Subject

Gaussian mixture model based clustering with TANAGRA: the EM algorithm.

In the **Gaussian mixture model-based clustering**, each cluster is represented by a Gaussian distribution. The entire dataset is modeled by a mixture (a linear combination) of these distributions.

The **EM (Expectation Maximization) algorithm** is used in practice to find the "optimal" parameters of the distributions that maximize the likelihood function.

The number of clusters is a parameter of the algorithm. But we can also detect the "optimal" number of clusters by evaluating several values, i.e. testing 1 cluster, 2 clusters, etc. and choosing the best one (which maximizes the likelihood or another criterion such as AIC or BIC).

Dataset

We use a synthetic dataset in a two dimensional space[1]. We aim to discover two clusters (Figure 1).

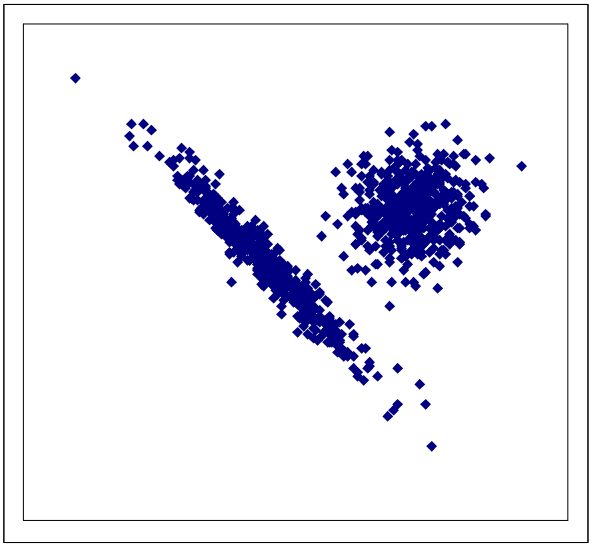


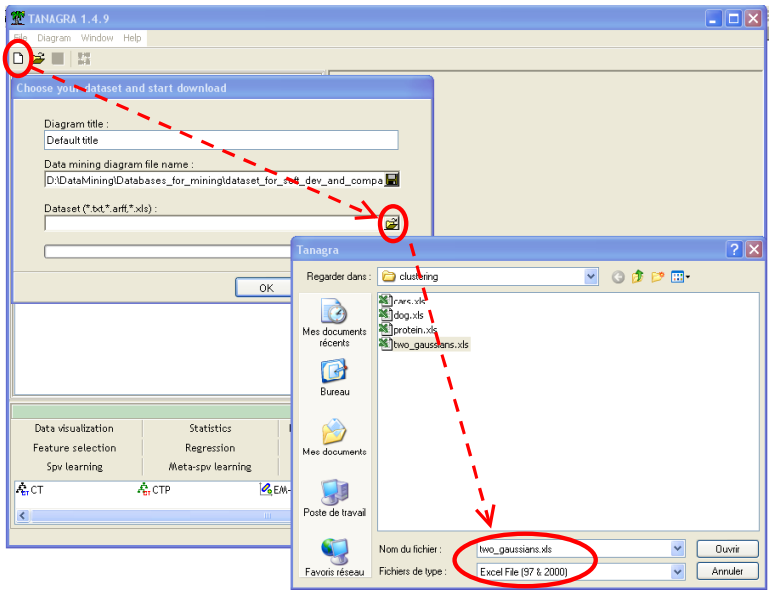
Figure 1: Two Gaussian with different parameters (means and shapes – covariance matrices)

[1] This dataset comes from the free distribution of « FAST EM Clustering » (AUTONLAB -- http://www.autonlab.org/autonweb/10466.html).

