# Comparing Several Methods to Fit Finite Mixture Models to Grouped Data by the EM Algorithm

Joanna Wengrzik and Jürgen Timm

*Abstract*—**The EM algorithm is a standard tool for maximum likelihood estimation in finite mixture models. Common approaches assume continuous data for its application. Frequently in practice, data is only available in grouped form, i.e. the frequencies of observations in fixed intervals are reported. The fitting of two-component Gaussian mixture to such data is considered in this paper. The aim is to compare several methods for fitting mixtures to grouped data via the EM algorithm, as well as to propose some new methods based on modifications of existing ones. Furthermore, the influence of different widths of intervals on the estimation is investigated. Finally, an example is presented.**

*Index Terms*—**EM Algorithm, Finite Mixture, Grouped Data**

## I. Introduction

FINITE mixture models are being increasingly used to model the distribution of heterogenous populations, which arises when subpopulations with different density functions occur.

A $k$-component mixture model for the density function $g(x)$ of a random variable $X$ has the form $g(x) = \sum_{j=1}^{k} \pi_j f_j(x)$, where $\pi_1, ..., \pi_k$ denote the mixing proportions which sum to one, and $f_j(x)$ denote the component density functions. Typically, the component density functions are specified up to a vector of unknown parameters, say $\theta_j$. In many applications, the component density functions $f(x, \theta_j)$ are taken to belong to the same parametric family, for example, the Gaussian. In the case of a two-component Gaussian mixture, the parameter vector is $\theta_j = (\mu_j, \sigma_j)$, where $j = 1, 2$.

Furthermore, $x_1, ..., x_n$ denote an observed sample of size $n$. A mixture model can be fitted to these data by maximum likelihood via the Expectation-Maximization algorithm of Dempster, Laird and Rubin [1]. There is a large amount of literature dealing with fitting a mixture model when the individual data points are available, see Everitt and Hand [3], Titterington, Smith and Markov [13], McLachlan and Peel [7], McLachan and Krishnan [9] and McLachlan and Basford [11].

In practice, however, individual data points are frequently not given and data is only available in grouped form, i.e. the frequencies of observations in fixed intervals are reported. Dempster, Laird, and Rubin [1] showed how the EM algorithm can be used for such data, but they did not consider mixture models in this context. Schader and Schmid [12] compared different algorithms for grouped data, but not for mixture models.

Using the EM algorithm to fit a mixture model to grouped data, as Dempster et al. suggested, the solution to the M-step of the algorithm does not exist in closed form. Several methods concerning this problem have been proposed, but no comparative study appears to have been made. This paper focuses on this issue. In particular, the first results of a simulation study are presented that compares several approaches of the EM algorithm in case of a two-component Gaussian mixture, which are reviewed in Section II. The results of the simulation study are shown in Section III. Section IV presents an example, and finally a brief summary is given in Section V.

## II. Fitting mixture to grouped data

Consider a two component Gaussian mixture

$$g(x, \Psi) = \sum_{j=1}^{2} \pi_j f(x, \theta_j),$$

where $\Psi = (\pi_1, \pi_2, \theta_1, \theta_2)$ contains the six parameters which need to be estimated: the mixing proportions $\pi_1$ and $\pi_2$, as well as the parameters $\theta_1 = (\mu_1, \sigma_1)$ and $\theta_2 = (\mu_2, \sigma_2)$ of the Gaussian distributions. Instead of the individual observations, only the number of observations $y = (n_1, ..., n_m)$ falling into the intervals $[a_0, a_1), [a_1, a_2), ..., [a_{m-1}, a_m)$ are available.

Given data grouped like a histogram, a theoretical mixed distribution $g(x, \Psi)$ can be fitted by finding parameters $\widehat{\Psi}$ that make the discrepancy between the theoretical distribution and the observed grouped data as small as possible. One possibility to solve this problem is the appliance of the EM algorithm, which requires the calculation of the conditional complete log likelihood function. In this context, the observed data $y$ is viewed as being incomplete. By introducing a missing data vector $z$, the E-step of the EM algorithm for the complete data $w = (y, z)$ is generally given by:

$$Q(\Psi, \Psi^{(t)}) = E_{\Psi^{(t)}}[\log L_c(\Psi, w)|y],$$

where $L_c$ denotes the complete log likelihood function and $t$ indicates the $t$th iteration. On the M-step the intent is to choose a value of $\Psi$ that maximizes the $Q$-function.

