Clustering with EM: complex models vs. robust estimation

Christophe Saint-Jean, Carl Frélicot, Bertrand Vachon

Clustering with EM: complex models vs. robust estimation

C. Saint-Jean, C. Frélicot, B. Vachon

L3I - UPRES EA 1216
Avenue de Marillac,17042 La Rochelle Cedex 1
{csaintje,cfrelico,bvachon}@univ-lr.fr

Abstract. Clustering multivariate data that are contaminated by noise is a complex issue, particularly in the framework of mixture model estimation because noisy data can significantly affect the parameters estimates. This paper addresses this problem with respect to likelihood maximization using the Expectation-Maximization algorithm. Two different approaches are compared. The first one consists in defining mixture models that take into account noise. The second one is based of robust estimation of the model parameters in the maximization step of EM. Both have been tested separately, then jointly. Finally, a hybrid model is proposed. Results on artificial data are given and discussed.

Keywords: Clustering, Expectation-Maximization, Robustness, M-estimation

1 Introduction

Clustering techniques are successfully applied in many areas and some software can be now downloaded from the Internet, e.g. EMMIX by G. McLachlan and al. [4], MCLUST by C. Fraley and A. Raftery [5]. It aims at describing relationships between objects in order to group them in homogeneous clusters. Let $\chi = x_1, x_2, \ldots, x_N$ be an observed p-dimensional random sample of size N. In mixture model theory, each x_k $(k = 1, N)$ is assumed to be a realization of a p-dimensional random vector X with the C-components mixture probability density function (*pdf*):

$$f(x;\Theta) = \sum_{i=1}^{C} \pi_i f(x;\Theta_i) \tag{1}$$

where $f(x;\Theta_i)$ denotes the p-dimensional pdf of the i^{th} component and pairs (π_i, Θ_i) $(i = 1, C)$ are the model parameters. A priori probabilities π_i sum up to one. If a normal mixture model is assumed, $\Theta_i = (\mu_i, \Sigma_i)^T$ with mean μ_i and covariance matrix Σ_i. Assuming independent features of X, the model parameters $\Theta = (\pi_1, \ldots, \pi_C, \Theta_1^T, \ldots, \Theta_C^T)^T$ can be estimated by maximizing the

likelihood $\mathcal{L}(\Theta)$ using the Expectation-Maximization (EM) iterative algorithm due to Dempster, Laird and Rubin [3]:

$$\mathcal{L}(\Theta) = P(\chi|\Theta) = \prod_{k=1}^{N} \sum_{i=1}^{C} \pi_i f(x_k; \Theta_i) \tag{2}$$

Resulting clustering is not robust, i.e. it is too much sensitive to outliers or noise. In this paper, we address the problem of clustering noisy data. To face such a problem, one can choose either to perform robust estimates or to use more complex mixture models. In section 2, we briefly recall the EM algorithm. Robust M-estimates that can be used are presented in section 3; and we focus on the normal case. Next, we present a comparative study of both strategies on artificial two-dimensional data.

2 EM algorithm

In clustering problems, observed data can be regarded as being incomplete data because the labelling is unknown. Complete data $y_k = (x_k, z_k)$ can be defined by introducing for all observation x_k the realization $z_k = (z_{k1}, \ldots, z_{kC})$ of a C-dimensional random variable Z representing the labels of x_k, i.e. z_{ki} is equal to 1 when x_k arises from the i^{th} component and 0 otherwise. Then, the maximization of the likelihood (2) can be replaced by an easier one, namely the complete likelihood $\mathcal{L}_c(\Theta) = P(X, Z|\Theta)$ maximization. This is achieved by the EM algorithm that performs iterative maximization of the complete-data likelihood expectation:

$$Q(\Theta; \Theta^{(t)}) = E[\mathcal{L}_c(\Theta)|\chi, \Theta^{(t)}] \tag{3}$$

where (t) is an iteration index. Each EM iteration consists of two steps. Computation of $Q(\Theta; \Theta^{(t)})$ corresponds to the so-called **E-Step** (Expectation Step). Assuming independent z_k, the complete-data log-likelihood is:

$$\log(\mathcal{L}_c(\Theta)) = \sum_{k=1}^{N} \sum_{i=1}^{C} z_{ki} \log(\pi_i f(x_k; \Theta_i)) \tag{4}$$