Clustering with the EM Algorithm

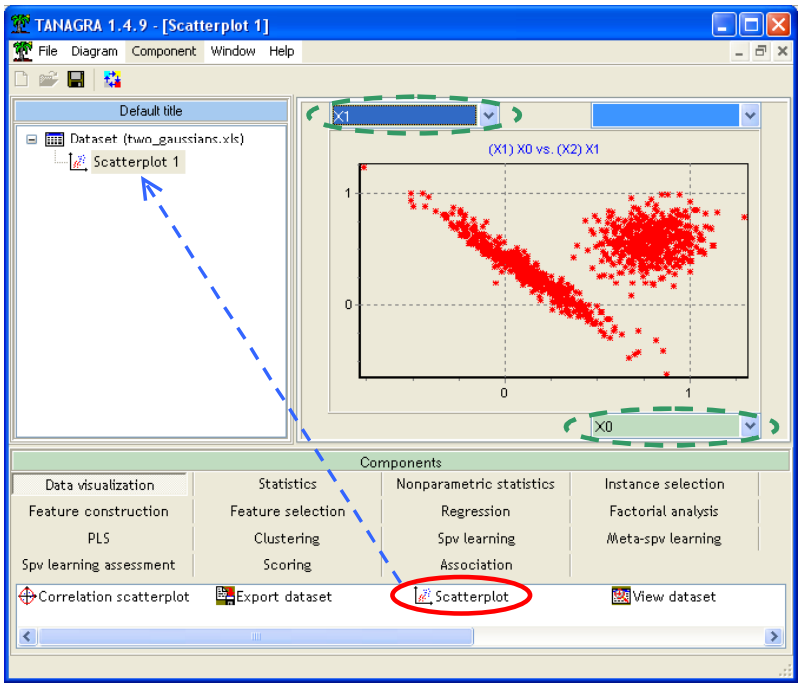
Create a diagram

We create a new diagram (FILE/NEW) and import the TWO_GAUSSIANS.XLS dataset.



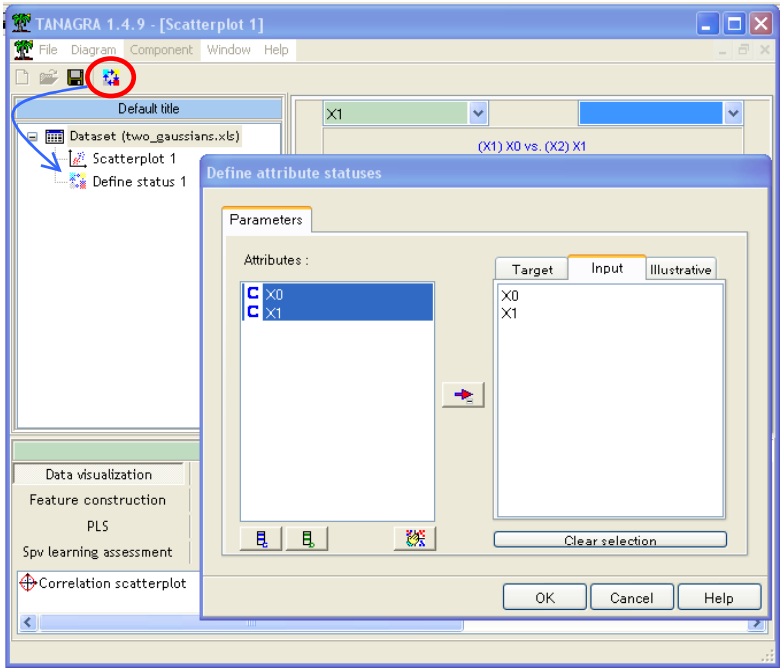
Plotting the examples

We use the SCATTERPLOT component in order to plotting the examples in a scatter plot. We distinguish well the two clusters.



INPUT attributes

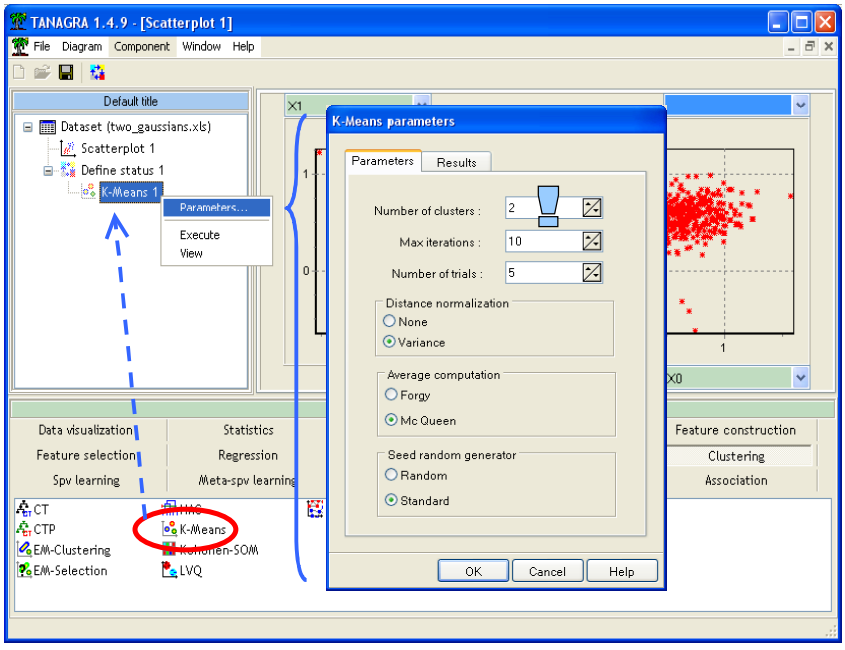
We use the DEFINE STATUS component in order to select the input attributes (X0 and X1).



