Robustness of ML estimators of location-scale mixtures

Christian Hennig[1]

Fachbereich Mathematik - SPST,
Universität Hamburg,
Bundesstr. 50, D-20146 Hamburg, Germany

Abstract. The robustness of ML estimators for mixture models with fixed and estimated number of components s is investigated by the definition and computation of a breakdown point for mixture model parameters and by considering some artificial examples. The ML estimator of the Normal mixture model is compared with the approach of adding a "noise component" (Fraley and Raftery (1998)) and by mixtures of t-distributions (Peel and McLachlan (2000)). It turns out that the estimation of the number of mixture components is crucial for breakdown robustness. To attain robustness for fixed s, the addition of an improper noise component is proposed. A guideline to choose a lower scale bound is given.

1 Introduction

Maximum likelihood (ML)-estimation based on mixtures of Normal distributions (NMML) is a flexible and widely used technique for cluster analysis (see, e.g., Fraley and Raftery (1998)).

Observations $x_1, \ldots, x_n$ are modeled as i.i.d. according to the density

$$f_\eta(x) = \sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x), \text{where } f_{a,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-a}{\sigma}\right), \qquad (1)$$

where $\eta = (s, a_1, \ldots, a_s, \sigma_1, \ldots, \sigma_s, \pi_1, \ldots, \pi_s)$ is the parameter vector, the number of components $s \in I\!N$ may be known or unknown, (a_j, σ_j) pairwise distinct, $a_j \in I\!R$, $\sigma_j > 0$, $\pi_j \geq 0$, $j = 1, \ldots, s$ and $\sum_{j=1}^{s} \pi_j = 1$. For the Normal mixture model, $f = \varphi$ is the density of the standard Normal distribution. Often mixtures of multivariate Normals are used, but for the sake of simplicity, I restrict considerations to the case of one-dimensional data in this paper.

As many other ML-techniques based on the Normal distribution, NMML is not robust against gross outliers, at least if the number of components s is treated as fixed: The estimators of the parameters $a_1, \ldots, a_s$ are weighted means of the observations where the weights for each observation sum up to one (see Redner and Walker (1984)), which means that at least one of these parameters can get arbitrarily large if a single extreme point is added to a dataset.

There are some ideas to overcome the robustness problems of Normal mixture. The software MCLUST (Fraley and Raftery 1998) allows the addition of a mixture component accounting for "noise", modeled as a uniform distribution on the convex hull (the range in one dimension, respectively) of the data, i.e., the data is modeled as generated by

$$f_\zeta(x) = \sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x) + \pi_0 \frac{1(x \in [x_{min}, x_{max}])}{x_{max} - x_{min}}, \tag{2}$$

for given $x_{min}, x_{max} \in I\!R$, where $\zeta = (s, a_1, \ldots, a_s, \sigma_1, \ldots, \sigma_s, \pi_0, \pi_1, \ldots, \pi_s)$, $\pi_0, \ldots, \pi_s \geq 0$, $\sum_{j=0}^{s} \pi_j = 1$ and $1(\ldots)$ is the indicator function. The corresponding ML procedure will be denoted by NMN in the following.

The software EMMIX (Peel and McLachlan (2000)) can be used to fit a mixture of t-distributions instead of Normals by ML (t_νMML), i.e., $f = f_\nu$ being the density of the t-distribution with ν degrees of freedom in (1).

Note that the presented theory will hold if f is any continuous density f that is symmetrical about its only mode 0 and that is > 0 on $I\!R$.

There are some alternatives for robust estimation of mixture components, see McLachlan and Peel (2000, p. 222 ff.) and the references given therein.

While a clear gain of stability can be demonstrated for these methods in various examples (see e.g. Banfield and Raftery (1993), McLachlan and Peel (2000, p. 231 ff.)), there is a lack of theoretical justification of their robustness.

In Section 2, I give a formal definition of a breakdown point for estimators of mixture parameters. The breakdown point goes back to Hampel (1971) and measures the smallest amount of contamination that can spoil an estimator completely.