Therefore:

$$Q(\Theta; \Theta^{(t)}) = \sum_{k=1}^{N} \sum_{i=1}^{C} E[z_{ki}|\chi, \Theta^{(t)}] \log(\pi_i f(x_k; \Theta_i)) \tag{5}$$

and the E-Step reduces to estimating $E[z_{ki}|\chi, \Theta^{(t)}]$. Let $\hat{z}_{ki}$ be this estimate.
The second step **M-Step** (Maximization Step) of EM consists in finding the value of Θ that maximizes $Q(\Theta; \Theta^{(t)})$:

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta; \Theta^{(t)}) \tag{6}$$

The EM algorithm increases monotonically the likelihood (see [9] by C.F.J. Wu for details).

When a p-dimensional normal mixture model is assumed, the parameters of each component (μ_i, Σ_i) as well as the prior probability π_i are iteratively estimated using $\hat{z}_{ki} = \frac{\pi_i f(x_k;\Theta_i)}{\sum_{j=1}^{C} \pi_j f(x_k;\Theta_j)}$:

$$\hat{\pi}_i^{(t+1)} = \frac{\sum_{k=1}^{N} \hat{z}_{ki}}{N} \tag{7}$$

$$\hat{\mu}_i^{(t+1)} = \frac{\sum_{k=1}^{N} \hat{z}_{ki} x_i}{\sum_{k=1}^{N} \hat{z}_{ki}} \tag{8}$$

$$\hat{\Sigma}_i^{(t+1)} = \frac{\sum_{k=1}^{N} \hat{z}_{ki} (x_k - \hat{\mu}_i^{(t+1)})(x_k - \hat{\mu}_i^{(t+1)})^T}{\sum_{k=1}^{N} \hat{z}_{ki}} \tag{9}$$

3 Robust estimation

Let $a = (a_1, \ldots, a_m)$ be a parameter to be estimated within a sample. Let e_k be the difference between an observed x_k and its predicted value $\hat{x}_k$, namely an error. e_k is a realization on a random variable e whose probability distribution is J. Assuming independent samples, the likelihood to be maximized is a product. Optimal a can be obtained by minimizing the following cost function:

$$C(a) = \sum_{k=1}^{N} \rho(\frac{x_k - \hat{x}_k(a)}{\sigma_k}) \tag{10}$$

where $\rho = \log(J^{-1})$ and σ_k is a weighting factor. This is achieved by solving the differential equations:

$$\frac{\partial C(a)}{\partial a_j} = \sum_{k=1}^{N} \frac{1}{\sigma_k} \psi(\frac{x_k - \hat{x}_k(a)}{\sigma_k}) \frac{\partial \hat{x}_k(a)}{\partial a_j} = 0 \tag{11}$$

where $\psi(x) = \frac{d\rho}{dx}(x)$. Different M-estimate models are shown in Table 1 where $w(x) = \frac{\psi(x)}{x}$ is a weight function increasing as the error decreases.

Table 1. Different M-estimates

Model	$\rho(x)$	$\psi(x)$	$w(x)$
Legendre	x^2	$2x$	2
Median	$\lvert x \rvert$	$sgn(x)$	$\frac{1}{\lvert x \rvert}$
Cauchy	$\frac{c^2}{2} \log(1 + (\frac{x}{c})^2)$	$\frac{x}{1+(\frac{x}{c})^2}$	$\frac{1}{1+(\frac{x}{c})^2}$
Huber [6]	$\begin{cases} \frac{x^2}{2} \text{ if } \lvert x \rvert \le c \\ c\lvert x \rvert - \frac{c^2}{2} \text{ else} \end{cases}$	$\begin{cases} x \text{ if } \lvert x \rvert \le c \\ c\ sgn(x) \text{ else} \end{cases}$	$\begin{cases} 1 \text{ if } \lvert x \rvert \le c \\ \frac{c}{\lvert x \rvert} \text{ else} \end{cases}$

Such models can be used when estimating the parameters (μ_i, Σ_i) of the components of a normal mixture by EM [1]. Equation (8) is simply replaced by the following procedure:

- $\tau = 0$
  $\tilde{\mu}_i^{(t+1,\tau)} = \hat{\mu}_i^{(t+1)}$
  $\tilde{\Sigma}_i^{(t+1,\tau)} = \hat{\Sigma}_i^{(t+1)}$
- *repeat* $\forall k = 1, N : e_k = (x_k - \tilde{\mu}_i^{(t+1,\tau)})^T \tilde{\Sigma}_i^{-1(t+1,\tau)} (x_k - \tilde{\mu}_i^{(t+1,\tau)})$
  $w(e_k) = \frac{\psi(e_k)}{e_k}$

$$\tilde{\mu}_i^{(t+1,\tau)} = \frac{\sum_{k=1}^{N} w(e_k)\ \hat{z}_{ki}\ x_k}{\sum_{k=1}^{N} w(e_k)\ \hat{z}_{ki}} \tag{12}$$

$$\tilde{\Sigma}_i^{(t+1,\tau)} = \frac{\sum_{k=1}^{N} w(e_k)\ \hat{z}_{ki}\ (x_k - \tilde{\mu}_i^{(t+1,\tau)})(x_k - \tilde{\mu}_i^{(t+1,\tau)})^T}{\sum_{k=1}^{N} w(e_k)\ \hat{z}_{ki}} \tag{13}$$

  $\tau \leftarrow \tau + 1$
  until reached bound

It is worthy of note that the property of monotonous increase of the likelihood is lost by EM in case of robust estimation. The model parameters are no more updated by solving $\frac{dQ(\Theta;\Theta^{(t)})}{d\Theta} = 0$. However, these estimates (7) – (9) are good initial values for the robust ones (12) – (13).
Since the weight functions w are monotonous decreasing functions of the error e_k, the more τ, the more robust but the less precise estimates are provided. Therefore the number of iterations is bounded, e.g.:

- $\tau < \tau_{max}$
- $\frac{|\tilde{\mu}_i^{(t+1,\tau+1)} - \tilde{\mu}_i^{(t+1,\tau)}|}{\tilde{\mu}_i^{(t+1,\tau)}} < \epsilon$
- rate of samples having a quite zero weight $\alpha < \alpha_{max}$

We have combined the first and the third conditions in our experiments.

4 Mixture models

In this study, we have used normal and uniform components for computation convenience. Let $\mathcal{N}$ denotes the p-dimensional gaussian pdf and $\mathcal{U}$ the uniform one defined on a given hypercube H. Here are the more or less complex mixture models that we have tested:

1. C normal components where the parameters are estimated via (7) – (9):

$$f(x;\Theta) = \sum_{i=1}^{C} \pi_i \, \mathcal{N}(x;\mu_i,\Sigma_i) \tag{14}$$

2. C normal and one uniform components, γ being user defined:

$$f(x;\Theta) = \gamma \, \mathcal{U}(x;H) + \sum_{i=1}^{C} \pi_i \, \mathcal{N}(x;\mu_i,\Sigma_i) \tag{15}$$

3. C normal components where robust estimates (12) – (13) are used:

$$f(x;\Theta) = \sum_{i=1}^{C} \pi_i \, \mathcal{N}(x;\mu_i,\Sigma_i) \tag{16}$$

4. C normal and one uniform components with robust estimates:

$$f(x;\Theta) = \gamma \, \mathcal{U}(x;H) + \sum_{i=1}^{C} \pi_i \, \mathcal{N}(x;\mu_i,\Sigma_i) \tag{17}$$

5. C normal components with robust estimates and one additional mixture:

$$f(x;\Theta) = \sum_{i=1}^{C+1} \pi_i f(x;\Theta_i) \tag{18}$$

$$f(x;\Theta_i) = (1-\gamma_i) \, \mathcal{N}(x;\mu_i,\Sigma_i) \qquad \forall i = 1, C \tag{19}$$

$$f(x;\Theta_{C+1}) = \sum_{i=1}^{C} \gamma_i \, \mathcal{N}(x;\mu_i,\alpha_i\Sigma_i) \tag{20}$$

In this latter model we propose, each of the C components is a linear combination of two normal pdf. The first one intends to track cluster kernel points while the second one is supposed to deal with surrounding outliers via multiplicative coefficients α_i. All these second modes are summed up to compose a $(C+1)^{th}$ component. The combination coefficients γ_i are user-defined as well as α_i.
Our model differs from previous work, e.g. G. Mac Lachlan and D. Peel [8] or Y. Kharin [7], in mixing robust estimation of the parameters in (19) and classical estimation of ones in (20).

