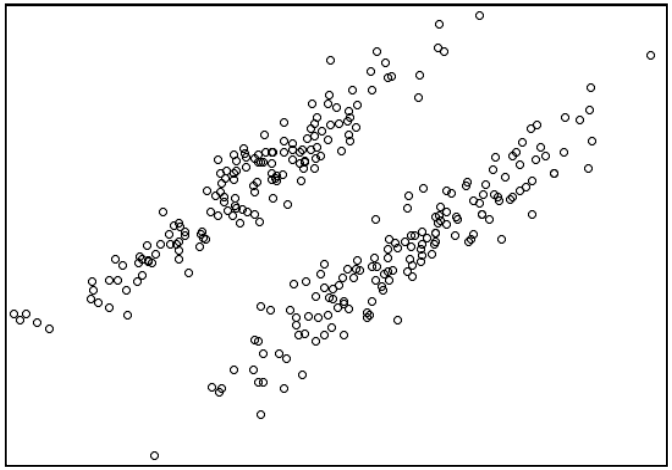
4. Стандартизация переменных
   (*Сравнимы ли масштабы переменных?*)

5. Трансформация переменных
   (*Переменные коррелируют? Кластеры не сферичны?*)

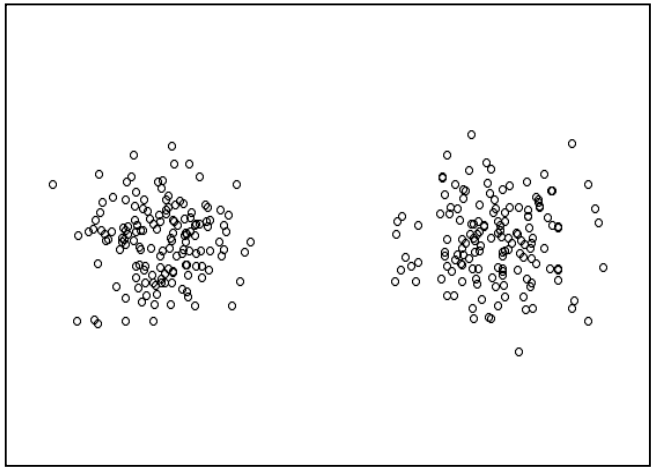
PROC ACECLUS

Общий вид процедуры ACECLUS:

```
PROC ACECLUS DATA=SAS-data-set <options>;
      VAR variables;
RUN;
```

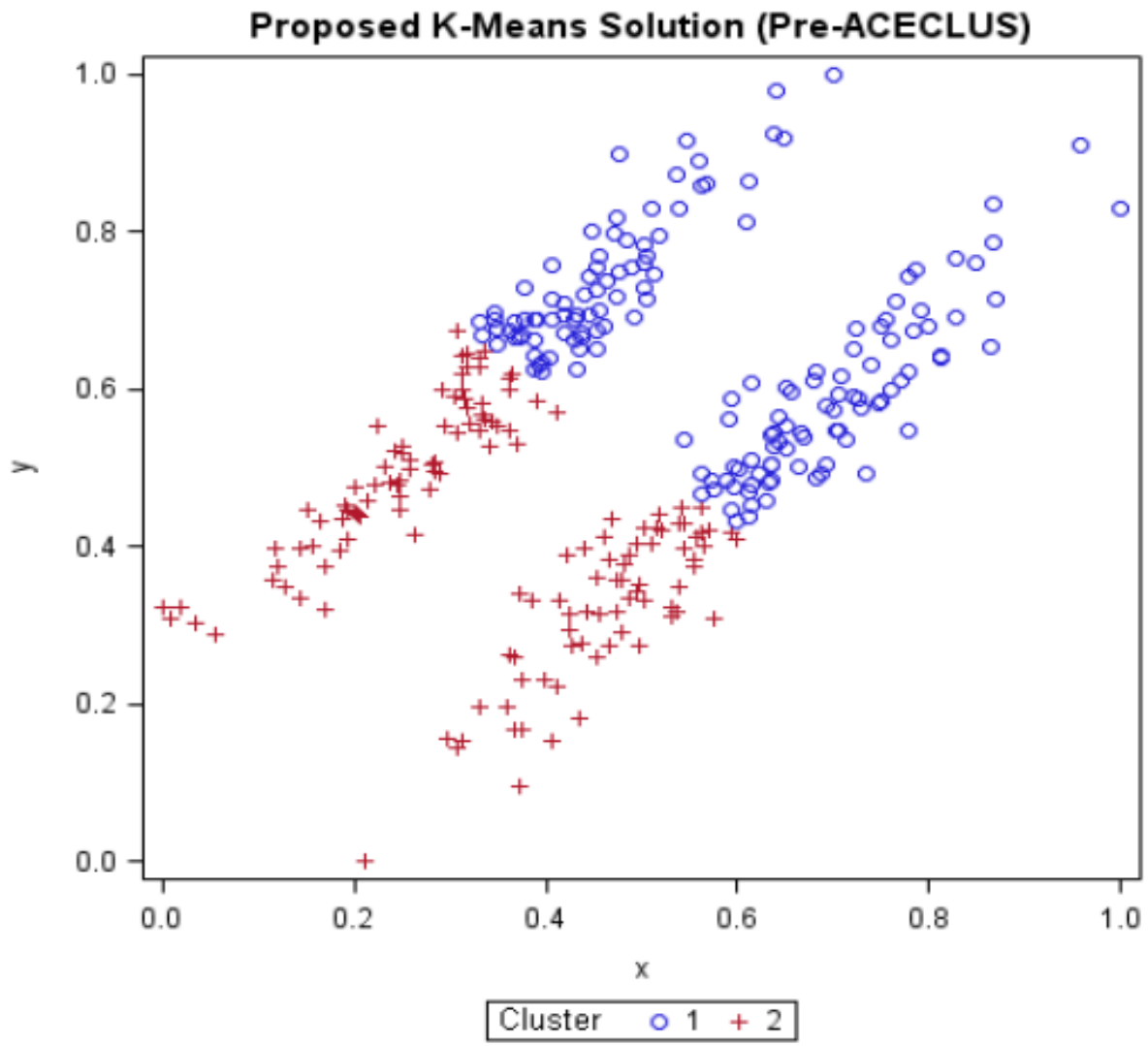


До ACECLUS

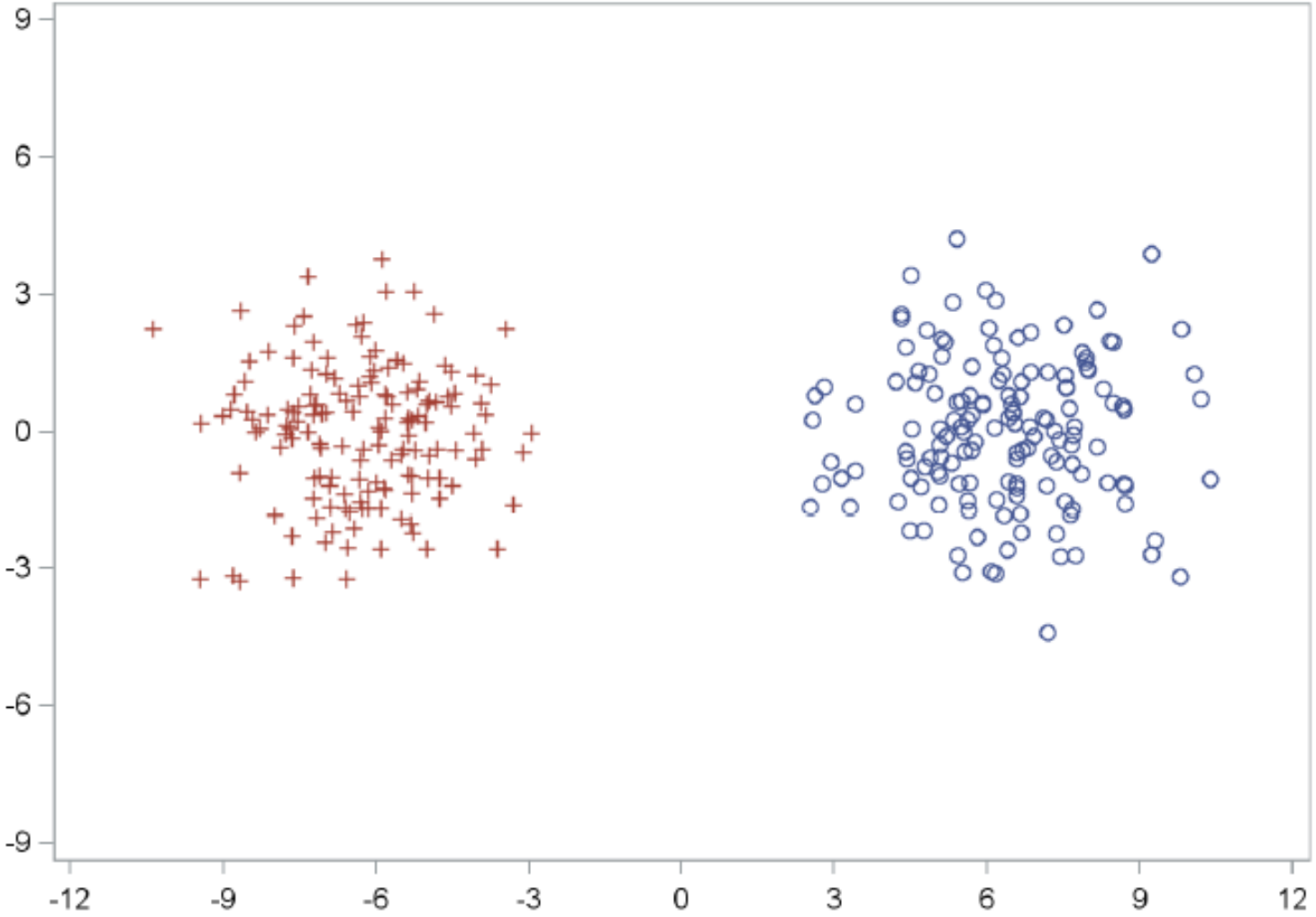


После ACECLUS

PROC ACECLUS



PROC ACECLUS



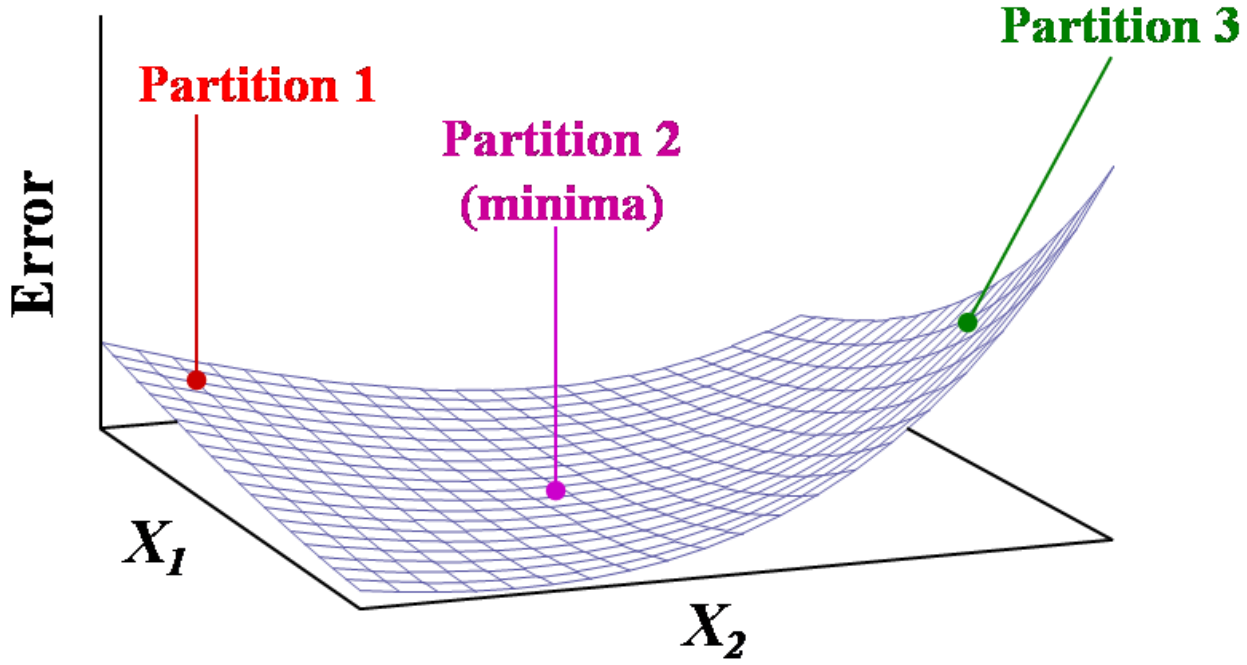
Partitive Clustering

Partitive Clustering

Алгоритм K-Means

Partitive Clustering: optimization

- Кластеризация разделением оптимизирует некоторую штрафную функцию, например:
  - межкластерное расстояние
  - внутрикластерную однородность (похожесть)



Семейства алгоритмов

- Оптимизирующие естественный критерий группировки(*K-means*)
- Параметрическое семейство алгоритмов (*Expectation-Maximization*)
- Непараметрическое семейство алгоритмов (*Kernel-based*)

Критерий естественной группировки

- Поиск наилучшего разбиения множества объектов на кластеры можно практически всегда можно свести к оптимизации **критерия естественной группировки**:
  - Максимизировать межкластерную сумму квадратов расстояний, или
  - Минимизировать внутрикластерную сумму квадратов расстояний между объектами.

- Большое межкластерное расстояние говорит о хорошей разделенности кластеров

- Малое внутрикластерное расстояние – признак однородности объектов внутри группы

Cross-Cluster Variation Matrix

$$W = \begin{bmatrix} W_{11} W_{12} W_{13} \cdots \\ W_{21} W_{22} W_{23} \cdots \\ W_{31} W_{32} W_{33} \cdots \\ \cdots\cdots\cdots\cdots W_{nn} \end{bmatrix}$$

The Trace Function

$$trace(\mathbf{W}) = \sqrt{\sum SS(w)_i} = \sqrt{\left(\frac{n}{m}\right)\sum\left(x_i - c_i\right)^2}$$

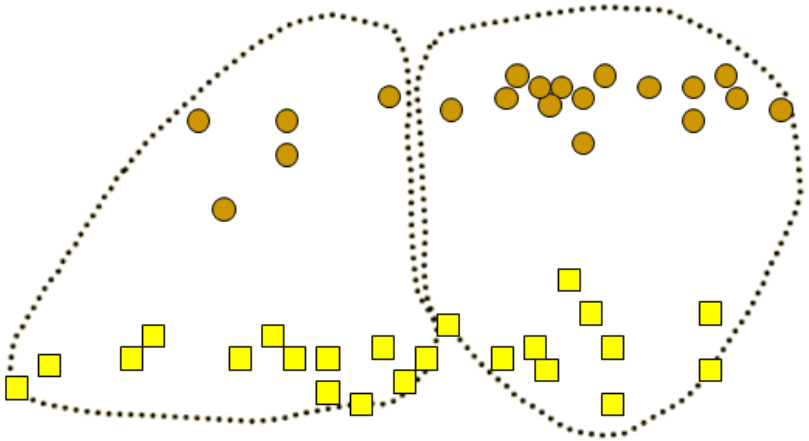
- Trace summarizes matrix **W** into a single number by adding together its diagonal (variance) elements.
- Simply adding matrix elements together makes trace very efficient, but it also makes it scale dependent

+

- Ignores the off-diagonal elements, so variables are treated as if they were independent (uncorrelated).
- Diminishes the impact of information from correlated variables.

-

Basic Trace(W) Problems

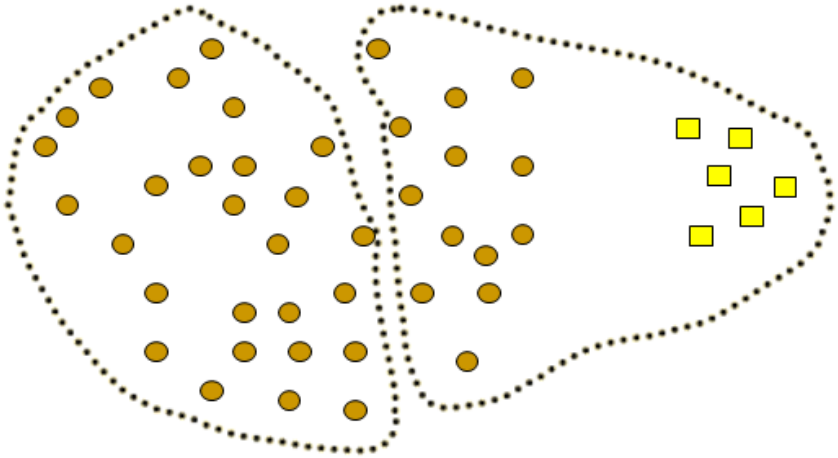


Spherical Structure Problem

- Because the trace function only looks at the diagonal elements of **W**, it tends to form spherical clusters
- Use data transformation techniques

Similar Size Problem

- *Trace*(**W**) also tends to produce clusters with about the same number of observations
- Alternative clustering techniques exist to manage this problem.