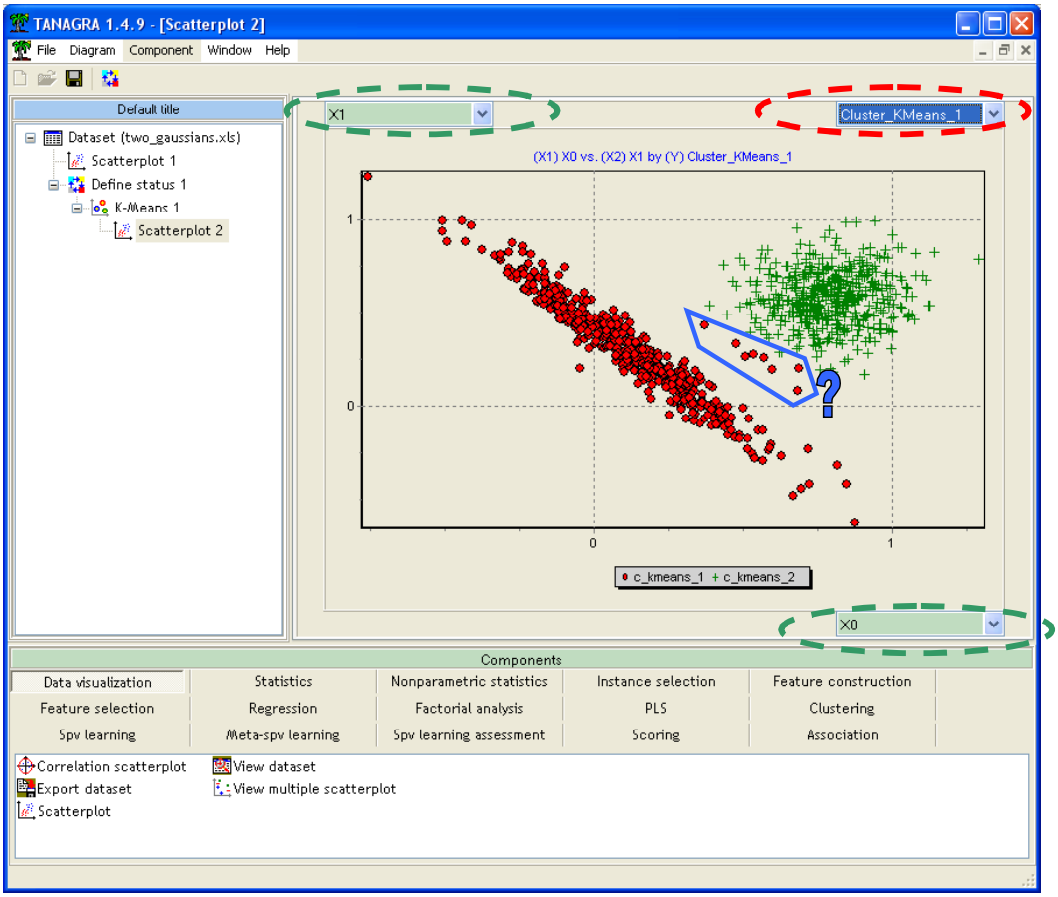
K-MEANS clustering algorithm

In the first step, we use the K-MEANS algorithm in order to create clusters. The aim is to have a reference, which will enable us to compare the later results.

We insert the K-MEANS (CLUSTERING tab) into the diagram. We modify the number of clusters parameter.