In Section 3.1, some results about the parameter breakdown of the mixture based clustering techniques are given. The number of components s is assumed to be known here. It is shown that for all techniques introduced above r outliers can make $r < s$ mixture components break down.

To attain a better breakdown behavior, I suggest the maximization of a kind of "improper likelihood" in Section 3.2 where "noise" is modeled by an improper uniform distribution on the real line.

In Section 3.3, the case of an estimated number of mixture components s is treated. I consider s as estimated by the maximization of the Bayesian information criterion BIC(s) (Schwarz (1978)):

$$\text{BIC}(s) = 2L_{n,s}(\eta_{n,s}) - k \log n, \tag{3}$$

where $L_{n,s}$ denotes the log-likelihood function for n points and s mixture components under one of the models (1) or (2) and $\eta_{n,s}$ is the corresponding ML estimator. k denotes the number of free parameters, i.e., $k = 3s - 1$ for (1) and $k = 3s$ for (2). For alternative methods to estimate s, I refer to Chapter 6 of McLachlan and Peel (2000).

With estimated s, all treated methods are able to isolate gross outliers as new mixture components on their own and are therefore very stable against extreme outliers. Breakdown can happen only because additional points inside the area of the estimated mixture components of the original data can lead to the estimation of a smaller number of components.

An important problem in ML estimation for mixture models is the convergence the log-likelihood function to ∞ if one of the σ_j^2 converges to 0. In order to get well defined estimators, the log-likelihood function has to be maximized under a restriction on the scale parameters. The simplest possible restriction is $\min_j \sigma_j \geq \sigma_0 > 0$, which is used to obtain the results given below. The choice of σ_0 is discussed in Section 4.

Some examples are given in Section 5. They illustrate that the stability of the methods depends on the scale restriction and the internal stability of the dataset.

The full theory and all proofs are given in Hennig (2002).

2 Breakdown point definitions

The classical meaning of breakdown for finite samples is that an estimator can be driven as far away from its original value as possible by addition of arbitrarily unfortunate points, usually by gross outliers. Donoho and Huber (1983) distinguish this “addition breakdown point” from breakdown by replacement of points. I consider the former definition here.

Breakdown means that estimators that can take values on the whole range of $\mathbb{R}^p$, can leave every compact set. If the value range of a parameter is bounded, breakdown means that addition of points can take the parameter arbitrarily close to the bound, e.g., a proportion parameter to 0.

A breakdown of an estimator of mixture (or cluster) parameters can be understood in two ways: A situation where at least one of the mixture components explodes is defined as breakdown in Garcia-Escudero and Gordaliza (1999). In contrast to that, Gallegos (2003) defines breakdown in cluster analysis as a situation where *all* clusters explode simultaneously. The definition given here is flexible enough to account for all these situations.

Definition 1. Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of estimators of η in model (1), of ζ in model (2), respectively, on $\mathbb{R}^n$ for fixed $s \in \mathbb{N}$. Let $r \leq s$, $\mathbf{x}_n = (x_1, \ldots, x_n)$ be a dataset, where

$$\forall \hat{\eta} = \arg\max_{\eta} L_{n,s}(\eta, \mathbf{x}_n) : \ \hat{\pi}_j > 0, \ j = 1, \ldots, s. \tag{4}$$

The r**-components breakdown point** of E_n is defined as

$$\begin{gathered} B_{r,n}(E_n, \mathbf{x}_n) = \min_g \{ \tfrac{g}{n+g} : \ \exists j_1 < \ldots < j_r \\ \forall \ D = [\pi_{min}, 1] \times C, \ \pi_{min} > 0, \ C \subset \mathbb{R} \times \mathbb{R}^+ \text{ compact} \\ \exists \ \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}), \ \hat{\eta} = E_{n+g}(\mathbf{x}_{n+g}) : \ (\hat{\pi}_j, \hat{a}_j, \hat{\sigma}_j) \notin D, \ j = j_1, \ldots, j_r \}. \end{gathered}$$

The proportions π_j are defined not to break down if they are bounded away from 0, which implies that they are bounded away from 1 if $s > 1$.