Partitive Clustering

Алгоритм K-Means
PROC FASTCLUS

The *K*-Means Methodology

The three-step *k*-means methodology:

1. Select (or specify) an initial set of cluster seeds

The *K*-Means Methodology

The three-step *k*-means methodology:

1. Select (or specify) an initial set of cluster seeds

2. Read the observations and update the seeds (known after the update as reference vectors). Repeat until convergence is attained

The *K*-Means Methodology

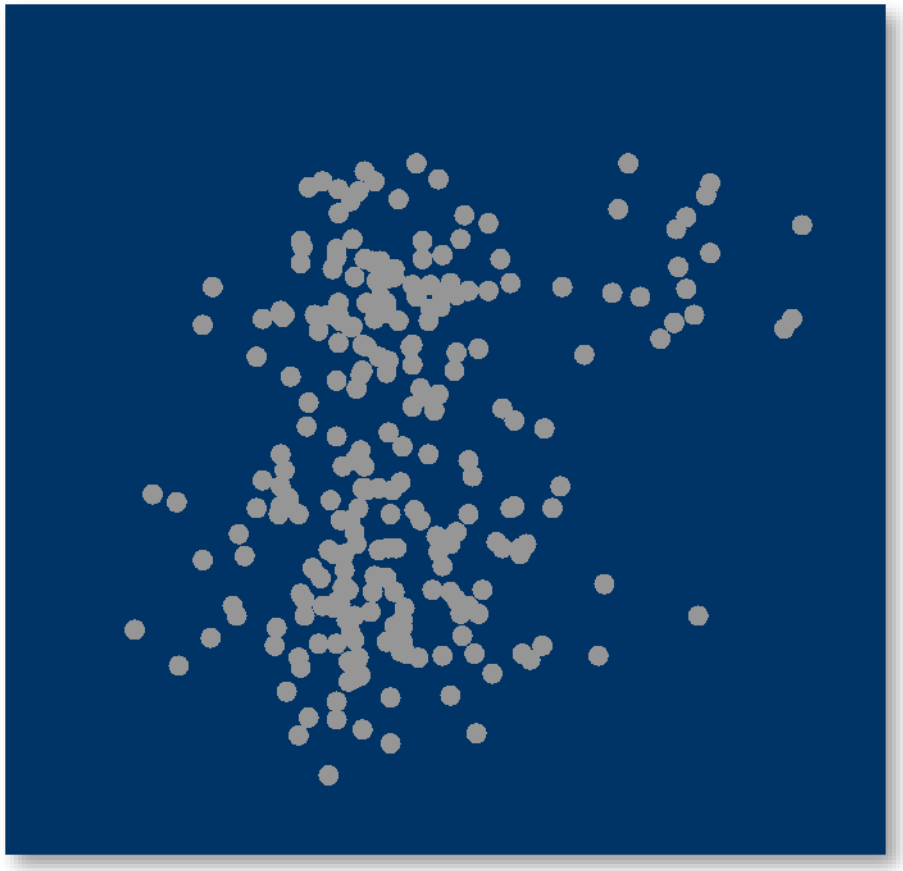
The three-step *k*-means methodology:

1. Select (or specify) an initial set of cluster seeds

2. Read the observations and update the seeds (known after the update as reference vectors). Repeat until convergence is attained

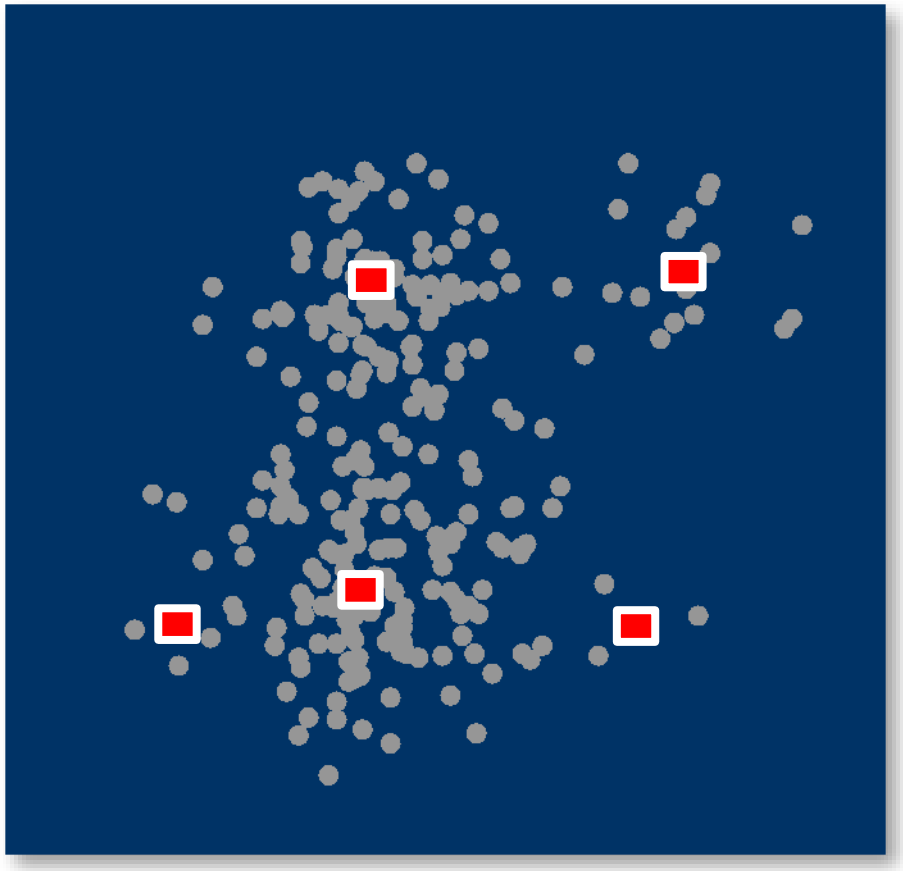
3. Make one final pass through the data, assigning each observation to its nearest reference vector

k-Means Clustering Algorithm



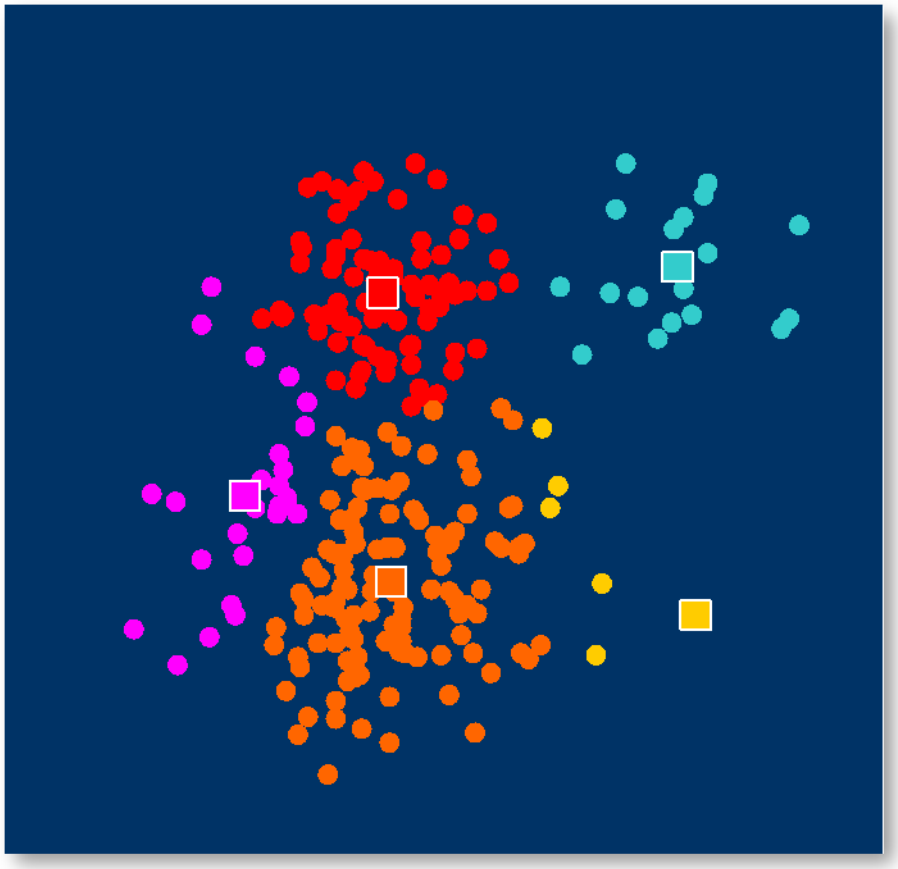
1. **Select inputs.**
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Re-assign cases.
6. Repeat steps 4 and 5 until convergence.

k-Means Clustering Algorithm



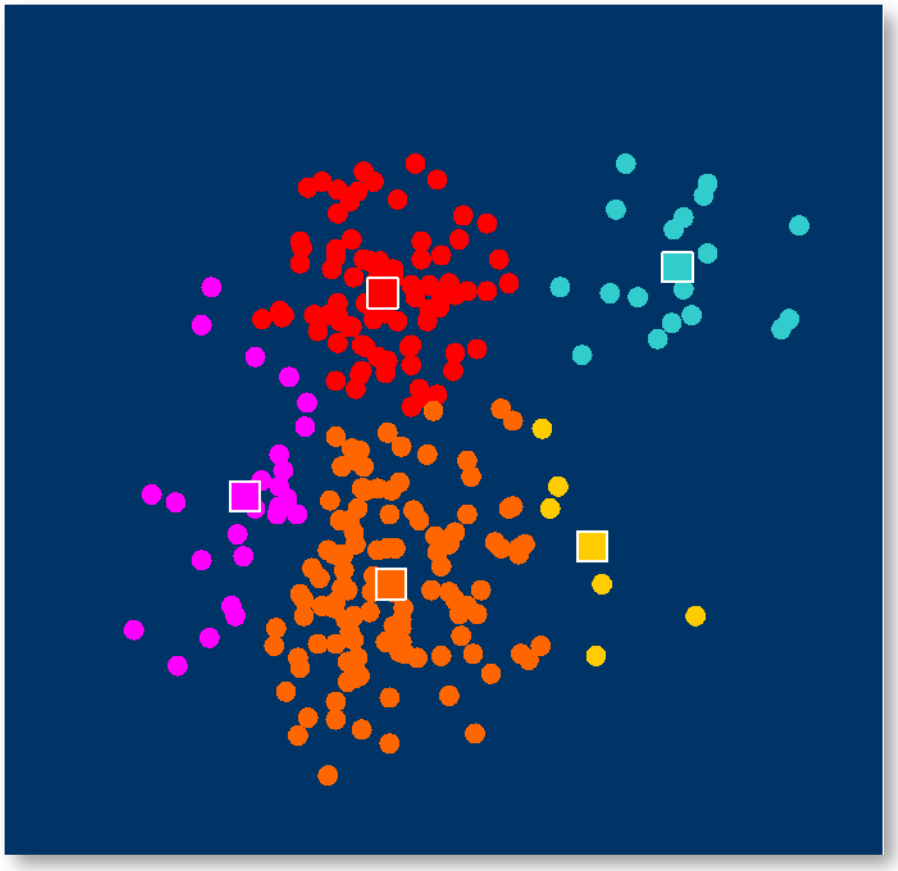
1. Select inputs.
2. **Select *k* cluster centers.**
3. Assign cases to closest center.
4. Update cluster centers.
5. Re-assign cases.
6. Repeat steps 4 and 5 until convergence.

k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. **Assign cases to closest center.**
4. Update cluster centers.
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

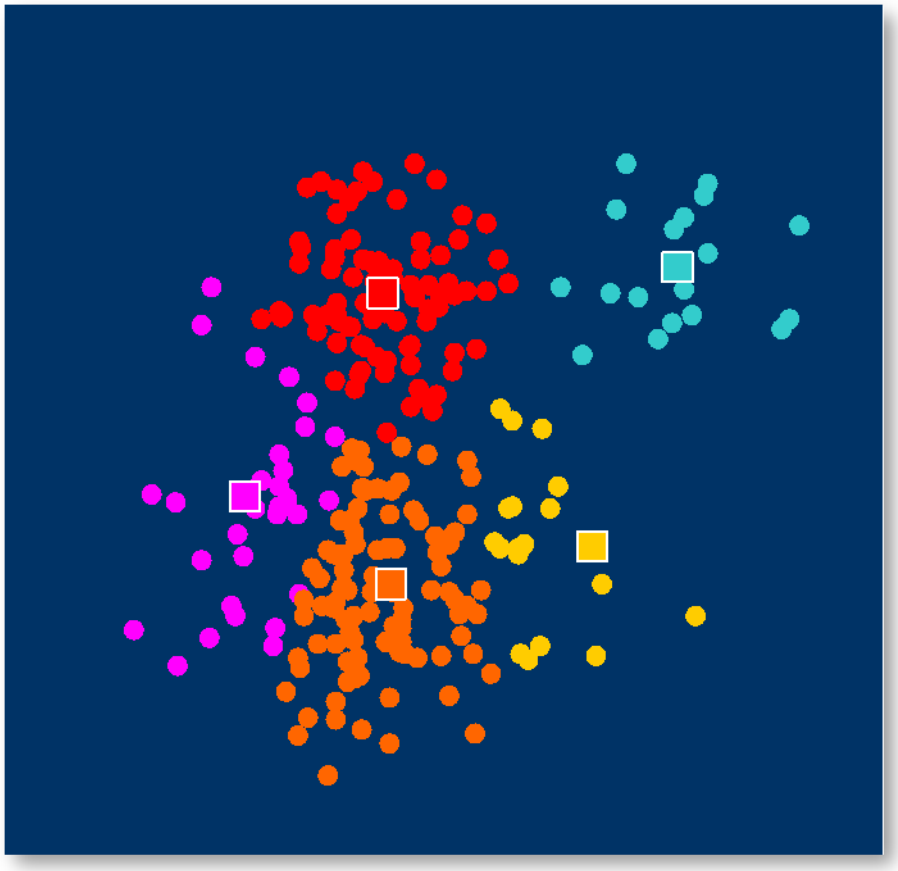
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

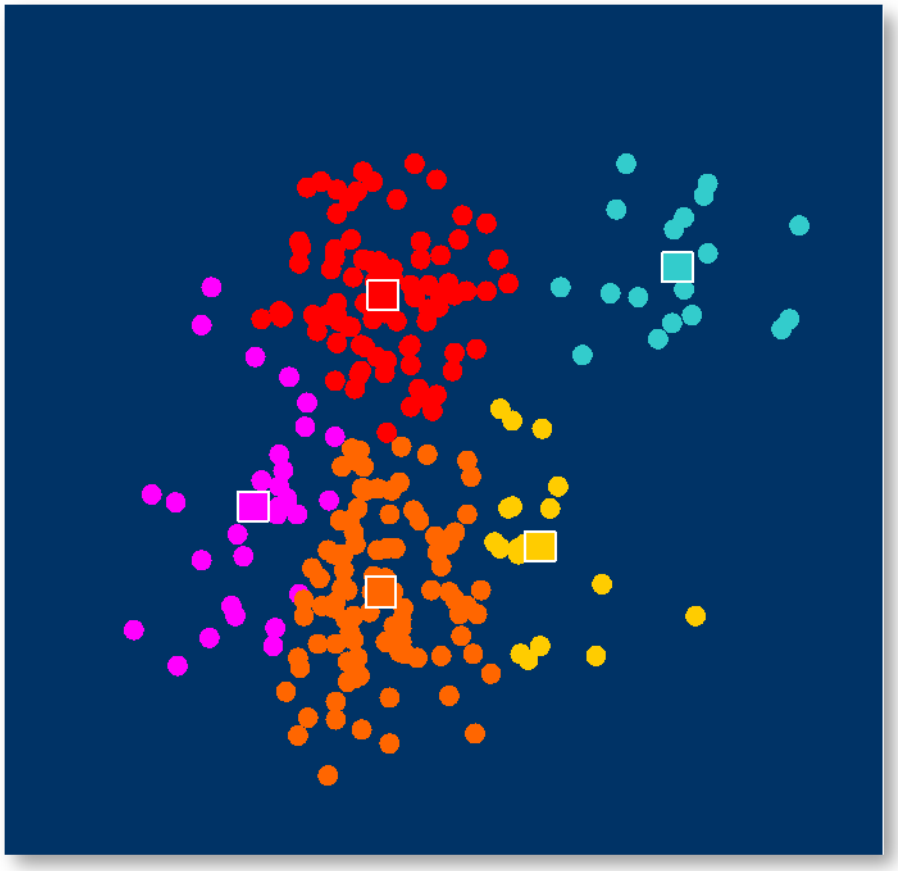
...

k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. Repeat steps 4 and 5 until convergence.