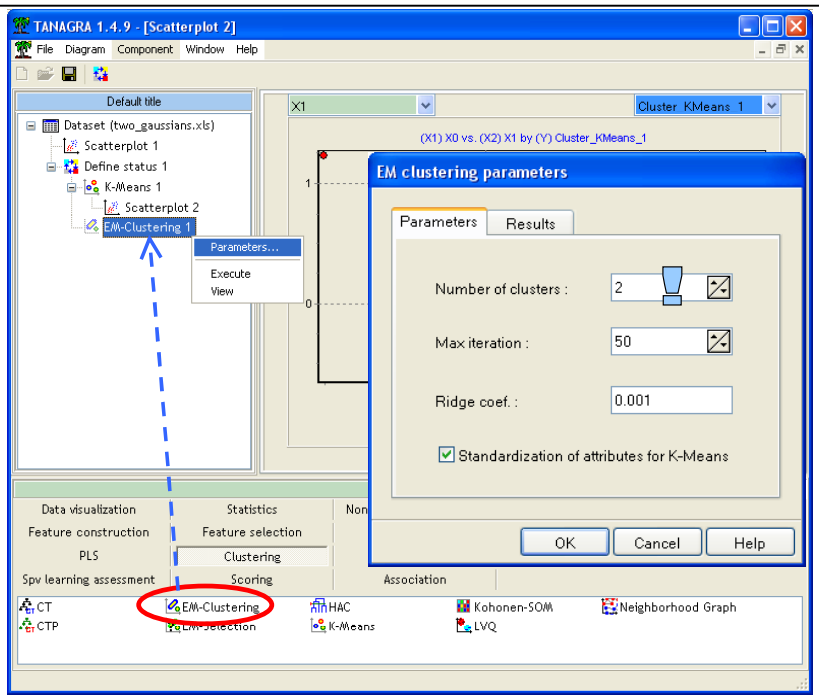
In order to visualize the clusters, we insert again a SCATTERPLOT component. We color the points with their cluster membership.



The K-MEANS method roughly found the two clusters. But we see that some examples seem visually misclassified. It is not surprising. The K-MEANS estimates only the barycentre of the clusters. The underlying hypothesis is that the shape of the clusters (covariance matrix) is spherical. This hypothesis is obviously wrong for our dataset.

Gaussian mixture based clustering

We insert the EM-CLUSTERING component in our diagram. The number of desired clusters is 2.



The component internally initializes the clusters with a K-MEANS approach, and then it optimizes the likelihood with the EM algorithm. The computation is stopped when the likelihood is not increasing anymore or when we have reached the maximum number of iterations[2]. TANAGRA shows the size of each cluster, their means and the quality of the clustering.

EM-Clustering 1

Parameters

EM parameters	
Clusters	2
Max Iteration	50
Ridge	0.001000
Seed random generator	Standard

Results

Clustering results

Clusters	2	
Cluster	**Description**	**Size**
cluster n°1	c_em_1	491
cluster n°2	c_em_2	509