The following methods, based on different previously proposed approaches as well as new ones, have been compared:

*Method 1*: The approach, proposed by MacDonald and Pitcher [5] and MacDonald and Green [6], was implemented in the R package *mixdist* by Du [2]. By introducing the variable $n^*_{ij}$ as missing data, which denotes the number of observations from the $j$th group falling into the $i$th interval,

Manuscript received February 2, 2011; revised March 11, 2011. The authors gratefully acknowledge the financial support by DFG (German Research Foundation) for the subprojects B2 "Methoden zu direkten Berücksichtigung der Verteilungsfunktion von Stoffwerten bei der FEM Simulation von Mikroumformprozessen" within the Collaborative Research Center SFB 747 "Mikrokaltumformen - Prozesse, Charakterisierung, Optimierung".

J.Wengrzik and J.Timm are with the Competence Center for Clinical Trials, University for Bremen, Germany, e-mail: wengrzik@uni-bremen.de.

the log likelihood function for the complete data $w = (y, n^*)$ is given as

$$\log L_c(\Psi, w) = \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log \pi_j + \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log P_{ij}(\theta_j),$$

with $P_{ij}(\theta_j) = \int_{a_{i-1}}^{a_i} f(x, \theta_j)dx$. The E-step can be calculated in a straightforward way: The unknown $n_{ij}^*$ are replaced by their expected values given the observed data $y$, which is $E(Z_{ij}|y) = \pi_j n_i P_{ij}(\theta_j^{(t)})/P_i(\theta_j^{(t)})$, where $P_i(\theta_j^{(t)}) = \sum_{j=1}^{k} P_{ij}(\theta_j^{(t)})$. However, the maximization of the Q-function, that is required in the M-step, cannot be done analytically, so that another iterative procedure is necessary. In this case the Newton-Raphson algorithm is used.

*Method 2*: An easier approach to the previous method is to approximate the probability $P_{ij}(\theta_j)$ that an individual observation $x_i$ falls into the *j*th interval by $h \cdot f(\bar{a}_i, \theta_j)$, where $h$ is the width of the interval and $\bar{a}_i = (a_{i-1}+a_i)/2$ indicates the midpoint of the *i*th interval. For this approximation the log likelihood for the complete data $w = (y, n^*)$ is:

$$\begin{aligned}\log L_c(\Psi, w) &= \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log \pi_j + \\ &\quad \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log[h \cdot f(\bar{a}_i, \theta_j)].\end{aligned}$$

After replacing the $n_{ij}^*$ by their conditional expectation $E(Z_{ij}|y) = \pi_j n_i f(\bar{a}_i, \theta_j^{(t)})/g(\bar{a}_i, \theta_j^{(t)}) =: e_{ij}^{(t)}$, the derivation of the Q-function in the M-step can be done analytically. The iterative estimators are given by:

$$\begin{aligned}\pi_j^{(t+1)} &= \frac{1}{m}\sum_{i=1}^{m} e_{ij}^{(t)} \\ \mu_j^{(t+1)} &= \sum_{i=1}^{m} e_{ij}^{(t)}\bar{a}_i / \sum_{i=1}^{m} e_{ij}^{(t)} \\ \sigma_j^{2(t+1)} &= \sum_{i=1}^{m} e_{ij}^{(t)}(\bar{a}_i - \mu_j^{(t+1)})^2 / \sum_{i=1}^{m} e_{ij}^{(t)}\end{aligned}$$

*Method 3*: The simplest approach being examined is the transformation of grouped data into individual data. For each interval $[a_i, a_{i+1})$ the midpoint $\bar{a}_i = (a_{i-1} + a_i)/2$ is replicated $n_i$ times, where $n_i$ indicates the number of observations per interval. For this transformed data $x^* = (x_1^*, x_2^*, ..., x_n^*)$ the EM algorithm for finite mixture models has been applied, where the missing data is given by an indicator variable $z_{ij} \in \{0, 1\}$, which indicates whether an observation $x_i^*$ arose or did not arise from the *j*th component. The log likelihood for the complete data $w = (x^*, z)$ is given by:

$$\log L_c(\Psi, w) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log f(x_i^*, \theta_j).$$

Considering the conditional expectation $E(Z_{ij}|x) = \pi_j f(x_i^*, \theta_j^{(t)})/g(x_i^*, \theta_j^{(t)}) =: e_{ij}^{*(t)}$, the solutions of the maximization have closed forms, that are similar to those of method 2.