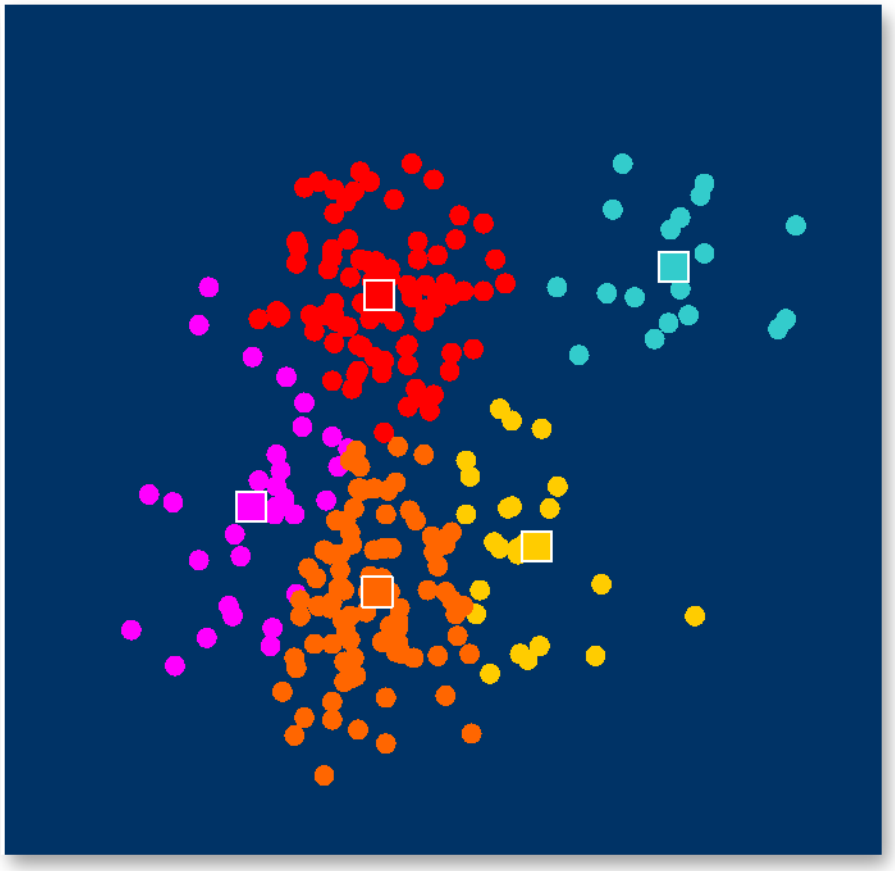
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

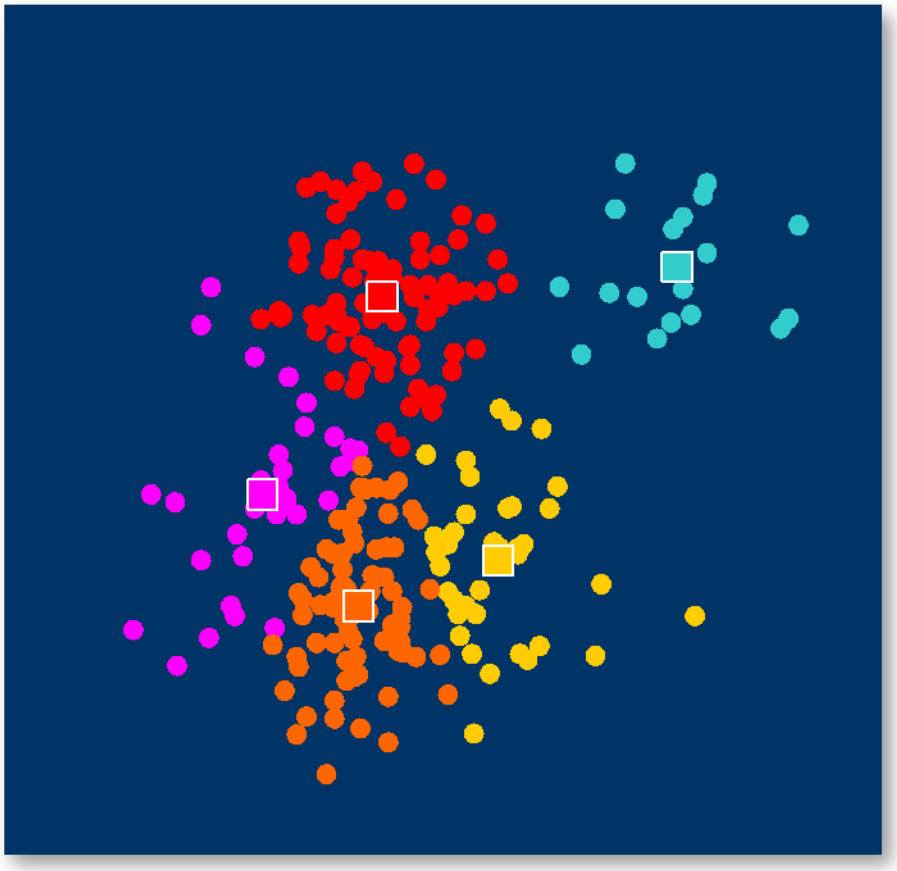
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. **Repeat steps 4 and 5 until convergence.**

...

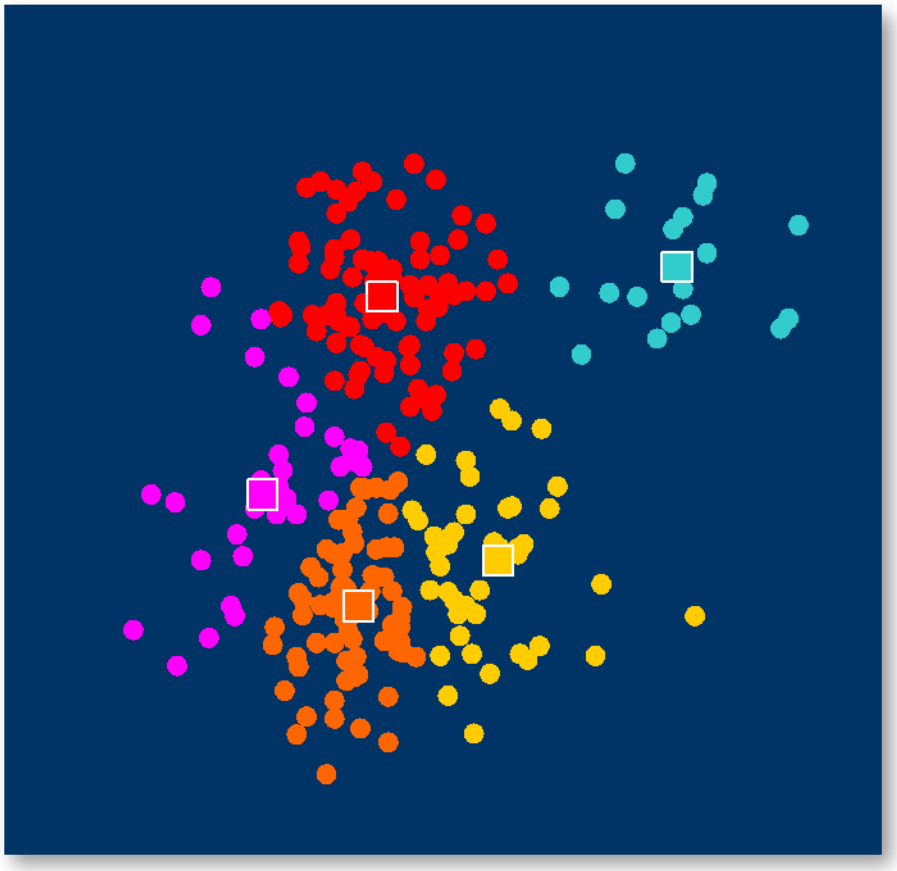
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

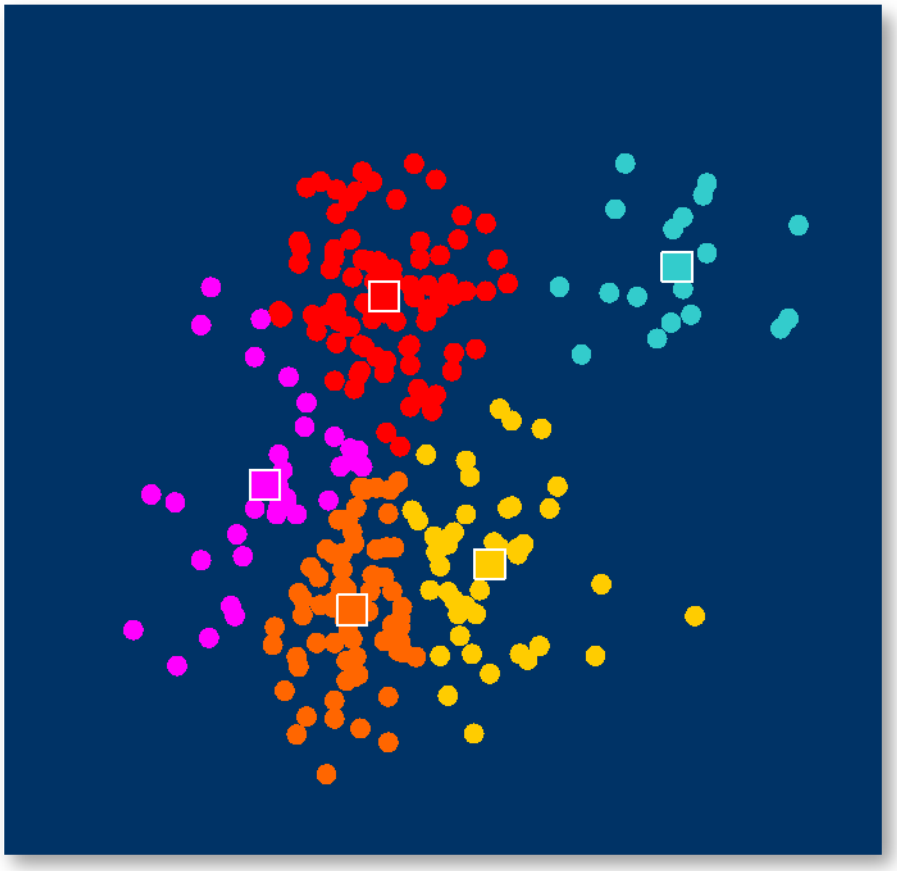
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. **Repeat steps 4 and 5 until convergence.**

...

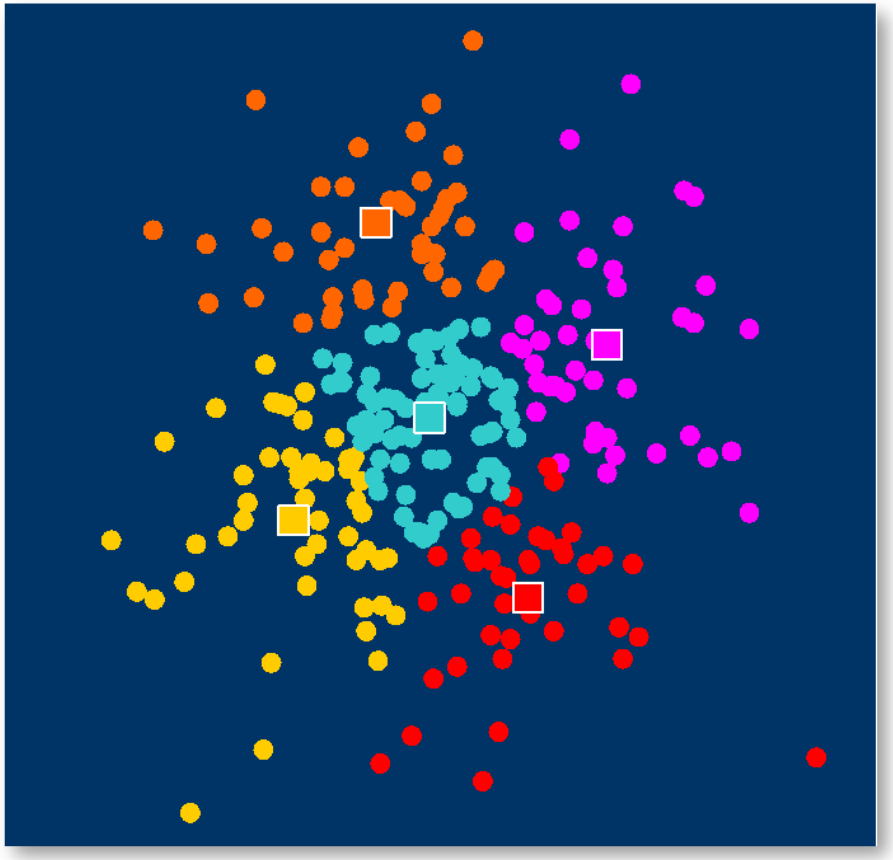
k-Means Clustering Algorithm



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

Segmentation Analysis



When no clusters exist, use the *k*-means algorithm to partition cases into contiguous groups.

The FASTCLUS Procedure

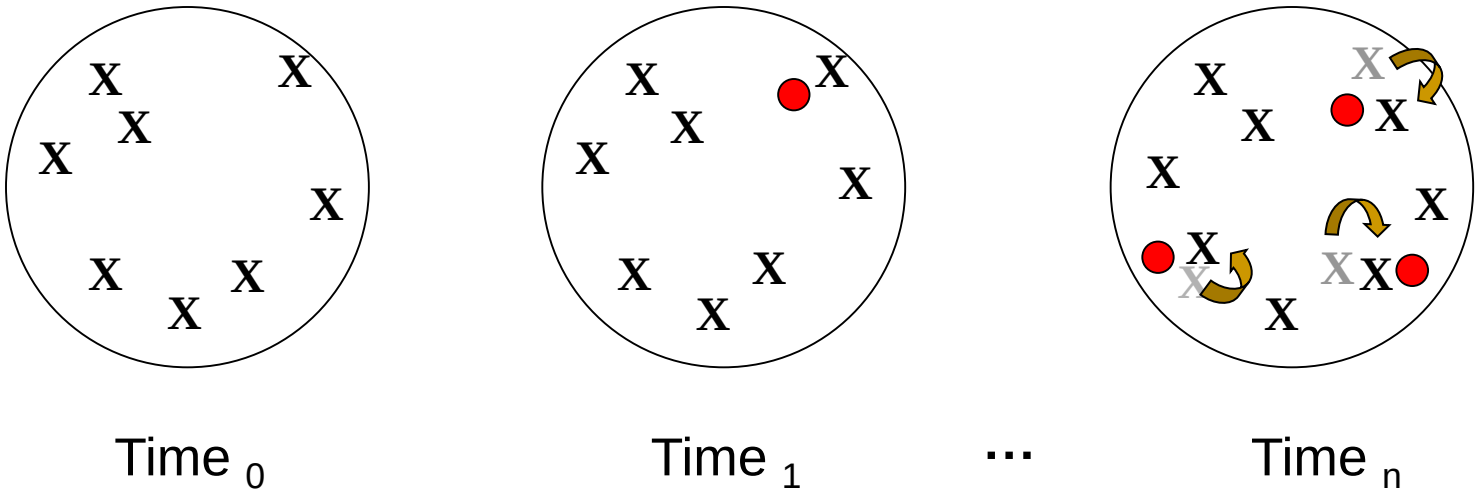
General form of the FASTCLUS procedure:

```
PROC FASTCLUS DATA=SAS-data-set
            <MAXC=>|<RADIUS=><options>;
      VAR variables;
RUN;
```

Because PROC FASTCLUS produces relatively little output, it is often a good idea to create an output data set, and then use other procedures such as PROC MEANS, PROC SGPLOT, PROC DISCRIM, or PROC CANDISC to study the clusters.