Clustering quality criterion

Criterion	Value
Log-likelihood	510.4171
AIC	-998.8341
BIC	-944.8488

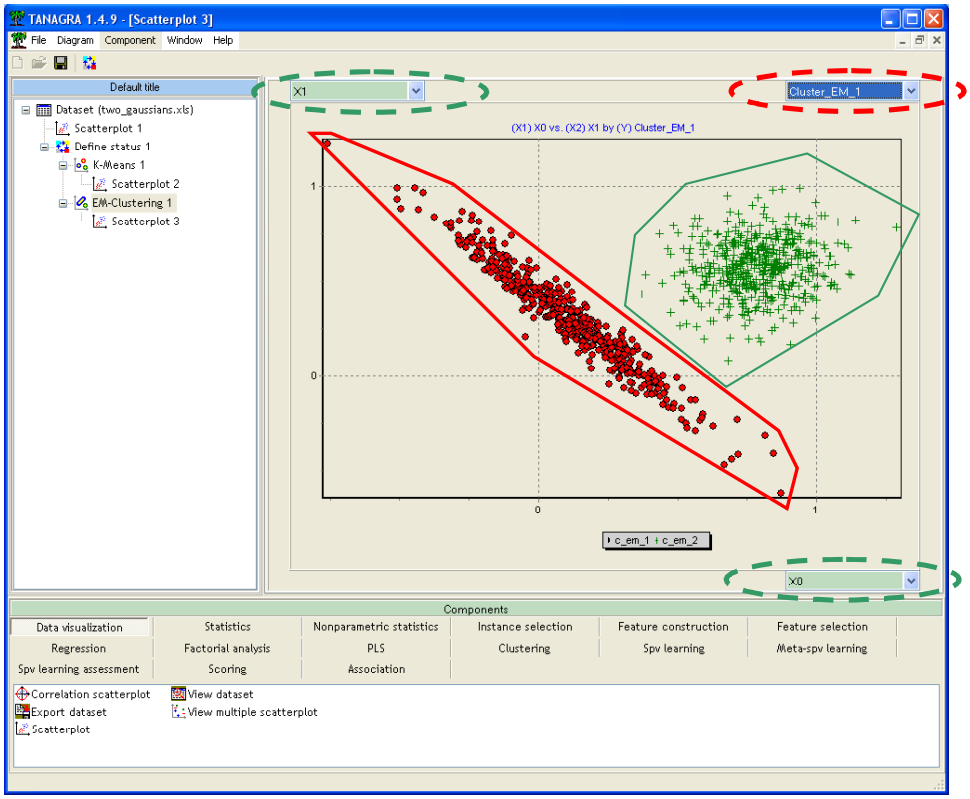
Mean of clusters

Attribute	Cluster_1	Cluster_2
X0	0.1121	0.7862
X1	0.3013	0.5745

Computation time : 16 ms.

[2] For more details on the approach, see http://fr.wikipedia.org/wiki/Algorithme_esp%C3%A9rance-maximisation and http://en.wikipedia.org/wiki/Expectation-maximization_algorithm

We add again a SCATTERPLOT component in the diagram. We see that the results of the clustering are more in adequation with our visual insight.



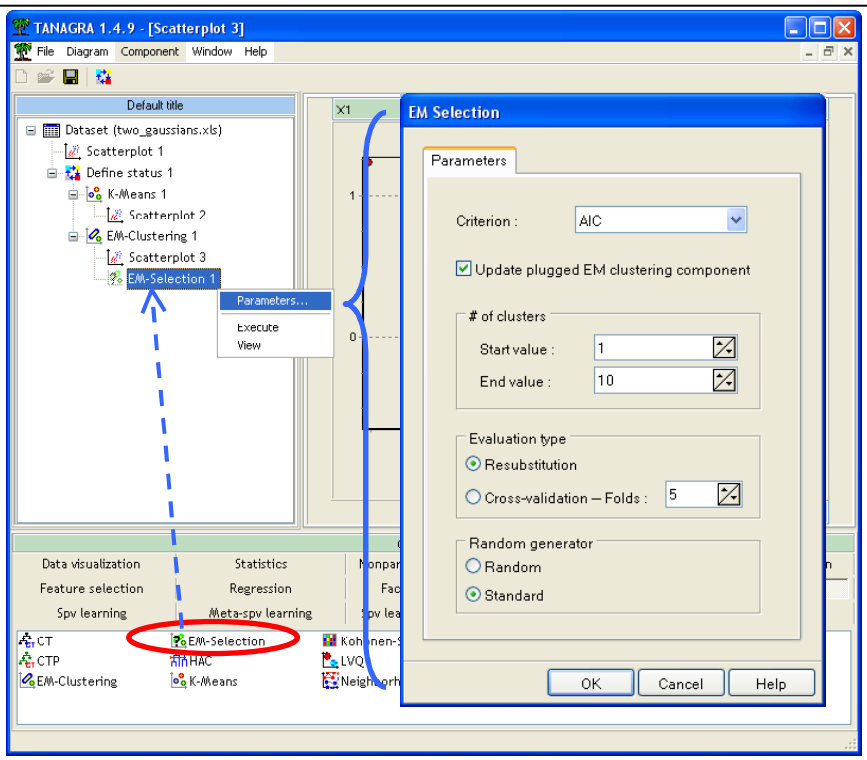
Gaussian mixture model is at least as well as K-MEANS for clustering. But it is more complex, we have to estimate more parameters, and the risk of overfitting is higher, especially when we are in a high dimensional space and fewer examples.