5 Experiments

We have generated two data sets in order to test the robustness of the presented mixture models. Both consist of three 2-dimensional gaussian classes and uniformly distributed samples supposed to be noisy points, as shown on figure 1. In the first set the classes are well-separated while they strongly overlap in the second one. Noisy patterns represent respectively 30% and 16.7% of each data set.

Table 2. Data sets parameters

	μ_1	Σ_1	r_1	μ_2	Σ_2	r_2	μ_3	Σ_3	r_3
Data set #1	$\begin{pmatrix}7\\9\end{pmatrix}$	$\begin{pmatrix}3 & -2\\-2 & 1.5\end{pmatrix}$	$\frac{-2\sqrt{2}}{3}$	$\begin{pmatrix}6\\3\end{pmatrix}$	$\begin{pmatrix}3 & 0\\0 & 1\end{pmatrix}$	0	$\begin{pmatrix}3\\7\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	0
Data set #2	$\begin{pmatrix}3\\5\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 3\end{pmatrix}$	0	$\begin{pmatrix}6\\9\end{pmatrix}$	$\begin{pmatrix}2 & 1.5\\1.5 & 2\end{pmatrix}$	$\frac{3}{4}$	$\begin{pmatrix}6\\5\end{pmatrix}$	$\begin{pmatrix}3.25 & -3\\-3 & 3.25\end{pmatrix}$	$\frac{-12}{13}$

Table 2 summarizes the theoretical parameters of the classes for both generated datasets. The correlation coefficients r_i that describe the clusters orientations are given. As most of partitioning methods for clustering data, the number of mixture components has to be set. One can assess this value [2] but we preferred to choose it manually ($C = 3$) in order to concentrate on robust estimation effect. The fitted parameters provided by EM significantly depend on initial values. For each model and each data set, we have run EM 50 times with same random initial values $\hat{\mu}_i$ and identity matrices for covariance matrices. We also have tried different values of the coefficients involved in the different models and kept the best results according to the maximum posterior probability criterion for the only gaussian points. Only the Cauchy M-estimate has been tested. In addition, the resulting cluster correlation coefficients r_i are estimated.

Table 3. Data set #1 – Results (final estimates values and errors)

	Model #1	Model #2	Model #3	Model #4	Model #5
μ_1	$\begin{pmatrix}7.07\\8.91\end{pmatrix}$	$\begin{pmatrix}7.18\\8.87\end{pmatrix}$	$\begin{pmatrix}7.17\\8.86\end{pmatrix}$	$\begin{pmatrix}7.14\\8.84\end{pmatrix}$	$\begin{pmatrix}7.16\\8.86\end{pmatrix}$
Σ_1	$\begin{pmatrix}5.32 & -1.69\\-1.69 & 2.28\end{pmatrix}$	$\begin{pmatrix}2.49 & -1.59\\-1.59 & 1.56\end{pmatrix}$	$\begin{pmatrix}0.89 & -0.63\\-0.63 & 0.59\end{pmatrix}$	$\begin{pmatrix}1.72 & -1.2\\-1.2 & 1.17\end{pmatrix}$	$\begin{pmatrix}2.77 & -1.36\\-1.36 & 1.47\end{pmatrix}$
r_1	-0.48	-0.81	-0.87	-0.87	-0.67
μ_2	$\begin{pmatrix}6.07\\2.74\end{pmatrix}$	$\begin{pmatrix}6.12\\2.73\end{pmatrix}$	$\begin{pmatrix}6.21\\2.73\end{pmatrix}$	$\begin{pmatrix}6.09\\2.73\end{pmatrix}$	$\begin{pmatrix}6.09\\2.73\end{pmatrix}$
Σ_2	$\begin{pmatrix}4.46 & -0.01\\-0.01 & 1.13\end{pmatrix}$	$\begin{pmatrix}4.06 & 0.08\\0.08 & 1.04\end{pmatrix}$	$\begin{pmatrix}1.77 & -0.01\\-0.01 & 0.41\end{pmatrix}$	$\begin{pmatrix}3.31 & 0.00\\0.00 & 0.89\end{pmatrix}$	$\begin{pmatrix}3.45 & -0.00\\-0.00 & 0.90\end{pmatrix}$
r_2	-0.00	0.04	-0.01	-0.00	-0.00
μ_3	$\begin{pmatrix}2.77\\6.90\end{pmatrix}$	$\begin{pmatrix}2.98\\7.00\end{pmatrix}$	$\begin{pmatrix}2.99\\7.06\end{pmatrix}$	$\begin{pmatrix}2.97\\7.03\end{pmatrix}$	$\begin{pmatrix}2.86\\6.99\end{pmatrix}$
Σ_3	$\begin{pmatrix}1.72 & -0.38\\-0.38 & 1.26\end{pmatrix}$	$\begin{pmatrix}2.05 & -0.37\\-0.37 & 1.66\end{pmatrix}$	$\begin{pmatrix}1.03 & -0.08\\-0.08 & 0.76\end{pmatrix}$	$\begin{pmatrix}1.78 & -0.2\\-0.2 & 1.22\end{pmatrix}$	$\begin{pmatrix}1.56 & -0.31\\-0.31 & 1.12\end{pmatrix}$
r_3	-0.26	-0.20	-0.09	-0.13	-0.23
E	3.43%	3.43%	2.%	3.43%	2.86%