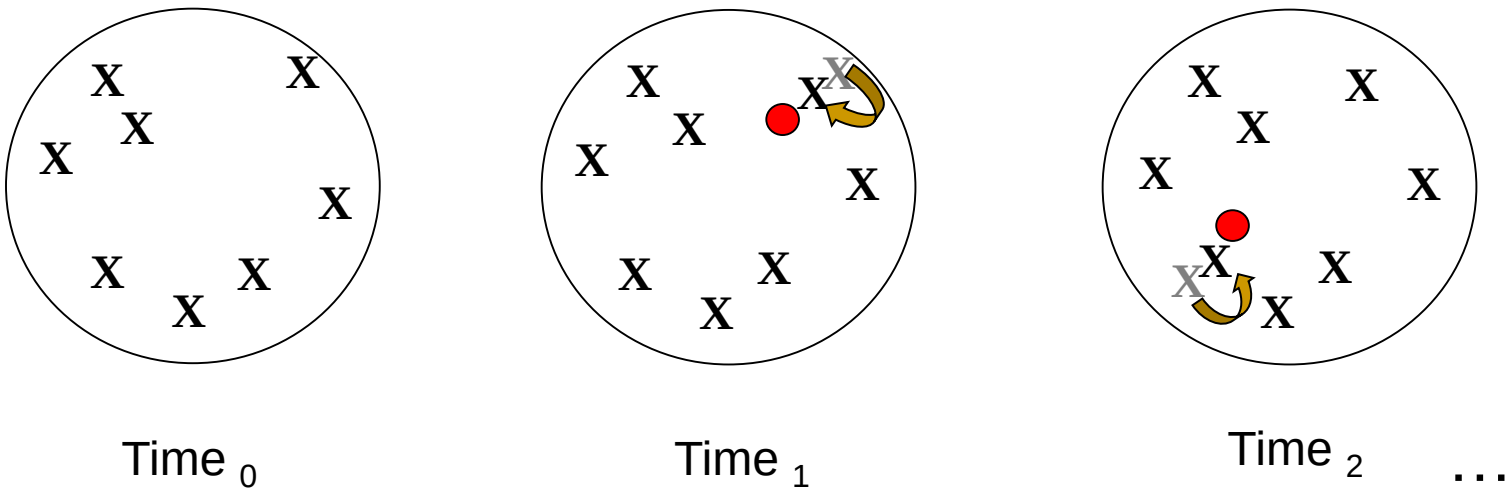
The MAXITER= Option

- The MAXITER= option sets the number of K-Means iterations (the default number of iterations is 1)



The DRIFT Option

The DRIFT option adjusts the nearest reference vector as each observation is assigned.



The LEAST= Option

The LEAST = option provides the argument for the Minkowski distance metric, changes the number of iterations, and changes the convergence criterion.

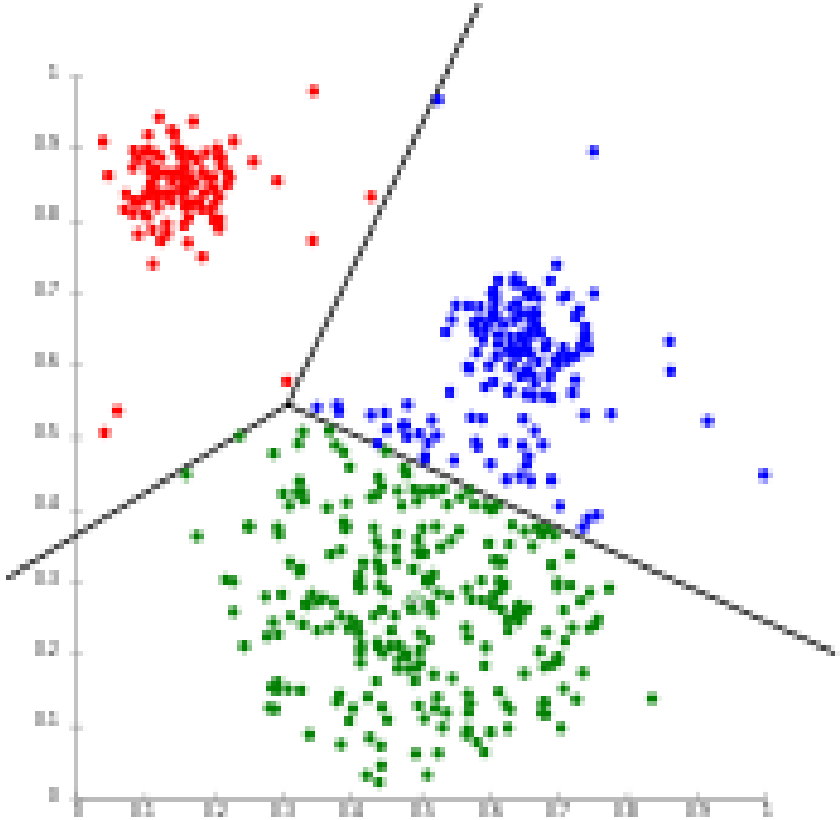
Option	Distance	Max Iterations	Converge=
default	EUCLIDEAN	1	.02
LEAST=1	CITY BLOCK	20	.0001
LEAST=2	EUCLIDEAN	10	.0001

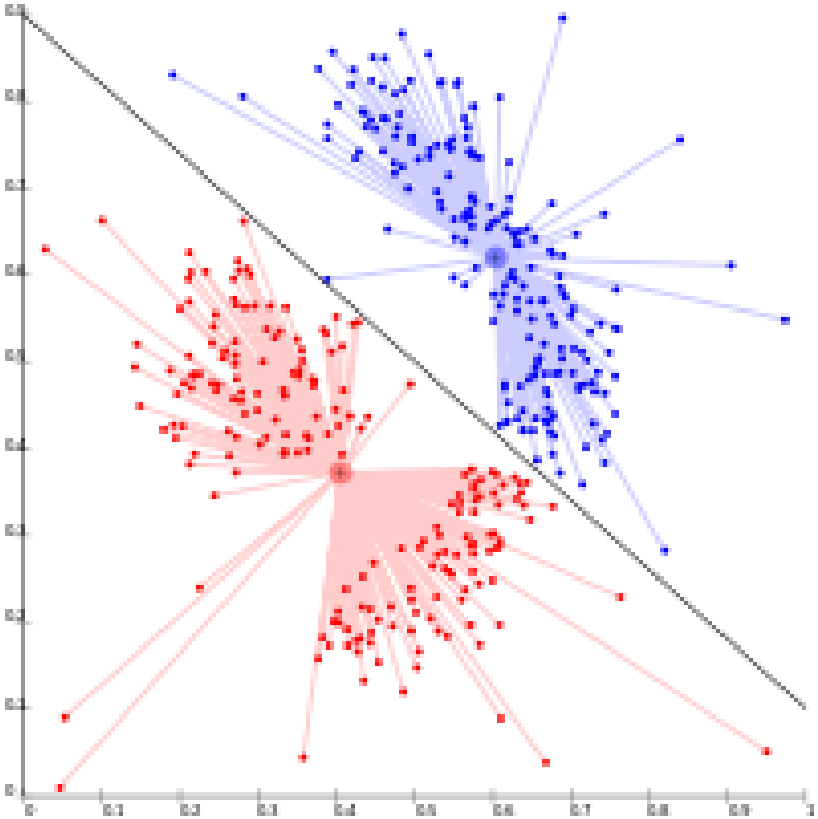
What Value of *k* to Use?

The number of seeds, *k*, typically translates to the final number of clusters obtained. The choice of *k* can be made using a variety of methods.

- Subject-matter knowledge (There are most likely five groups.)
- Convenience (It is convenient to market to three to four groups.)
- Constraints (You have six products and need six segments.)
- Arbitrarily (Always pick 20.)
- Based on the data (combined with Ward's method).

Problems with K-Means





Не всегда оптимальное разбиение пространства

Плотность выборки? Нет, не слышал!

Grocery Store Case Study: Census Data

Analysis goal:

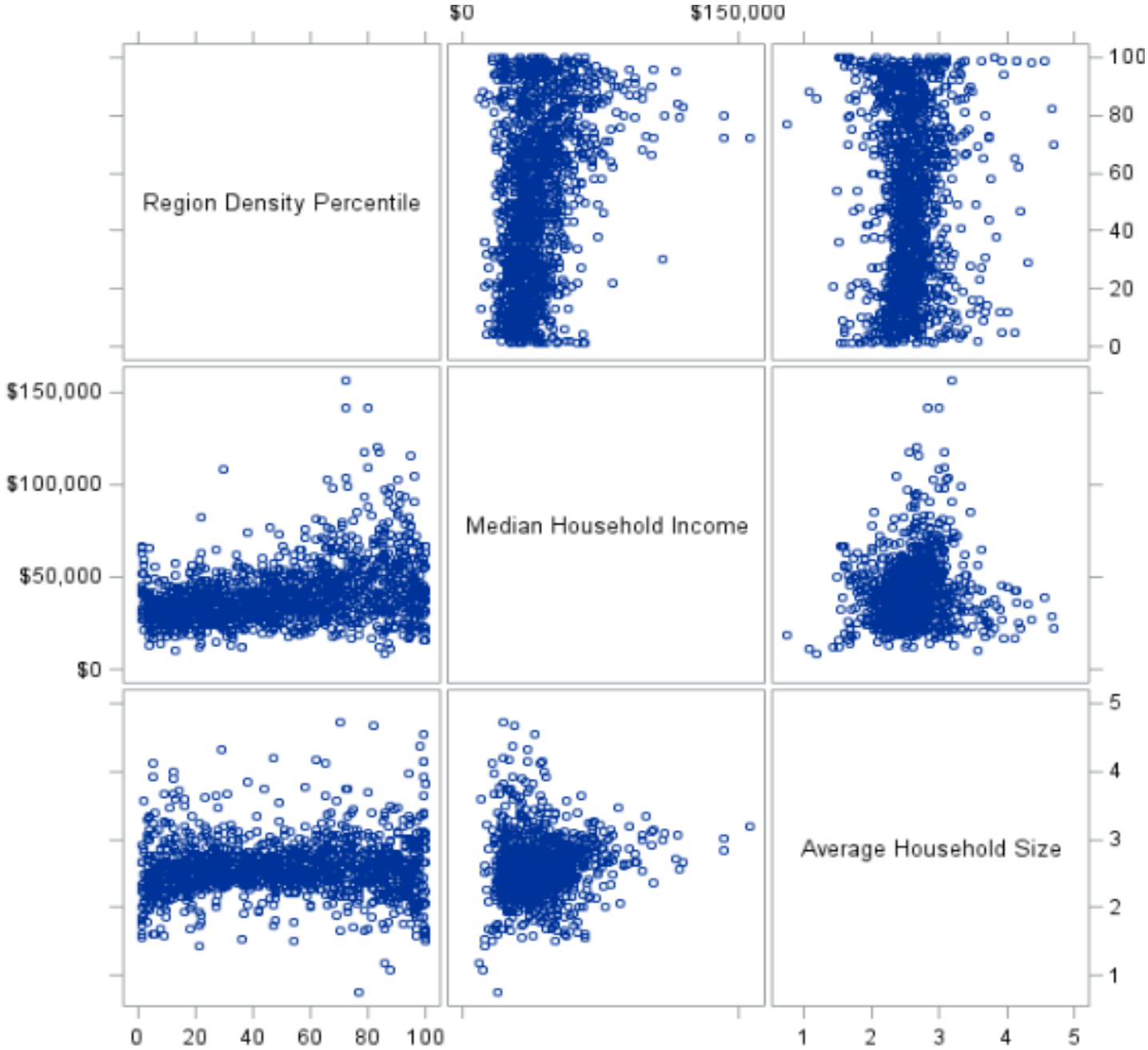
Where should you open new grocery store locations?
Group geographic regions based on income, household size, and population density.

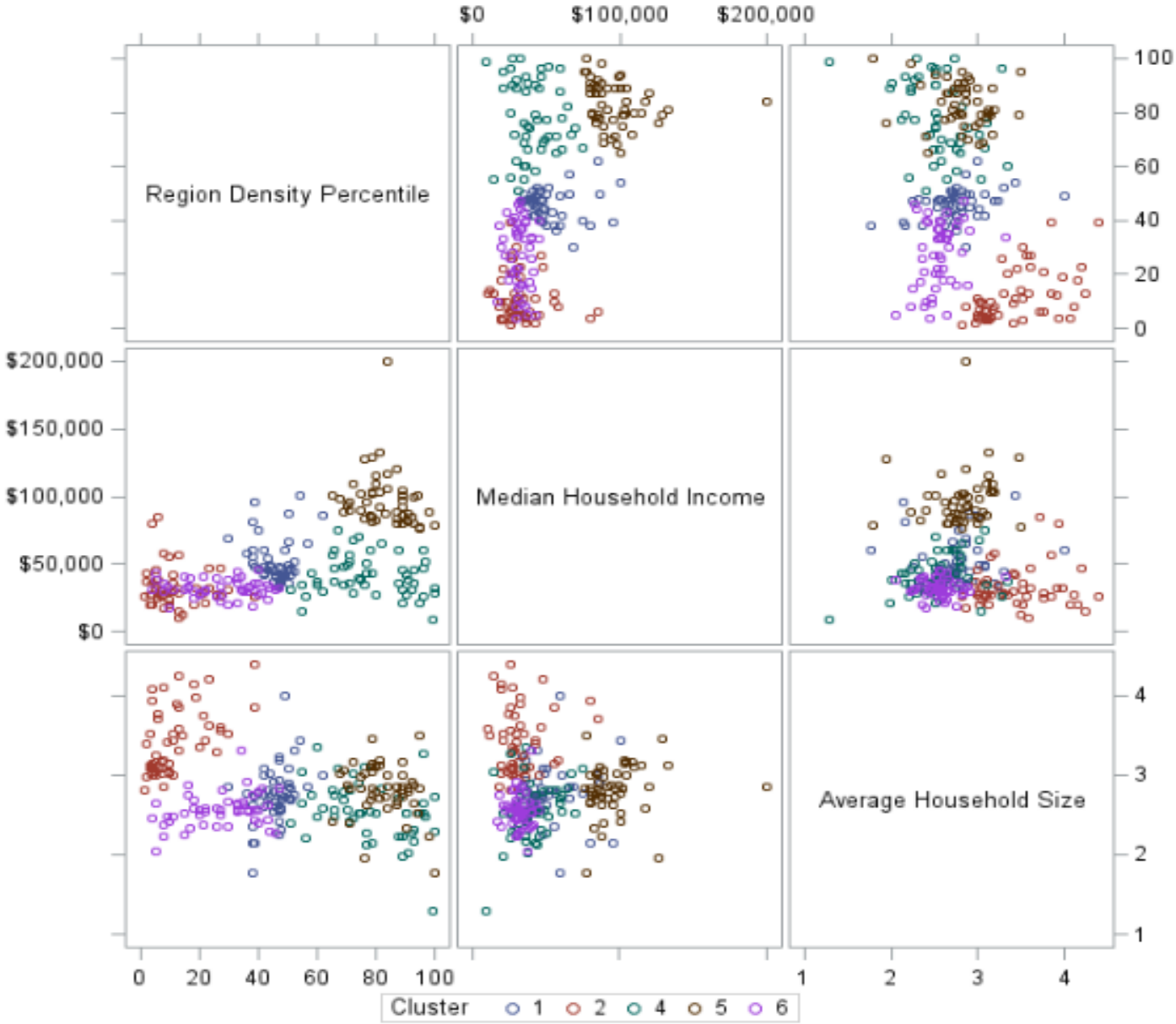
Analysis plan:

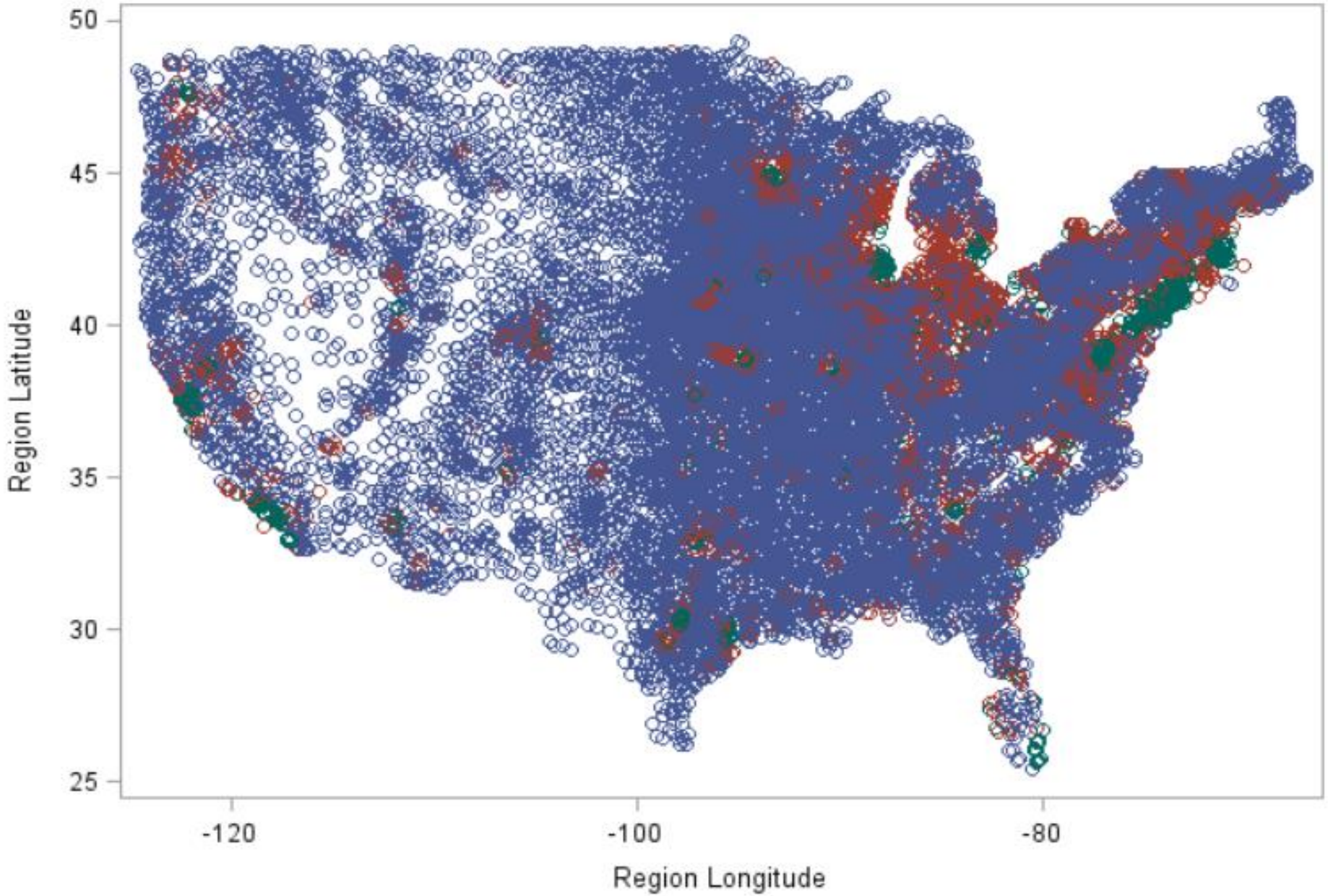
- **Explore the data.**
- **Select the number of segments to create.**
- **Create segments with a clustering procedure.**
- **Interpret the segments.**
- **Map the segments.**

K-Means Clustering for Segmentation

This demonstration illustrates the concepts discussed previously.


$0
$150,000
Region Density Percentile
100
80
60
40
20
0
$150,000
$100,000
$50,000
$0
Median Household Income
5
4
3
2
1
Average Household Size
0
20
40
60
80
100
1
2
3
4
5


Region Density Percentile
Median Household Income
Average Household Size
$0
$100,000
$200,000
$150,000
$50,000
0
20
40
60
80
100
1
2
3
4
Cluster
1
2
4
5
6

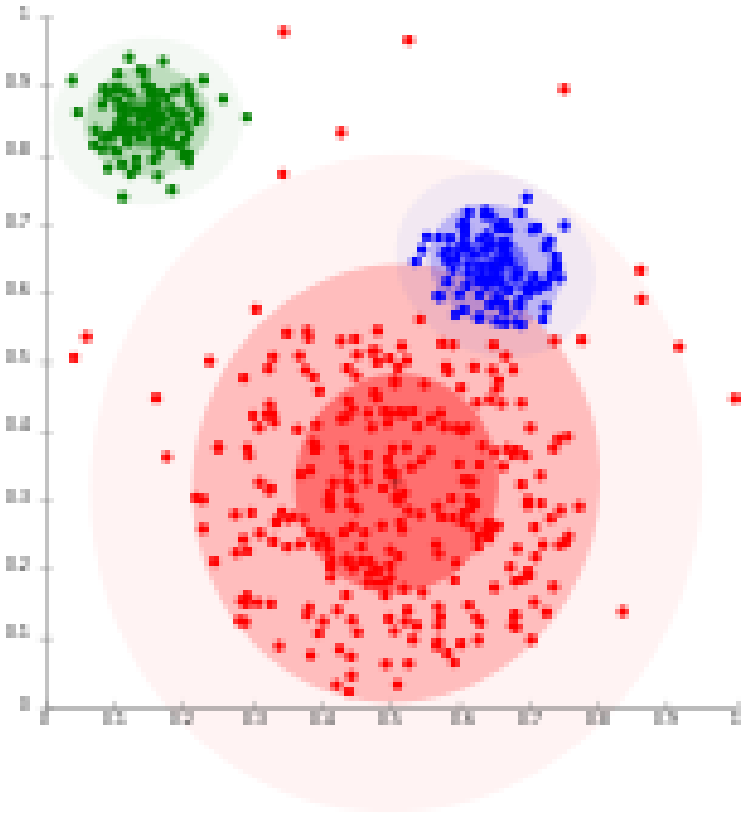

50
45
40
35
30
25
Region Latitude
-120
-100
-80
Region Longitude

Partitive Clustering

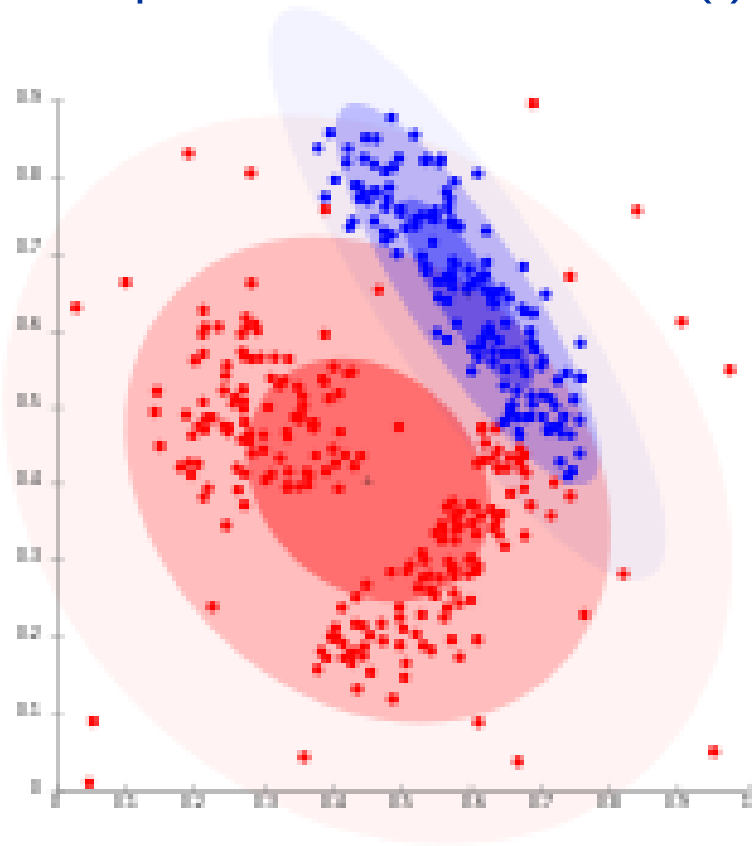
Непараметрическая кластеризация

Parametric vs Non-Parametric Clustering

Expectation-Maximization (+)



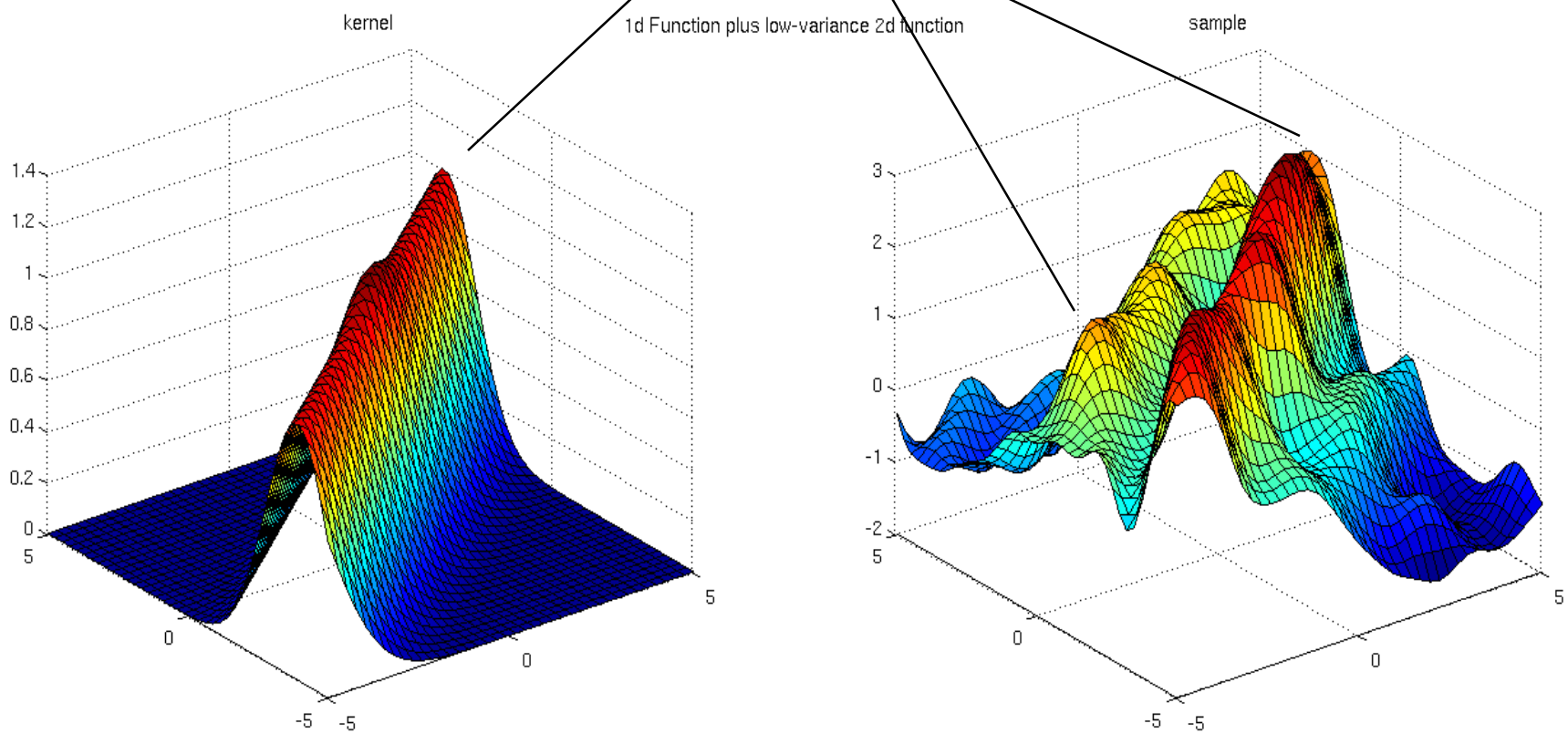
Expectation-Maximization (-)



Параметрические алгоритмы плохи на density-based кластерах

Developing Kernel Intuition

Modes



Advantages of Nonparametric Clustering

- It still obtains good results on compact clusters.
- It is capable of detecting clusters of unequal size and dispersion, even if they have irregular shapes.
- It is less sensitive (but not insensitive) to changes in scale than most clustering methods.
- It does not require that you guess the number of clusters present in the data.

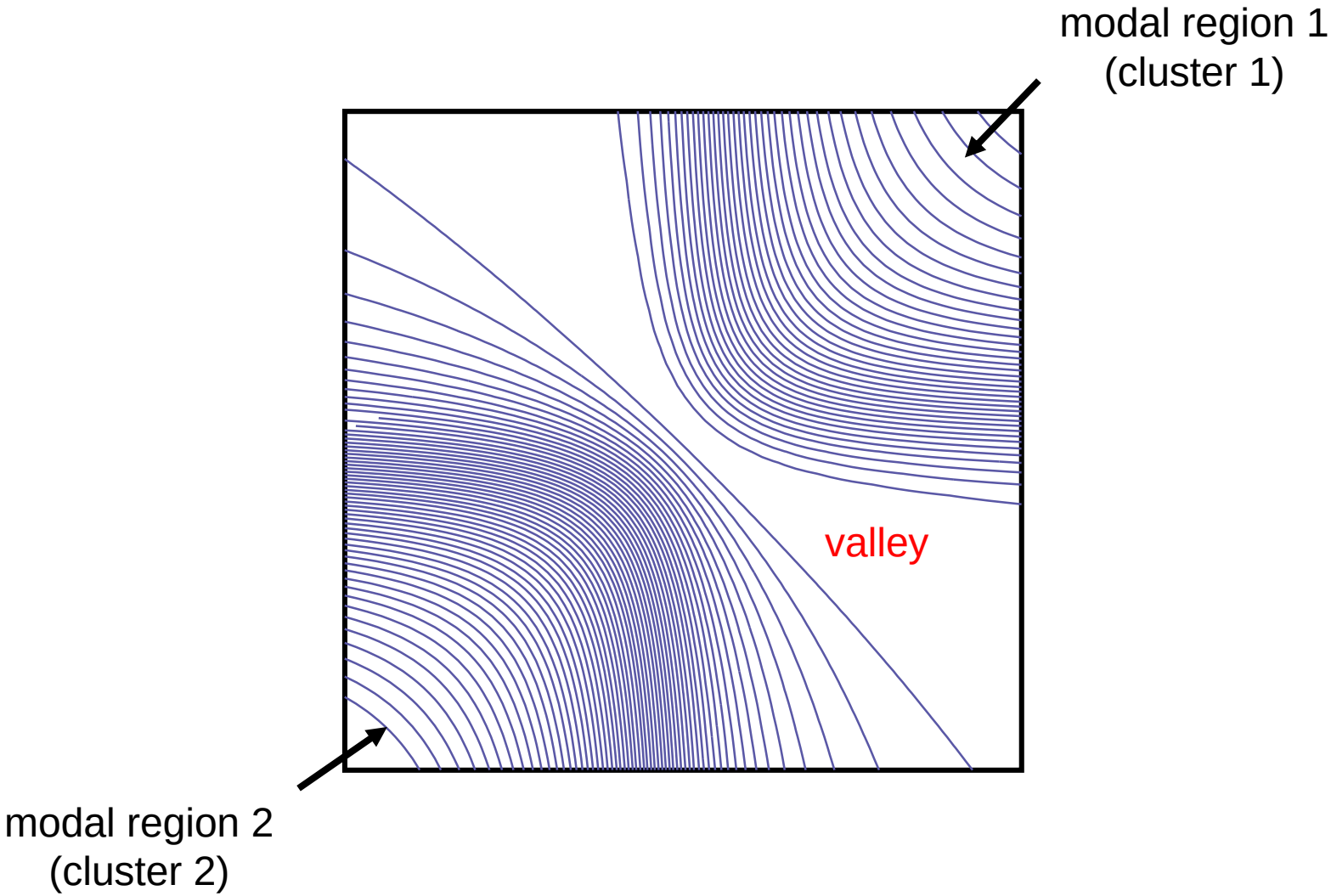
```
PROC MODECLUS DATA=SAS-data-set
        METHOD=method <options>;
      VAR variables;
RUN;
```