Detecting automatically the number of clusters

Choosing the right number of clusters is a crucial question in a clustering problem. With the Gaussian mixture model, we have a criterion that we must optimize: the likelihood. So, we can evaluate various values of the number of clusters and select the best one.

In order to obtain an unbiased result, we must insert two enhancements in this basic outline: we use a resampling method (cross-validation) for obtaining an honest likelihood estimate; the likelihood must be counterbalanced with the model complexity (number of clusters) for avoiding the solutions with a high number of clusters (i.e. AIC Akaike or BIC Schwartz criteria).

The EM-SELECTION component must be inserted under the EM-CLUSTERING component. It tests various numbers of clusters and select the best one according to the chosen criterion.



In our dataset, we use a resubstitution AIC estimate. We evaluate a number of clusters from 1 to 10. The associated EM-CLUSTERING component is automatically updated.

We obtain the following results.

EM-Selection 1

Parameters

Parameter	Value
Criterion	AIC
# clusters -- Start value	1
# clusters -- End value	10
Evaluation type	Resubstitution
Folds for CV	5
Random generator	Standard

Results

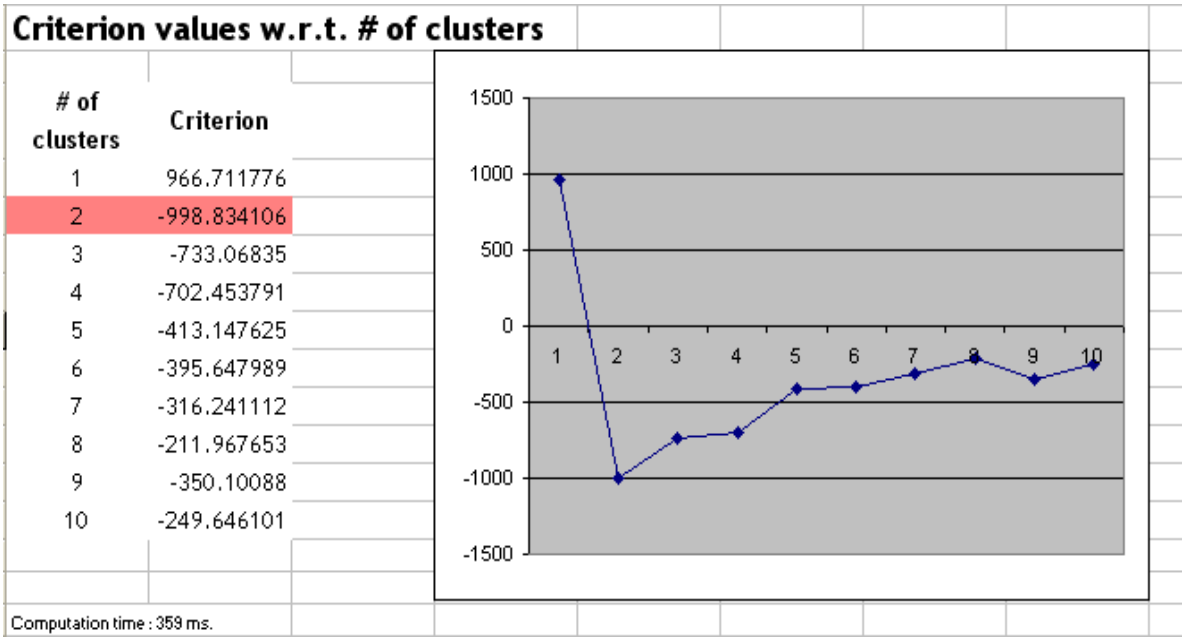
Criterion values w.r.t. # of clusters

# of clusters	Criterion
1	966.711776
2	-998.834106
3	-733.068350
4	-702.453791
5	-413.147625
6	-395.647989
7	-316.241112
8	-211.967653
9	-350.100880
10	-249.646101

Computation time : 359 ms.
Created at 28/08/2006 10:03:10

Selecting two clusters seems also numerically the best solution.

We can copying the results in a spreadsheet and plotting the criterion according to the number of clusters. It allows you to compare the various solutions.



Conclusion

Gaussian mixture model with EM algorithm is a powerful approach for clustering. It is surprisingly rarely implemented in the data mining and statistical software.