*Method 4*: The second data transformation being considered is that the grouped data is transformed into individual data $x^* = (x_1^*, x_2^*, ..., x_n^*)$, which is uniformly distributed at every interval. Similar to method 3, this transformed data is used by the EM algorithm for finite mixture models.

*Method 5*: McLachlan and Jones [11], [10] proposed another approach. By introducing a random variable $x_{ij}$ as missing data vector, which includes individual observations per interval, and extending the complete data vector to a zero-one indicator variable $z_{ijl}$, which indicates the affiliation of the *l*th observation in the *i*th interval to the *j*th component, the log likelihood for the complete data $w = (x, y, z)$ is given by

$$\log L_c(\Psi, w) = \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n_i} z_{ijl}[\log \pi_j + \log f(x_{ij}, \theta_j)].$$

For this log likelihood function, the calculation of the E-step can be done analytically as well as the maximization in the M-step.

The similarity of method 2 and 3 gives reason to investigate this two methods separately. An analytical consideration of the conditional expectation of method 3 gives:

$$\begin{aligned}\sum_{i=1}^{n} e_{ij}^{*(t)} &= \sum_{i=1}^{n} \frac{\pi_j f(x_i^*, \theta_j^{(t)})}{g(x_i^*, \theta_j^{(t)})} = \\ &\quad \underbrace{\frac{\pi_j f(x_1^*, \theta_j^{(t)})}{g(x_1^*, \theta_j^{(t)})} + ... + \frac{\pi_j f(x_1^*, \theta_j^{(t)})}{g(x_1^*, \theta_j^{(t)})}}_{n_1 times} + ... + \\ &\quad \underbrace{\frac{\pi_j f(x_n^*, \theta_j^{(t)})}{g(x_n^*, \theta_j^{(t)})} + ... + \frac{\pi_j f(x_n^*, \theta_j^{(t)})}{g(x_n^*, \theta_j^{(t)})}}_{n_m times} \\ &= \sum_{i=1}^{m} \frac{\pi_j n_i f(\bar{a}_i, \theta_j^{(t)})}{g(\bar{a}_i, \theta_j^{(t)})} = \sum_{i=1}^{m} e_{ij}^{(t)},\end{aligned}$$

which is equal to the conditional expectation of method 2. Hence, both methods provide the same iterative estimators, although they uses two different approaches. However, the runtime is not equal, which can be seen more detailed in the next section, where the comparison is presented.

## III. A SIMULATION COMPARISON

The methods described above are compared by considering several two component Gaussian mixtures. The first component is taken to be fixed with parameters $\mu_1 = 5$ and $\sigma_1 = 1$. The mixing proportions are chosen as $\pi_1 = \pi_2 = 0.5$. The standard deviation of the second component $\sigma_2$ is chosen equal to 1 while for the mean $\mu_2$ five different values are considered. Starting with a heavily overlapped distribution, where the mean $\mu_2$ is taken to be equal to 6, the second distribution is adjusted as long as there are two well-separated distributions, i.e. when $\mu_2$ is equal to 10. According to preliminary simulations, the sample size is chosen as $n = 100$. In the first investigation, the generated individual data is grouped with an interval width of $h = 1$. Iterations are stopped if the absolute difference in the log likelihood is smaller than $10^{-6}$. To avoid the convergence to local maxima, the true values are chosen as initial values.