Significance Tests

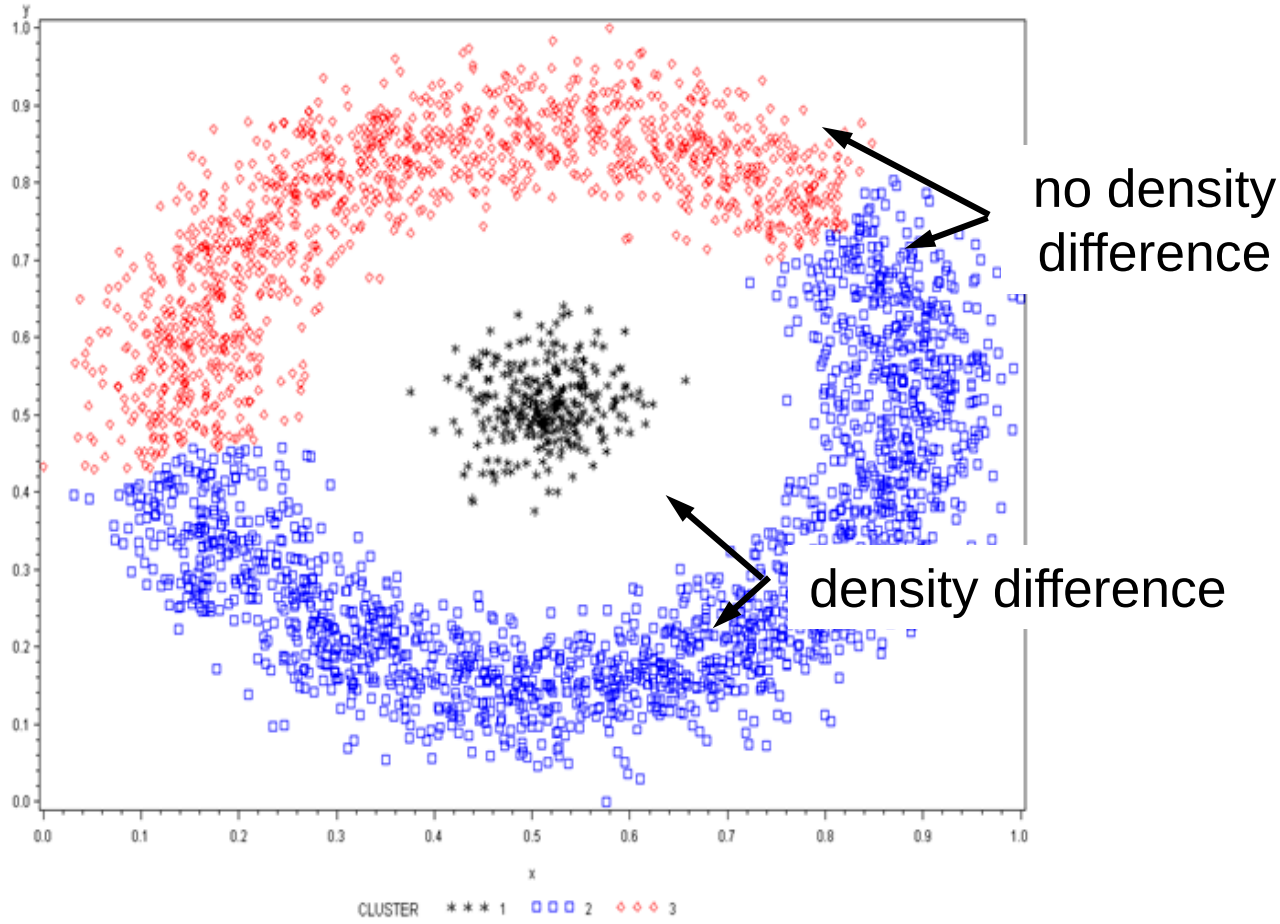
- If requested (the JOIN= option), PROC MODECLUS can hierarchically join non-significant clusters.

- Although a fixed-radius kernel (R=) must be specified, the choice of smoothing parameter is not critical.

Valley-Seeking Method

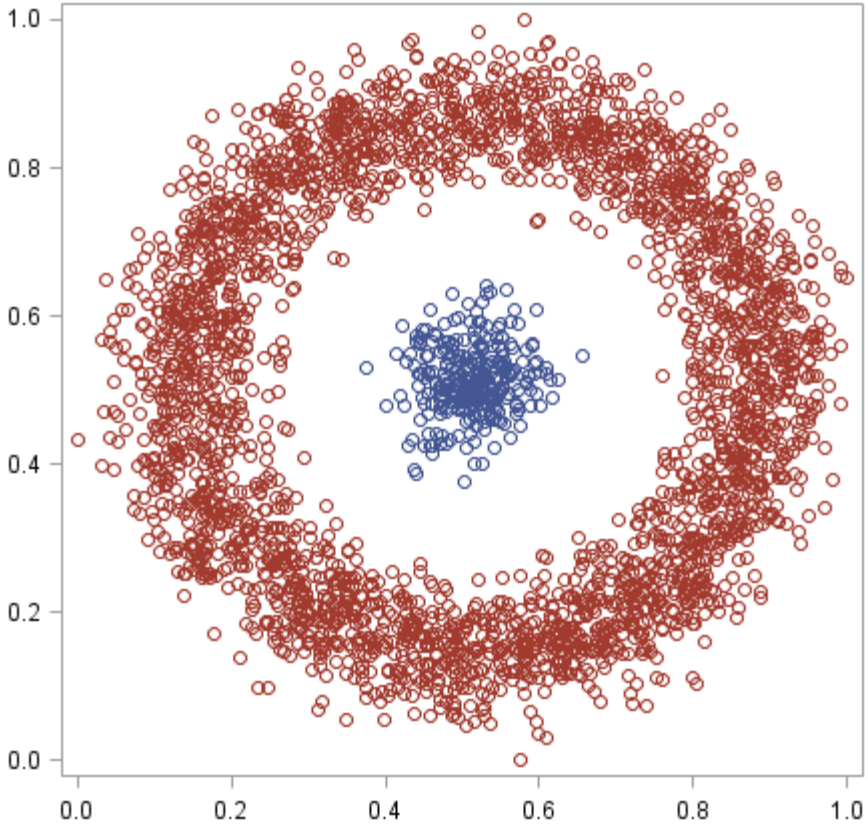


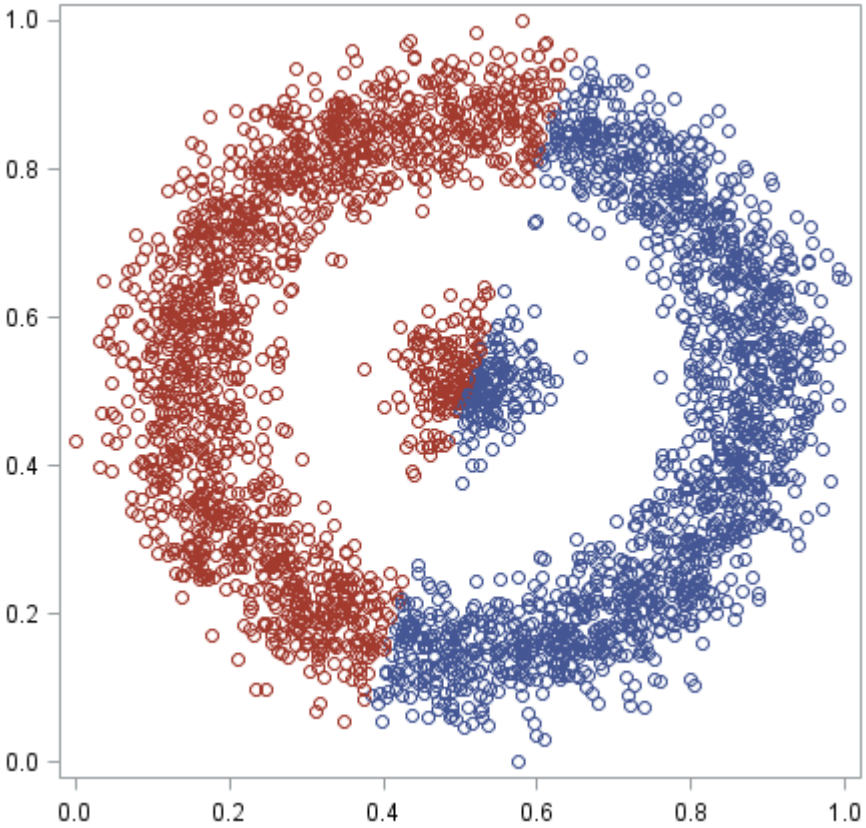
Saddle Density Estimation

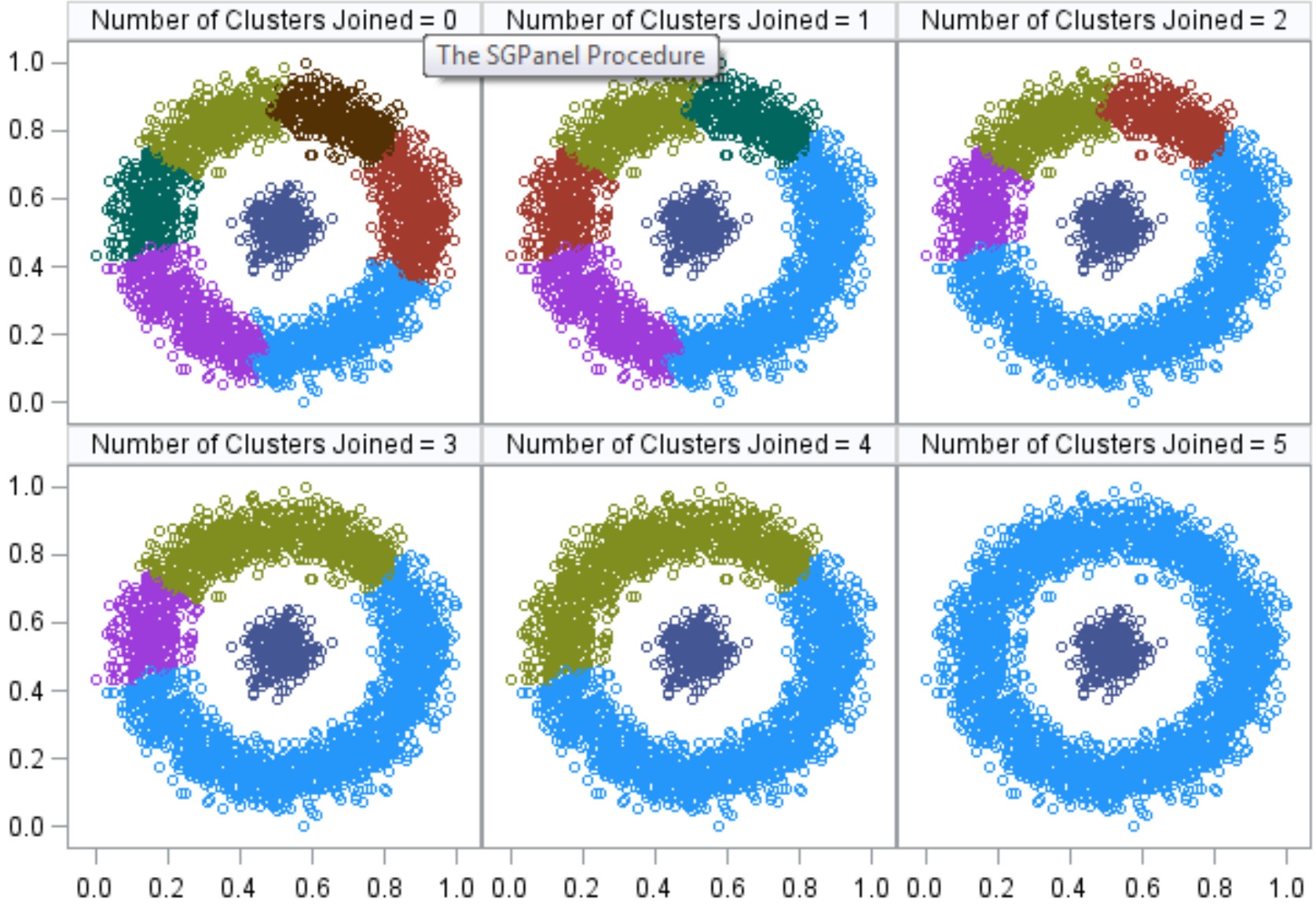


Hierarchically Joining Non-Significant Clusters

This demonstration illustrates the concepts discussed previously.

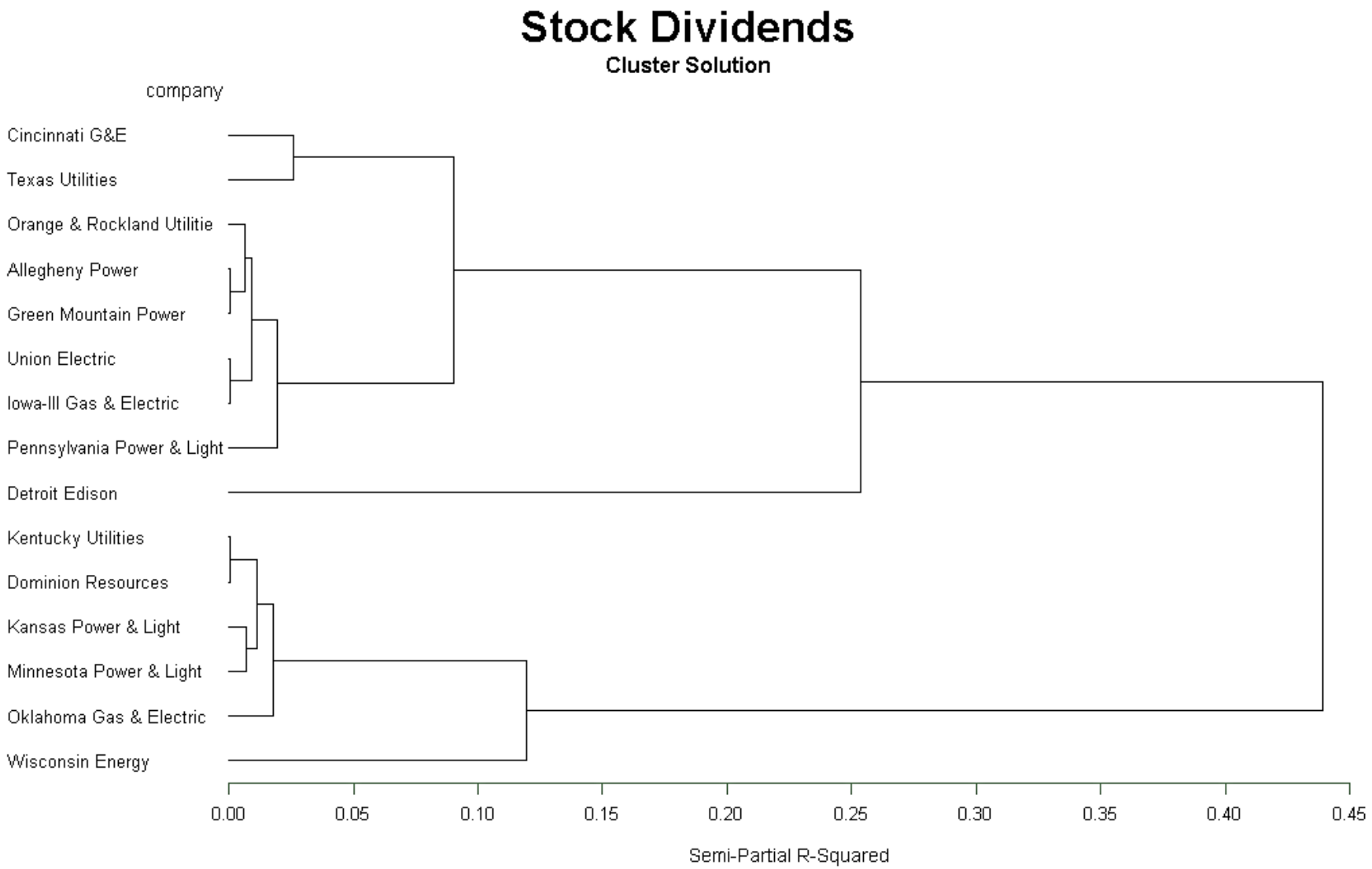

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Number of Clusters Joined = 0
Number of Clusters Joined = 1
Number of Clusters Joined = 2
Number of Clusters Joined = 3
Number of Clusters Joined = 4
Number of Clusters Joined = 5
The SGPanel Procedure
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0

Иерархическая кластеризация

Hierarchical Clustering



The CLUSTER Procedure

General form of the CLUSTER procedure:

```
PROC CLUSTER DATA=SAS-data-set
        METHOD=method <options>;
    VAR variables;
    FREQ variable;
    RMSSTD variable;
RUN;
```

The required METHOD= option specifies the hierarchical technique to be used to cluster the observations.

Cluster and Data Types

Hierarchical Method	Distance Data Required?
Average Linkage	Yes
Two-Stage Linkage	Some Options
Ward's Method	Yes
Centroid Linkage	Yes
Complete Linkage	Yes
Density Linkage	Some Options
EML	No
Flexible-Beta Method	Yes
McQuitty's Similarity	Yes
Median Linkage	Yes
Single Linkage	Yes

The TREE Procedure

General form of the TREE procedure:

```
PROC TREE DATA=<dendrogram> <options>;
RUN;
```

The TREE procedure either

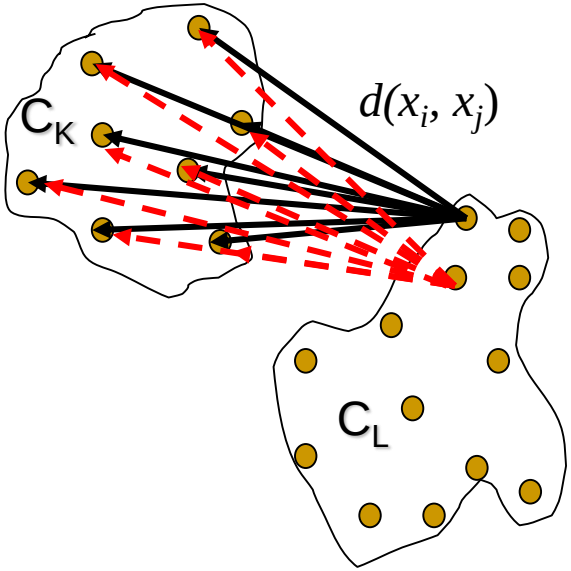
- displays the dendrogram (LEVEL= option), or
- assigns the observations to a specified number of clusters (NCLUSTERS= option).