Not surprisingly, all the models give a similar low error rate value on data set #1 (see Table 3). As the gaussian classes are well separated and as the noise rate is

low enough (30%) the optimal error rate has been obtained for a low value of the mixing coefficients γ or γ_i (models #2,#4 or #5). Therefore, robust estimation has a not a significant action. The Cauchy parameter c whose value describes the number of observed points contributing to the robust estimates compensates a low value of the mixing coefficients (if used). The smaller γ or γ_i are, the less noisy points are modelled. So the smaller Cauchy's parameter c is in order to filter enough. The means are very close to the theoretical ones whatever the model is. On the other hand, taking the noise into account (models #2,#3,#4 and #5) clearly improve the clusters shapes and orientations as reflected by the obtained covariance matrices and correlation coefficients. The optimal partition we obtained with model #3 is shown on Figure 2 (left hand-side). Obviously, all the outliers have been incorrectly clustered, the few ones in the upper right corner in particular. This can be explained by the fact that original noisy points do not enter into the error rate computation. Clusters resulting from the model we propose (#5) are shown on Figure 2 (right hand-side).

Table 4. Data set #2 – Results (final estimates values and errors)

	Model #1	Model #2	Model #3	Model #4	Model #5
μ_1	$\begin{pmatrix} 3.97 \\ 5.06 \end{pmatrix}$	$\begin{pmatrix} 2.84 \\ 4.65 \end{pmatrix}$	$\begin{pmatrix} 2.74 \\ 4.57 \end{pmatrix}$	$\begin{pmatrix} 3.02 \\ 5.20 \end{pmatrix}$	$\begin{pmatrix} 2.95 \\ 5.11 \end{pmatrix}$
Σ_1	$\begin{pmatrix} 5.60 & 0.03 \\ 0.03 & 4.54 \end{pmatrix}$	$\begin{pmatrix} 1.45 & -0.38 \\ -0.38 & 3.14 \end{pmatrix}$	$\begin{pmatrix} 0.96 & -0.37 \\ -0.37 & 1.85 \end{pmatrix}$	$\begin{pmatrix} 1.26 & -0.10 \\ -0.10 & 3.38 \end{pmatrix}$	$\begin{pmatrix} 1.21 & -0.17 \\ -0.17 & 3.06 \end{pmatrix}$
r_1	0.01	-0.18	-0.28	-0.05	-0.09
μ_2	$\begin{pmatrix} 6.06 \\ 9.13 \end{pmatrix}$	$\begin{pmatrix} 5.99 \\ 8.94 \end{pmatrix}$	$\begin{pmatrix} 6.03 \\ 9.04 \end{pmatrix}$	$\begin{pmatrix} 6.09 \\ 9.05 \end{pmatrix}$	$\begin{pmatrix} 6.10 \\ 9.06 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 2.02 & 1.19 \\ 1.19 & 1.41 \end{pmatrix}$	$\begin{pmatrix} 1.70 & 0.95 \\ 0.95 & 1.40 \end{pmatrix}$	$\begin{pmatrix} 1.07 & 0.50 \\ 0.50 & 0.81 \end{pmatrix}$	$\begin{pmatrix} 1.29 & 0.69 \\ 0.69 & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1.24 & 0.56 \\ 0.56 & 0.95 \end{pmatrix}$
r_2	0.70	0.61	0.53	0.59	0.52
μ_3	$\begin{pmatrix} 5.68 \\ 5.17 \end{pmatrix}$	$\begin{pmatrix} 5.61 \\ 5.30 \end{pmatrix}$	$\begin{pmatrix} 5.45 \\ 5.45 \end{pmatrix}$	$\begin{pmatrix} 6.40 \\ 4.65 \end{pmatrix}$	$\begin{pmatrix} 6.30 \\ 4.74 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 5.58 & -5.4 \\ -5.4 & 5.53 \end{pmatrix}$	$\begin{pmatrix} 3.80 & -3.37 \\ -3.37 & 3.45 \end{pmatrix}$	$\begin{pmatrix} 3.00 & -2.65 \\ -2.65 & 2.70 \end{pmatrix}$	$\begin{pmatrix} 1.90 & -1.82 \\ -1.82 & 2.06 \end{pmatrix}$	$\begin{pmatrix} 2.13 & -2.01 \\ -2.01 & 2.27 \end{pmatrix}$
r_3	-0.96	-0.93	-0.93	-0.92	-0.91
E	15.33%	12.67%	14.22%	10.67%	11.33%