In the comparison of the methods, the main focus is on the quality of the estimation of $\Psi$, measured by the discrepancy

TABLE I
MSEs FOR THE SIX ESTIMATED PARAMETERS FOR THE MODEL $\pi_1 = \pi_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 5$ AND $\mu_2 = 6$

| MSE | $\widehat{\pi_1}$ | $\widehat{\pi_2}$ | $\widehat{\mu_1}$ | $\widehat{\mu_2}$ | $\widehat{\sigma_1}$ | $\widehat{\sigma_2}$ | $\sum$ |
|---|---|---|---|---|---|---|---|
| Method 1 | 0.10906 | 0.10906 | 0.98357 | 0.92800 | 0.26321 | 0.24113 | 2.63403 |
| Method 2&3 | 0.00045 | 0.00045 | 0.01094 | 0.01187 | 0.00654 | 0.00622 | 0.03648 |
| Method 4 | 0.00047 | 0.00047 | 0.01378 | 0.01477 | 0.00892 | 0.00834 | 0.04676 |
| Method 5 | 0.00040 | 0.00040 | 0.00902 | 0.00965 | 0.00497 | 0.00471 | 0.02915 |

between the estimated and the true parameters. Therefore, the mean squared error (MSE) is considered, which is defined by

$$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + (E(\widehat{\theta}) - \theta)^2.$$

The simulation study provides the distribution for each parameter estimator, hence the mean and the standard deviation for the six parameters can be estimated, which is necessary for the MSE. The results for the first considered model, where $\mu_2 = 6$, are given in Table I. For comparable reasons, the cumulated MSEs for every method were calculated. As can be seen, method 1 provides the largest discrepancy between the true and the estimated parameters (2.634). These results correspond to those of Du [2]. It is proposed, that the difficulties arising with heavily overlapping components, can be solved by adding previous knowledge about the parameters. However, the other methods provide good results in spite of the heavy overlapping. The best results were achieved by method 5 with a cumulated MSE of 0.029. Method 2&3 and 4 provide acceptable values as well, 0.047 and 0.037, respectively.

Further, Table II provides the results of the cumulated MSEs, when $\mu_2$ is adjusted. It can be seen that the more the difference between $\mu_1$ and $\mu_2$ decreases, the more accurate the results from method 1 are. For the other methods the opposite seems to be true. To confirm this assumption, more research is needed. However, over all investigated models, method 5 provides the best results, followed by method 2 and method 4. Comparing method 1 and method 2 directly, which use the same approach, it can be seen, that the approximation in method 2 provides much better results, especially, when the means are close to each other.

TABLE II
CUMULATED MSEs FOR DIFFERENT MIXTURE MODELS

| MSE | Method 1 | Method 2&3 | Method 4 | Method 5 |
|---|---|---|---|---|
| $\mu_2 = 6$ | 2.6340 | 0.0365 | 0.0468 | 0.0292 |
| $\mu_2 = 7$ | 1.8948 | 0.0383 | 0.0458 | 0.0311 |
| $\mu_2 = 8$ | 0.4292 | 0.0438 | 0.0493 | 0.0356 |
| $\mu_2 = 9$ | 0.1371 | 0.0518 | 0.0601 | 0.0423 |
| $\mu_2 = 10$ | 0.0899 | 0.0627 | 0.0717 | 0.0512 |

As the most comparable results were achieved with $\mu_2 = 10$, this model is taken to investigate the influence of different interval widths. Five interval widths were considered: (a) $h = 0.1$, (b) $h = 0.2$, (c) $h = 0.5$, (d) $h = 1$, and (e) $h = 2$. The results are shown in Table III. For small interval widths, like $h = 0.1$ and $h = 0.2$, the results cannot prove major superiority of one method over another, but there is an indication for the slight benefit of method 5. By enlarging the interval width to $h = 1$ and $h = 2$ the benefit increases. While for the other methods the discrepancy between the estimated and true parameters increases for larger intervals, the accuracy of method 5 is retained.