Иерархическая кластеризация

Параметры алгоритма

Average Linkage

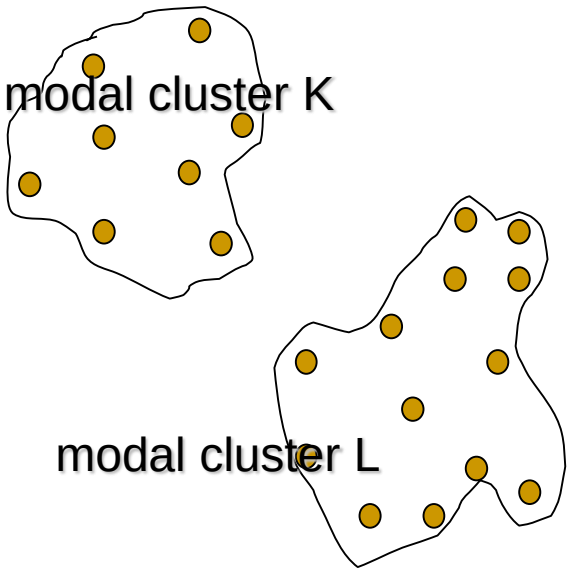
The distance between clusters is the average distance between pairs of observations.



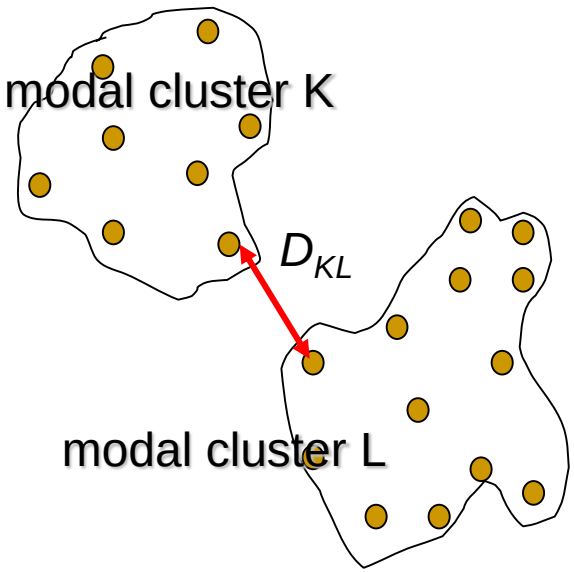
$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d\left(x_i, x_j\right)$$

Two-Stage Density Linkage

A nonparametric density estimate is used to determine distances, and recover irregularly shaped clusters.



1. Form ‘modal’ clusters



2. Apply single linkage

The Two Stages of Two-stage

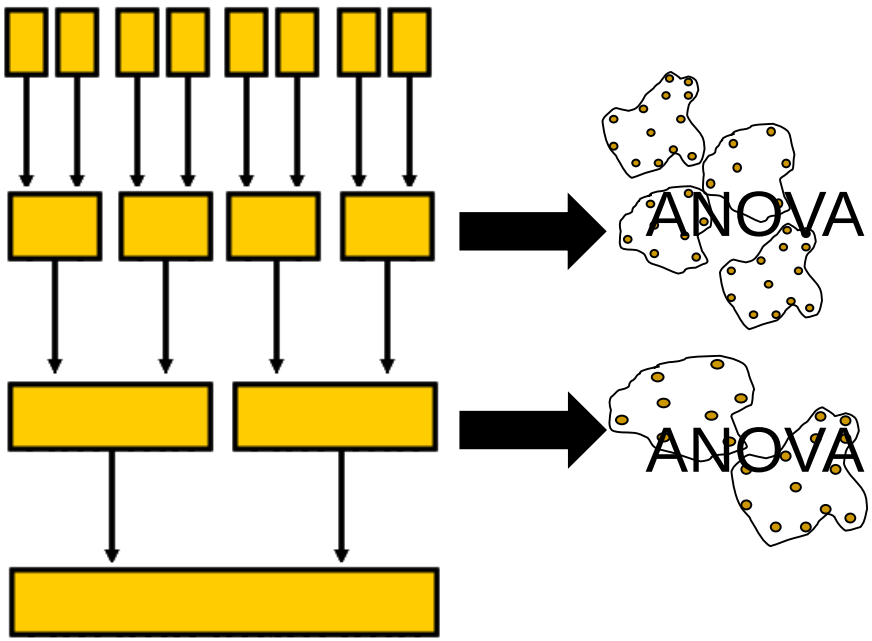
- ✓ The first stage, known as *density linkage*, constructs a distance measure, d*, based on kernel density estimates and creates modal clusters.

- ✓ The second stage ensures that a cluster has at least "***n***" members before it can be fused. Clusters are fused using *single linkage* (joins based on the nearest points between two clusters).

- ✓ The measure d* can be based on three methods. This course uses the *k-nearest neighbor m*ethod.

Ward's

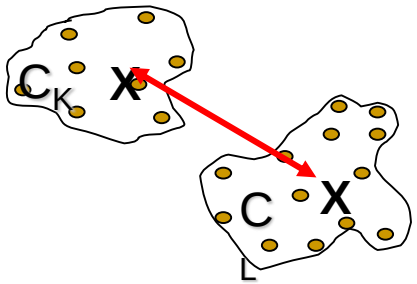
Ward's method uses ANOVA at each fusion point to determine if the proposed fusion is warranted.



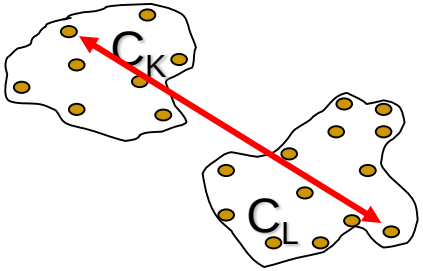
$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L} \right)}$$

Additional Clustering Methods

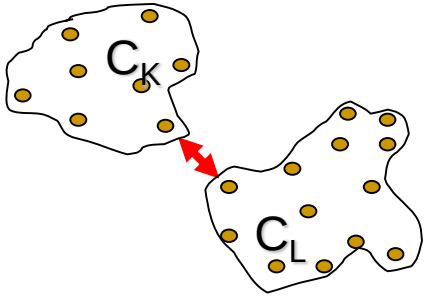
Centroid Linkage



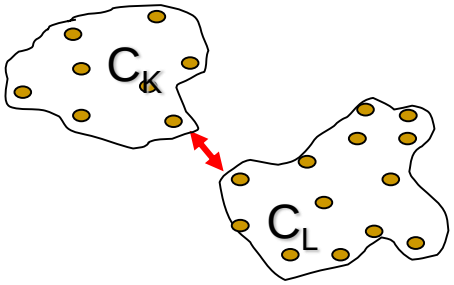
Complete Linkage



Density Linkage

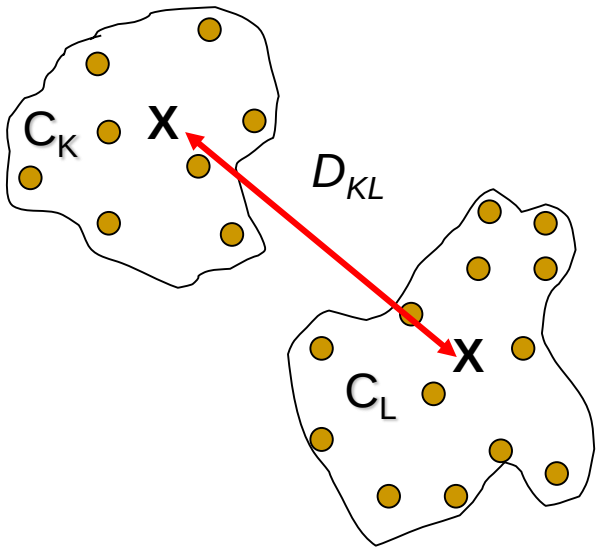


Single Linkage



Centroid Linkage

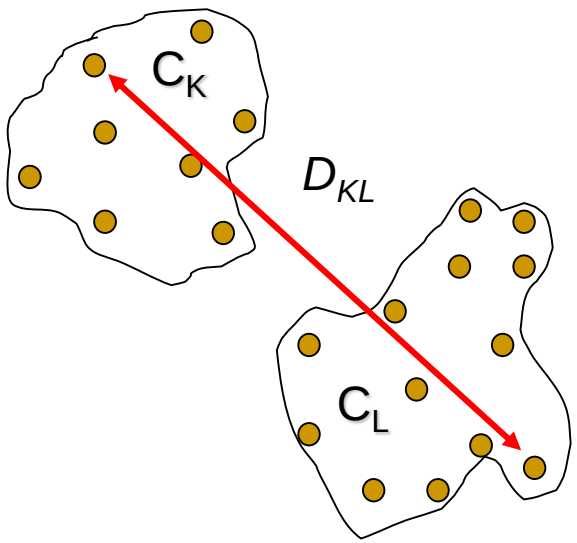
The distance between clusters is the squared Euclidean distance between cluster centroids $\bar{\mathbf{x}}_K$ and $\bar{\mathbf{x}}_L$



$$D_{KL} = \left\| \bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L \right\|^2$$

Complete Linkage

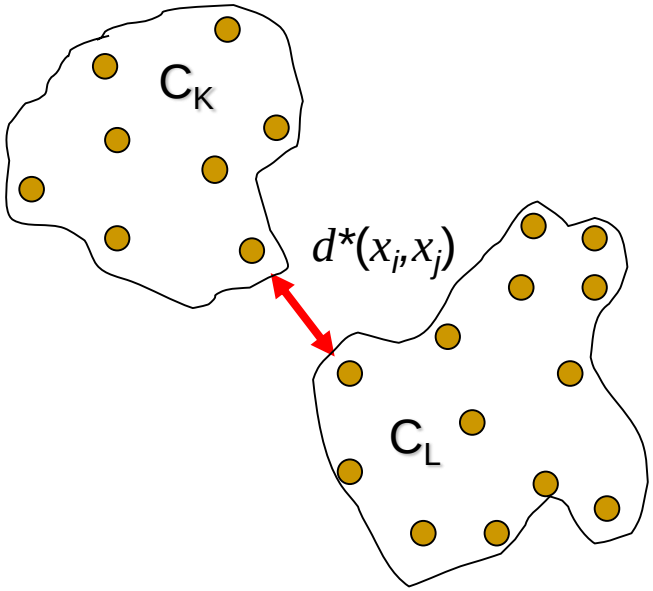
The distance between clusters is the maximum distance between two observations, one in each cluster.



$$D_{KL} = i \in C_K \, j \in C_L \max d(x_i, x_j)$$

Density Linkage

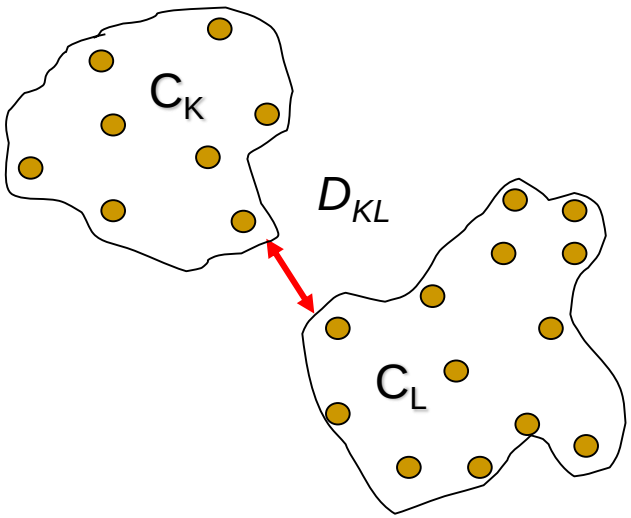
1. Calculate a new distance metric, *d**, using *k*-nearest neighbor, uniform kernel, or Wong's hybrid method.
2. Perform single linkage clustering with *d**.



$$d^*(x_i, x_j) = \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right)$$

Single Linkage

The distance between clusters is the distance between the two nearest observations, one in each cluster.



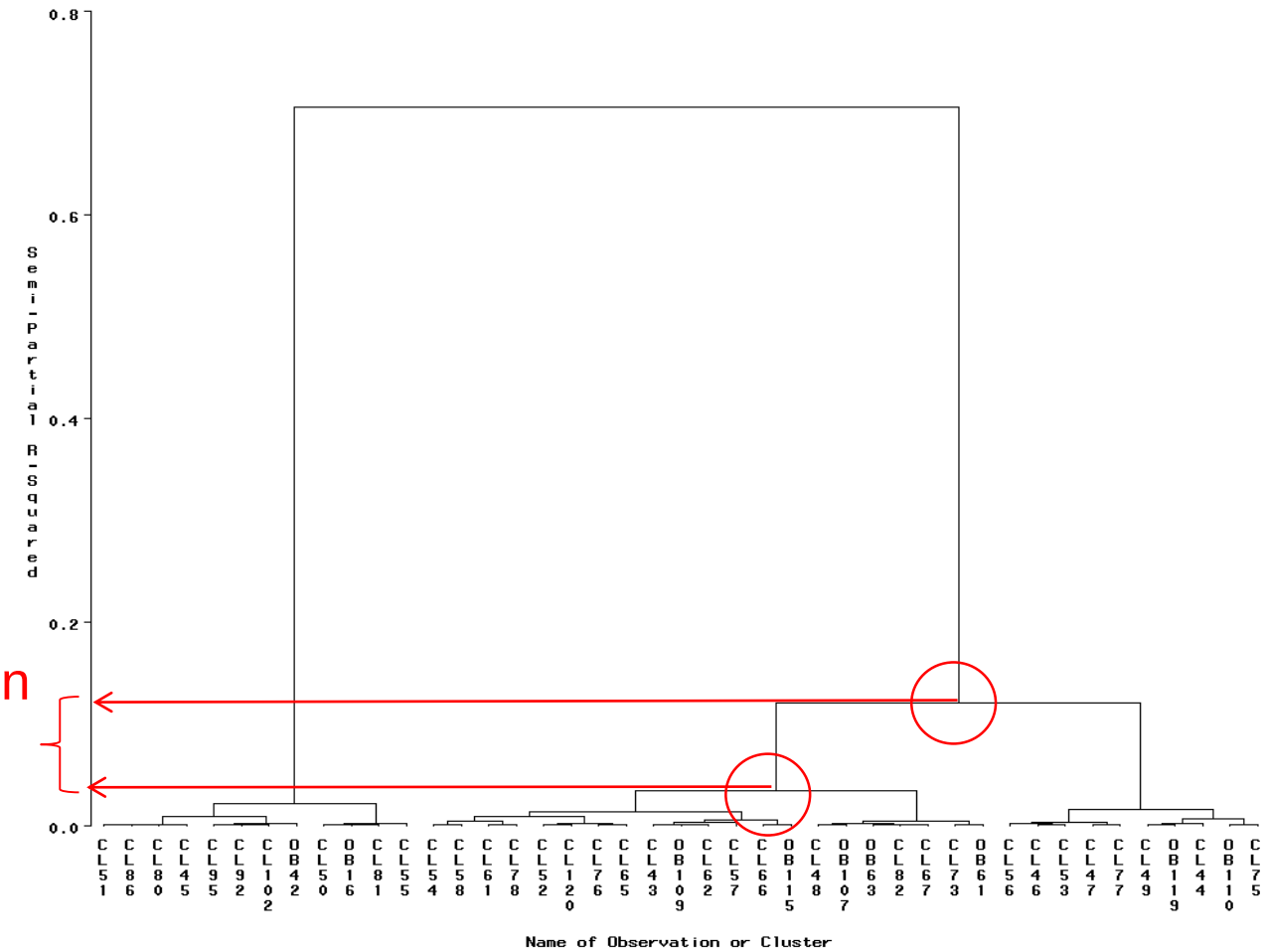
$$D_{KL} = i \in C_K \;\; j \in C_L \;\; \min d(x_i, x_j)$$

Оценка результатов кластеризации

Оптимальное количество кластеров

Interpreting Dendrograms

For interpreting **any** hierarchical clustering method



change in fusion level; prefer 3 clusters.