When faced to much more overlapping clusters (data set #2), the parameter estimation process is strongly disturbed, the means being attracted by the dense areas, as shown in Table 4. As expected, robust estimates tends to correct this trend, i.e. the obtained values are closer to the theoretical ones (models #3, #4 and #5). Consequently, the optimal Cauchy parameter values c are smaller than those selected on data set #1. Figure 3 shows the partitions that we have obtained with the best models involving robust estimates according to the parameter fitting and error as well (model #4 on the left hand-side, model #5 on

the right hand-side).

Table 5. Results over the 50 runs

	Error	Model #1	Model #2	Model #3	Model #4	Model #5
Data set#1	Mean	8.31%	9.87%	9.86%	11.87%	2.96%
	StDev	10.87	12.57	15.18	15.15	0.36
Data set#2	Mean	18.82%	13.27%	16.76%	13.95%	11.93%
	StDev	5.8	4.27	5.	6.97	0.2

In order to test the robustness to initial center locations, we have chosen the optimal parameter setting with respect to the maximum posterior probability criterion. Table 5 summarizes the means and standard deviations of the error we have obtained over the 50 different runs. According to a mean value close to the minimum one and a low standard deviation, our model (#5) outperforms all the others on both datasets. We think that the lower sensitivity of this model to initialization can be explained by the introduction of normal subcomponents that softens the tails of the resulting component. The ability of the model 5 to perform good clustering in spite of a bad initialization suggest us that it would be useful in many real situations.

6 Conclusion

In this paper, we have compared two different approaches to clustering multivariate data in the context of mixture of components likelihood maximization with the EM algorithm. Indeed, such algorithm often fails in finding accurate parameters when the data are mixed with noisy data. So, one can either take noisy data into account when defining the mixture model or use robust estimation techniques. We have noticed that both approaches can improve the results whatever the separability of the clusters is. Furthermore, in case of strong overlap, their joint use give better results. We have proposed such a model whose performances in terms of misclassification as well as accuracy of the parameters estimates are satisfactory. Moreover, we notice that this model is very robust to different initializations. Further investigation will concern the automatic selection of some coefficients involved in this model.

Acknowledgements
This work has been partially supported by the *Conseil Général de Charente-Maritime.*

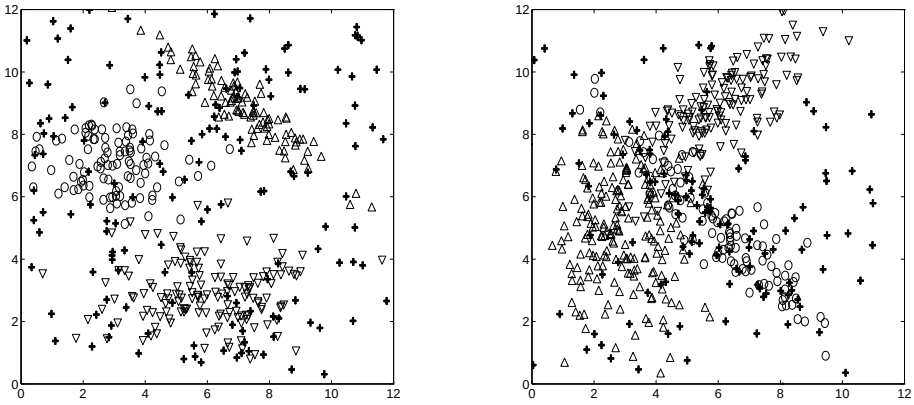

Fig. 1. Data sets #1 (left) and #2 (right)