In the situation for unknown s, I restrict considerations to the case of 1-components breakdown. Breakdown means that neither of the s mixture components estimated for $\mathbf{x}_n$ vanishes, nor that any of their scale and location parameters explodes to ∞ under addition of points. Further, breakdown of the proportions π_j to 0 is no longer of interest for estimated s according to the BIC, because if some π_j is small enough, s will simply be estimated as being smaller.

Definition 2. Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of estimators of η in model (1) or of ζ in model (2) on $\mathbb{R}^n$, where $s \in \mathbb{N}$ is unknown and estimated as well. Let $\mathbf{x}_n = (x_1, \ldots, x_n)$ be a dataset. Let s be the estimated number of components of $E_n(\mathbf{x}_n)$. The **breakdown point** of E_n is defined as

$$\begin{gathered} B_n(E_n, \mathbf{x}_n) = \min_g\{\tfrac{g}{n+g} : \forall C \subset \mathbb{R}^s \times (R^+)^s \text{ compact} \\ \exists\ \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}),\ \hat\eta = E_{n+g}(\mathbf{x}_{n+g}) : \\ \text{pairwise distinct } j_1, \ldots, j_s \text{ do not exist, such that } (\hat a_{j_1}, \ldots, \hat a_{j_s}, \hat\sigma_{j_1}, \ldots, \hat\sigma_{j_s}) \in C\}. \end{gathered}$$

This implies especially that breakdown occurs whenever $\hat s < s$, $\hat s$ being the estimated s for $\mathbf{x}_{n+g}$.

An alternative breakdown definition is given by Kharin (1996).

3 Breakdown results

3.1 Breakdown point for fixed s

Let $r < s$. The contribution of r added points $x_{n+1}, \ldots, x_{n+r}$ to the log-likelihood is, for model (1), $\sum_{i=n+1}^{r} \log\left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i)\right)$. It converges to $-\infty$ if the distances among these r points and between them and the original n points converge to ∞, and more than $s-r$ mixture components remain in a compact set about the originally estimated mixture. On the other hand, the log-likelihood is bounded from below, if the r additional points are fitted by r mixture components. This means that r additional points make r mixture components break down. The argument holds as well for NMN because the noise density also converges to 0.

Theorem 1. *Let $\mathbf{x}_n \in \mathbb{R}^n$, $s > 1$. Let $\eta_{n,s}$ be an ML estimator for model (1) or (2). For $r = 1, \ldots, s-1$,*

$$B_{r,n}(\eta_{n,s}, \mathbf{x}_n) \leq \frac{r}{n+r}. \tag{5}$$

For $r = s$, this remains true for the NMML and NMN, while t_νMML has a better s-components breakdown point of $\geq \frac{1}{\nu+1}$, see Hennig (2002).

3.2 An alternative for fixed s

An alternative can be constructed as a modification of NMN. The problem of NMN is that the noise component could be affected by outliers as well, as was shown in the previous section. This can be prevented when the density constant for the noise component is chosen as fixed beforehand, which leads to ML estimation for a mixture where some improper density component is added to catch the noise (NMI). That is, an estimator $\xi_{n,s}$ is defined as the maximizer of

$$L_{n,s}(\xi, \mathbf{x}_n) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i) + \pi_0 b \right), \tag{6}$$

where $b > 0$. This requires the choice of b. If the objective is cluster analysis and there is a maximum scale σ_{max}, above which a mixture component is no longer accepted as a cluster (compare Section 4), b could be chosen as the density value at the 0.025-quantile of $f_{0,\sigma_{max}}$, so that 95% of the points generated from such a distribution have a larger density value for it than for the noise component. For this estimator the breakdown point depends on the stability of the dataset $\mathbf{x}_n$. Breakdown can only occur if additional observations allow that the non-outliers can be fitted with advantage by fewer than s components, and this means that a relatively good solution for $r < s$ components must exist already for $\mathbf{x}_n$. This is formalized in (7). Let $L_{n,s} = L_{n,s}(\xi_{n,s}, \mathbf{x}_n)$. I consider only the breakdown of a single mixture component $B_{1,n}(\xi_{n,s}, \mathbf{x}_n)$.