TABLE III
CUMULATED MSEs FOR DIFFERENT INTERVAL WIDTHS

| MSE | Method 1 | Method 2&3 | Method 4 | Method 5 |
|---|---|---|---|---|
| $h = 0.1$ | 0.0721 | 0.0544 | 0.0544 | 0.0543 |
| $h = 0.2$ | 0.0733 | 0.0536 | 0.0535 | 0.0532 |
| $h = 0.5$ | 0.0761 | 0.0561 | 0.0568 | 0.0537 |
| $h = 1.0$ | 0.0899 | 0.0627 | 0.0717 | 0.0512 |
| $h = 2.0$ | 0.3774 | 0.0915 | 0.1908 | 0.0337 |

Finally, the runtime of the methods is presented in Table IV. Apparently, method 1 is the slowest, because of the second iteration in the M-step. The other methods need approximately the same amount of time. Comparing methods 2 and 3 directly, which both provide the same results, it can be seen, that method 3 is slightly faster than method 2. Because of this, and the easier implementation, method 3 should be preferred. Method 5, which provides the most accurate estimations, is just a little bit slower, but nevertheless this is an acceptable amount of time.

TABLE IV
TIME NEEDED FOR THE PERFORMANCE IN SECONDS

| | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 |
|---|---|---|---|---|---|
| Time | 396.62 | 22.09 | 21.97 | 21.57 | 25.97 |

Thus, over all simulations the best results were achieved by method 5. Because of the most accurate estimations, the acceptable time of performance and the straightforward calculation, this method can be recommended for the considered situations. Nevertheless, it needs to be mentioned that because of their easy calculation and their fast computation methods 3 and 4 are also excellent choices to fit a mixture model to grouped data. Both methods use a transformation that avoids the problems that arises with grouped data. The results may depend on special parameter choices, however. Ongoing simulations with various parameter sets are computed to ensure these results.

The application of the considered method may be demonstrated by an example from the material research.

## IV. EXAMPLE

A grain denotes a particle from granular materials. The grain size has an important influence on the material characteristics; therefore its investigation is an important research field. The grain size of microsheets consisting of steel is currently investigated by the SFB 747 “Mikrokaltumformen” in Bremen, Germany. The analyzed data contains intervals, which arise by classification of the grain sizes obtained as areas in a two-dimensional section with interval width $h = 1$. Further, the percentages of grain areas are given. Figure 1 shows a histogram of the grouped data as well as the four fitted mixture models with initial values $\pi_1 = 0.8$, $\pi_2 = 0.2$, $\mu_1 = 9$, $\mu_2 = 10.5$, $\sigma_1 = 1$, and $\sigma_2 = 1.5$. The corresponding estimators are given in Table V.

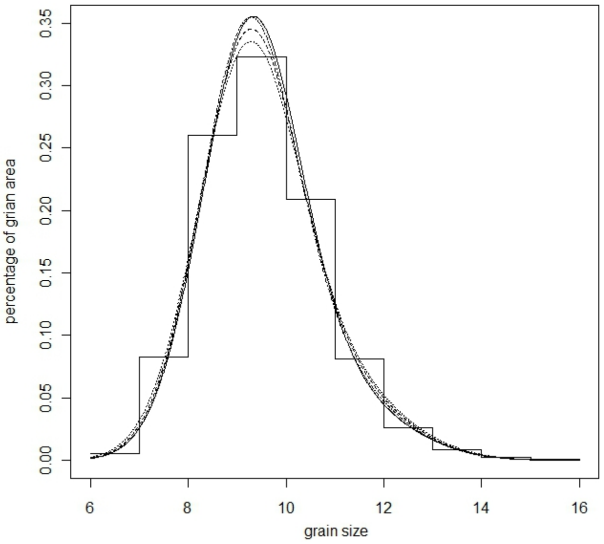

Fig. 1. A two-component Gaussian mixture is fitted to data, adapted from Köhler et. al. [4]. Solid: method 1, dashed: method 2&3, dotted: method 4, dotdashed: method 5.