Cubic Clustering Criterion

$$\mathrm{CCC} = \ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{\left(0.001+E(R^2)\right)^{1.2}}$$

- Sarle's Cubic Clustering Criterion compares observed and expected R^2 values.
- It tests the null hypothesis (H_0) that the data was sampled from uniform distribution across a hyper-box.
- CCC values greater than 2 suggest there is sufficient evidence of cluster structure (reject the H_0).
- *Join clusters in local **MAXIMA** of CCC*

Other Useful Statistics

- Pseudo-F Statistics

$$\mathrm{PSF} = \frac{\mathbf{B}/(g-1)}{\mathbf{W}/(n-g)} \longrightarrow$$

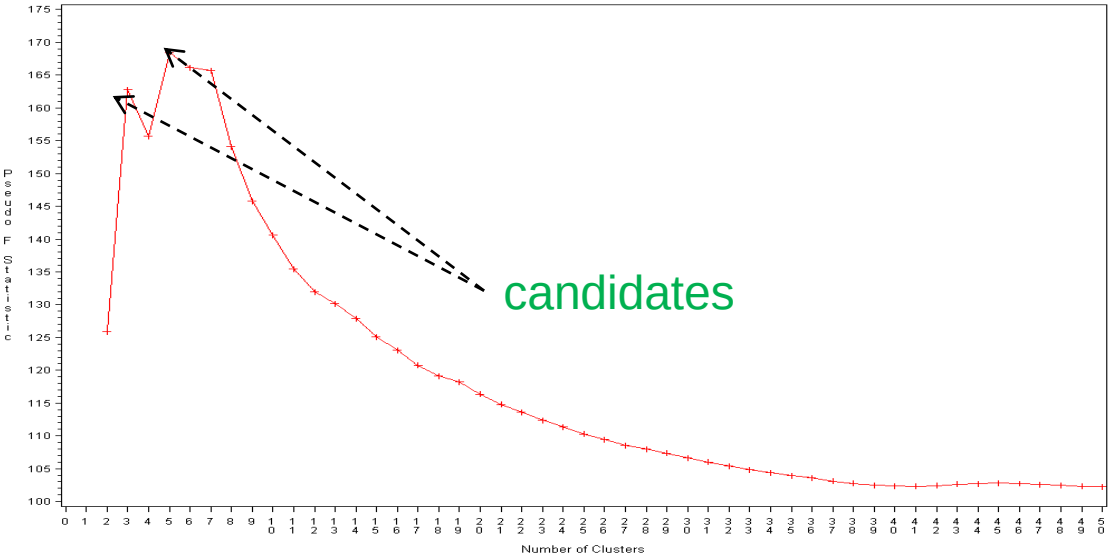
Join clusters if statistics is in local ***MAXIMUM***

- Pseudo-T2 Statistics

$$\mathrm{PST2} = \frac{\mathbf{W}_m - \mathbf{W}_k - \mathbf{W}_l}{\left(\mathbf{W}_k - \mathbf{W}_l\right) - (n_k + n_l - 2)}$$

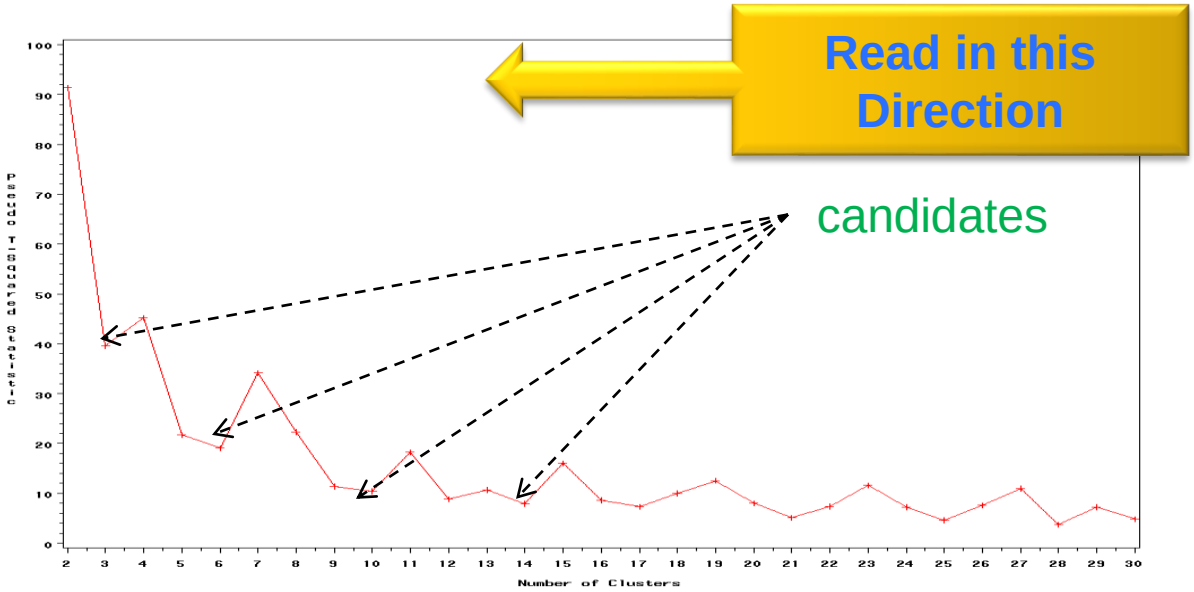
Join clusters if T2 statistics is in local ***MINIMUM***

Interpreting PSF and PST2



Pseudo-F Statistics

Pseudo-T2 Statistics



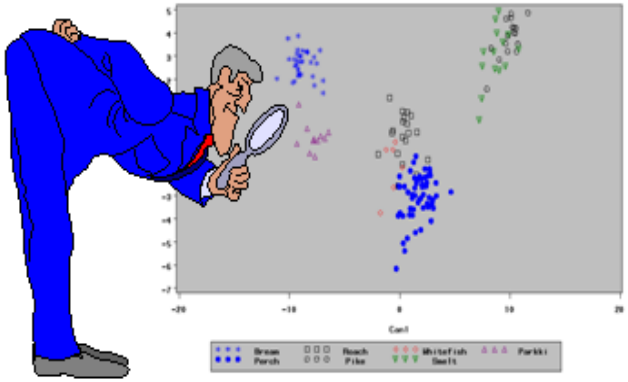
Оценка результатов кластеризации

Профилирование кластеров

Cluster Profiling

- Generation of unique cluster descriptions from the input variables.
- It can be implemented using many approaches:
  - Generate the "typical" member of each cluster.
  - Use ANOVA to determine the inputs that uniquely define each of the typical members.
  - Use graphs to compare and describe the clusters

- In addition, one can compare each cluster *against* the whole cluster population

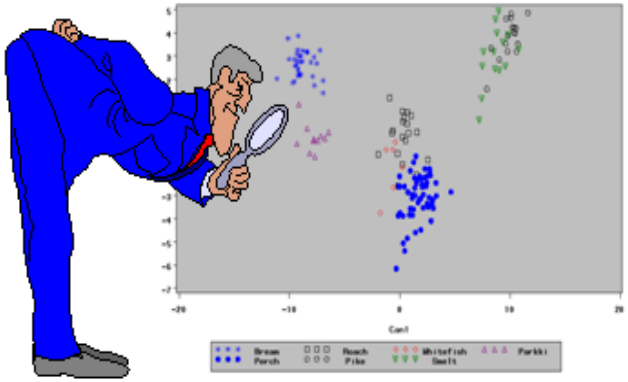


One-Against-All Comparison

1. For the cluster k classify each observation as being a member of cluster *k* (with a value of 1) or not a member of cluster *k* (with a value of 0)

2. Use logistic regression to rank the input variables by their ability to distinguish cluster *k* from the others

3. Generate a comparative plot of cluster *k* and the rest of the data.



Оценка результатов кластеризации

Применение модели кластеризации к новым наблюдениям

Scoring PROC FASTCLUS Results

1. Perform cluster analysis and save the centroids.

```
PROC FASTCLUS OUTSTAT=centroids;
```

2. Load the saved centroids and score a new file.

```
PROC FASTCLUS INSTAT=centroids
              OUT=SAS-data-set;
```

Scoring PROC CLUSTER Results

1. Perform the hierarchical cluster analysis.

```
PROC CLUSTER METHOD= OUTTREE=tree;
      VAR variables;
RUN;
```

2. Generate the cluster assignments.

```
PROC TREE DATA=tree N=nclusters OUT=treeout;
RUN;
```


continued...

Scoring PROC CLUSTER Results

3. Calculate the cluster centroids.

```
PROC MEANS DATA=treeout;
        CLASS cluster;
        OUTPUT MEAN= OUT=centroids;
RUN;
```

4. Read the centroids and score the new file.

```
PROC FASTCLUS DATA=newdata SEED=centroids
        MAXCLUSTERS=n MAXITER=0 OUT=results;
RUN;
```

Кейс

Happy Household Study

The Happy Household Catalog

A retail catalog company with a strong online presence monitors quarterly purchasing behavior for its customers, including sales figures summarized across departments and quarterly totals for 5.5 years of sales.

- HH wants to improve customer relations by tailoring promotions to customers based on their preferred type of shopping experience
- Customer preferences are difficult to ascertain based solely on opportunistic data.

Cluster Analysis as a Predictive Modeling Tool

The marketing team gathers questionnaire data:

- Identify patterns in customer attitudes toward shopping
- Generate attitude profiles (clusters) and tie to specific marketing promotions
- Use attitude profiles as the target variable in a predictive model with shopping behavior as inputs
- Score large customer database (n=48K) using the predictive model, and assign promotions based on predicted cluster groupings

Preparation for Clustering

1. Data and Sample Selection
2. Variable Selection (*What characteristics matter?*)
3. Graphical Exploration (*What shape/how many clusters?*)
4. Variable Standardization (*Are variable scales comparable?*)
5. Variable Transformation (*Are variables correlated? Are clusters elongated?*)

Data and Sample Selection

A study is conducted to identify patterns in customer attitudes toward shopping

Online customers are asked to complete a questionnaire during a visit to the company's retail Web site. A sample of 200 complete data questionnaires is analyzed.

Preparation for Clustering

1. Data and Sample Selection (*Who am I clustering?*)
2. Variable Selection
3. Graphical Exploration (*What shape/how many clusters?*)
4. Variable Standardization (*Are variable scales comparable?*)
5. Variable Transformation (*Are variables correlated? Are clusters elongated?*)

Variable Selection

This demonstration illustrates the concepts discussed previously.

clus06d01.sas

What Have You Learned?

Three variables will be used for cluster analysis:

HH5 → I prefer to shop online rather than offline

HH10 → I believe that good service is the most important thing a company can provide

HH11 → Good value for the money is hard to find

Preparation for Clustering

1. Data and Sample Selection (*Who am I clustering?*)
2. Variable Selection (*What characteristics matter?*)
3. Graphical Exploration
4. Variable Standardization (*Are variable scales comparable?*)
5. Variable Transformation (*Are variables correlated? Are clusters elongated?*)

Graphical Exploration of Selected Variables

This demonstration illustrates the concepts discussed previously.

clus06d02.sas

Preparation for Clustering

1. Data and Sample Selection (*Who am I clustering?*)
2. Variable Selection (*What characteristics matter?*)
3. Graphical Exploration (*What shape/how many clusters?*)
4. Variable Standardization
5. Variable Transformation

What Have You Learned?

- ✓ Standardization is unnecessary in this example because all variables are on the same scale of measurement

- ✓ Transformation might be unnecessary in this example because there is not evidence of elongated cluster structure from the plots, and the variables have low correlation.

Selecting a Clustering Method

- With 200 observations, it is a good idea to use a hierarchical clustering technique.
- Ward's method is selected for ease of interpretation
- Select number of clusters with CCC, PSF and PST2
- Use cluster plots to assist in providing cluster labels

Hierarchical Clustering and Determining the Number of Clusters

This demonstration illustrates the concepts discussed previously.

clus06d03.sas

Profiling the Clusters

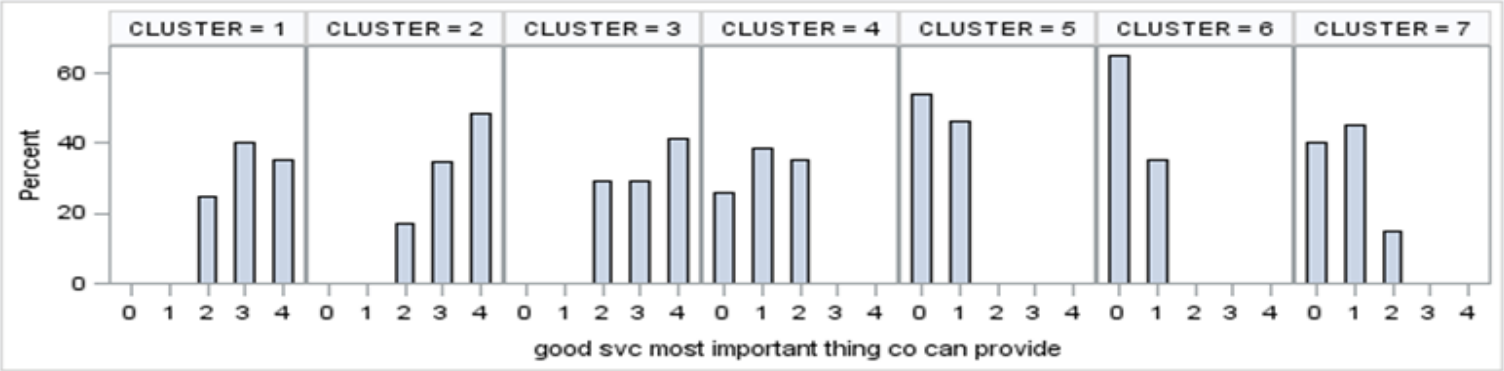
- There are seven clusters
- There are three marketing promotions
- Determine whether the seven cluster profiles are good complements to the three marketing promotions
- Otherwise try another number of clusters

Profiling the Seven-Cluster Solution

This demonstration illustrates the concepts discussed previously.

clus06d04.sas

What Have You Learned?



What Have You Learned?

Cluster 1. High on HH5, HH10.	Label: Discriminating online taste
Cluster 2. High on HH10 and HH11.	Label: Savings and service
Cluster 3. Low on HH5 and HH11. High on HH10.	Label: Values in-store service
Cluster 4. Low on HH5, HH10, High on HH11	Label: Seeks in-store savings
Cluster 5. High on HH5, Low on HH10 and HH11	Label: Reluctant shopper, on-line
Cluster 6. Low on HH5, HH10, HH11	Label: Reluctant shopper, in-store
Cluster 7: High on HH5, HH11, Low on HH10	Label: Seeks on-line savings

What Will You Offer?

Offer 1: Coupon for free shipping if > 6mo since last purchase

Offer 2: Fee-based membership in exclusive club to get “valet” service, personal (online) shopper.

Offer 3: Coupon for product of a brand different from previously purchased.

1. Discriminating online tastes
2. Savings and service anywhere
3. Values in-store service
4. Seeks in-store savings
5. Reluctant shopper, online
6. Reluctant shopper, in-store
7. Seeks on-line savings

What Will You Offer?

Offer 1: Coupon for free shipping if > 6mo since last purchase

Offer 2: Fee-based membership in exclusive club to get “valet” service, personal (online) shopper.

Offer 3: Coupon for product of a brand different from previously purchased.

1. Discriminating online tastes
2. **Savings and service anywhere**
3. Values in-store service
4. **Seeks in-store savings**
5. Reluctant shopper, online
6. Reluctant shopper, in-store
7. **Seeks on-line savings**

Offer will be made based on cluster classification and a high customer lifetime value score.

Predictive Modeling

The marketing team can choose from a variety of predictive modeling tools, including logistic regression, decision trees, neural networks, and discriminant analysis

Logistic regression and NN should be neglected because of the small sample and large number of input variables

Discriminant analysis is used in this example

```
PROC DISCRIM DATA=data-set-1;
   <PRIORS priors-specification;>
   CLASS cluster-variable;
   VAR input-variables;
RUN;
```

Modeling Cluster Membership

This demonstration illustrates the concepts discussed previously.

clus0605.sas

Scoring the Database

Once a model has been developed to predict cluster membership from purchasing data, the full customer database can be scored.

Customers are offered specific promotions based on predicted cluster membership.

```
PROC DISCRIM DATA=data-set-1
        TESTDATA=data-set-2 TESTOUT=scored-data;
   PRIORS priors-specification;
   CLASS cluster-variable;
   VAR input-variables;
RUN;
```

Let's Cluster the World!