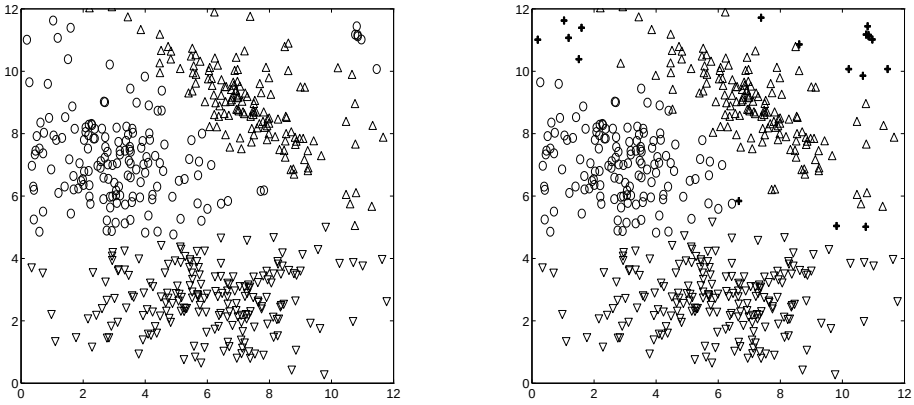

Fig. 2. Data set #1 clustering with models #3 (left) and #5 (right)

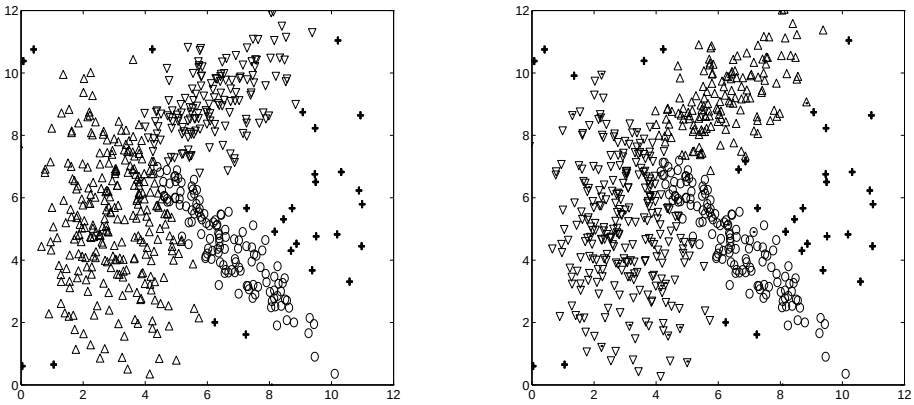

Fig. 3. Data set #2 clustering with models #4 (left) and #5 (right)

References

1. Campbell, N.A., Lopuhad, H.P., Rousseeuw, P.J.: On the calculation of a robust S-estimator of a covariance matrix. Delft University of Technology. Tech. Report **DUT-TWI-95-117** (1995)
2. Celeux, G., Soromenho, G.: An entropy criterion for assessing the number of clusters in a mixture model. J. of Classification **13** (1996) 195–212
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm with discussion. J. of the Royal Stat. Soc. **39** (1977) 1–38
4. McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P.: The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software **4**, No. **2**. (1999).
5. Fraley, C., Raftery, A.E.: MCLUST: Software for model-based clustering and discriminant analysis. Univ. of Wash. Tech. Report **TR-342** (1998)
6. Huber, P.J.: Robust statistics. John Wiley. New-York (1981)
7. Kharin, Y.: Robustness of clustering under outliers. LNCS **1280** (1997)
8. McLachlan, G.J., Peel, D.: Robust cluster analysis via mixtures of multivariate t-distributions. LNCS **1451** (1999) 658–667
9. Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. of Stat. **11** (1983) 95–103