TABLE V
ESTIMATION

| | $\widehat{\pi}_1$ | $\widehat{\pi}_2$ | $\widehat{\mu}_1$ | $\widehat{\mu}_2$ | $\widehat{\sigma}_1$ | $\widehat{\sigma}_2$ | p-value |
|---|---|---|---|---|---|---|---|
| Method 1 | 0.80 | 0.20 | 9.26 | 10.76 | 0.99 | 1.44 | 1.00 |
| Method 2&3 | 0.76 | 0.24 | 9.20 | 10.72 | 0.99 | 1.40 | 1.00 |
| Method 4 | 0.76 | 0.24 | 9.18 | 10.75 | 1.02 | 1.41 | 1.00 |
| Method 5 | 0.77 | 0.23 | 9.20 | 10.69 | 0.97 | 1.39 | 1.00 |

To investigate the goodness of fit of the models, $\chi^2$-tests were performed. As can be seen in Table V the p-values are close to 1, hence, the null hypothesis can not be rejected for any of the considered models.

If it is attempted to fit just one Gaussian distribution the p-value results in $p = 0.0241$, which indicates a less capable fit.

Hence, it seems that this material consists of two kinds of grain sizes. A consequence of this finding could be the incorporation of two components in the subsequent calculation of material parameters like elasticity and tensile strength.

## V. SUMMARY

A two-component Gaussian mixture model was fitted to grouped data by estimating the parameter via five methods based on the EM algorithm. The compared methods consist of previously published methods as well as new ones that are based on modifications of existing ones. Besides the comparison, the aim of the study was to investigate the influence of different interval widths. For well-separated distributions and small interval width, method 1, using an approach by MacDonald and Green [5], provides acceptable results. However, this method needs significant more time for the performance, and because throughout all simulations this method achieved the largest discrepancy between the estimated and the true values, this method seems to be less appropriate for the considered data sets.

Considering the straightforward calculation, the excellent estimators, and the small amount of time needed for its performance, method 5, using an approach from McLachlan and Jones [10], seems to be the best choice for estimating the parameter of a finite mixture from grouped data. Even with enlarged interval widths, this method provides the most accurate estimators.

Nevertheless, the simulations have shown that the new proposed methods could be recommended as well. Especially for small interval widths, method 3 and 4, which both use a data transformation, provide excellent estimators, and because of their easy implementation and their fast computation, these methods are good choices as well. The comparison of method 2 and 3 has shown, that even though both methods use a different approach, they provide the same results.

Finally, the example from the grain research has shown the practical relevance and it demonstrates the feasibility of the methods.

## REFERENCES

[1] A.P, Dempster, N.M. Laird and D.R. Rubin, “Maximum likelihood from incomplete data,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[2] J. Du, *Combined Algorithms for fitting finite mixture distributions*, McMaster University Hamilton, Ontaria, 2002.

[3] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, London, 1981.

[4] B. Köhler, H. Bomas, M. Hunkel, J. Lütjens and H.-W. Zoch, “Yield strength behaviour of carbon steel microsheets after cold forming and after annealing”, *Sciencedirect*, Scripta Materialia, vol. 62, pp. 548-551, 2010.

[5] P.D.M. MacDonald and T.J. Pitcher, “Age-groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures,” *Journal of Fisheries Research Board of Canada*, Vol. 36, pp. 987-1001, 1979.

[6] P.D.M. MacDonald and P.E.J. Green, *User’s Guide to Program MIX: An Interactive Program for Fitting Mixtures of Distributions*, Release 2.3. Ichthus Data Systems, 1988.

[7] G.J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.

[8] G.L. McLachlan and K.E. Basford, *Mixture Models*, Marcel Dekker, Inc., 1988.

[9] G.L. McLachlan and T. Krishnanm, *The EM Algorithm and Extentions*, Wiley, New York, 1997.

[10] G.J. McLachlan and P.N. Jones, “Algorithm AS 254: Maximum likelihood estimation from grouped and truncated Data with finite normal mixture model”, *Journal of Applied Statistics*, Vol. 39, No.2, pp. 273-312, 1990.

[11] G.J. McLachlan and P.N. Jones, “Fitting Mixture Models to Grouped an Truncated Data via the EM Algorithm”, *Biometrics*, Vol. 44, pp. 571-578, 1988.

[12] M. Schader and F. Schmidt, “Maximum-Likelihood-Schätzung aus gruppierten Daten - eine Übersicht”, *OR Spektrum*, Vol. 10, pp. 1-12, 1988.

[13] D.M. Titteringston, A.F.M. Smith and U.E. Markov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.