Theorem 2. *Let $\mathbf{x}_n \in I\!R^n$. Let a_j, σ_j, π_j denote the parameters of $\xi_{n,s}$ and $f_{max} = f(0)/\sigma_0 > b$. If*

$$\max_{r<s} L_{n,r} < \sum_{i=1}^{n} \log \left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i) + (\pi_0 + \frac{g}{n})b \right)$$
$$+ g \log(\pi_0 + \frac{g}{n})b + (n+g) \log \frac{n}{n+g} - g \log f_{max}, \tag{7}$$

then

$$B_{1,n}(\xi_{n,s}, \mathbf{x}_n) > \frac{g}{n+g}. \tag{8}$$

The meaning of (7) is illustrated in Section 5.

3.3 Breakdown point for unknown s

The treatment of s as unknown is favorable for robustness against outliers, because outliers can be fitted by additional mixture components. Generally, for large enough outliers the addition of a new mixture component for each outlier yields a better log-likelihood than any essential change of the original mixture components. That is, gross outliers are almost harmless except that they let the estimated number of components grow.

Breakdown can occur, however, because added points, usually not outlying, but inside the range of the original data, may lead to a preference of a solution with $r < s$ clusters. (9) of Theorem 3 gives a necessary condition for the impossibility of breakdown and may serve as a formalization of the "stability" of an s-components solution for a data set in terms of the differences between the optimal log-likelihoods for s and fewer components.

Theorem 3. *Let $\tau_n = (s, \eta_{n,s})$ be a maximizer of the BIC under (1) or (2). If*

$$\min_{r<s} \left[L_{n,s} - L_{n,r} - \frac{1}{2}(5g + 3s - 3r + 2n)\log(n+g) + n\log n \right] > 0, \qquad (9)$$

then

$$B_n(\tau_n, \mathbf{x}_n) > \frac{g}{n+g}. \qquad (10)$$

The meaning of (9) is illustrated in Section 5.

4 Choice of the scale restrictions

In most applications, sufficient prior information to specify the scale restriction constant σ_0 is not available. A common strategy to avoid a sensible specification of these constants in practice is to compute local maximizers of the log-likelihood from initial values which avoid very small values for the sigmas. This, however, avoids the isolation of single points as clusters, which is crucial for good breakdown behavior for estimated s.

Consider s as unknown. A sensible choice of the restriction constant should fulfill two objectives:

1. The constant should be so large that a data subset that looks like a homogeneous cluster is estimated as one component and no single point of it forms a "one-point-component" with a very small scale.
2. The constant should be so small that a gross outlier generates a new component instead of being merged with an otherwise homogeneous data subset.

α-outliers (with $\alpha > 0$ but very small) are defined by Davies and Gather (1993) with respect to an underlying model as points from a region of low density, chosen so that the probability of the occurrence of an outlier is $\leq \alpha$. For a standard Normal distribution, for example the points outside $[\Phi^{-1}(\frac{\alpha}{2}), \Phi^{-1}(1-\frac{\alpha}{2})]$ are the α-outliers, where Φ_{a,σ^2} denotes the cdf of the Normal distribution with parameters a, σ^2. For $\alpha_n = 1-(1-p)^{1/n}$, the probability of the occurrence of at least one α_n-outlier among n i.i.d. points from $\mathcal{N}(0,1)$ is equal to p.

The strategy is as follows: Choose $p = 0.05$, say, and consider the choice of σ_0 for the NMML with unknown s. The following definition is used to generate reproducible benchmark datasets:

Definition 3. $\Phi^{-1}_{a,\sigma^2}(\frac{1}{n+1}), \ldots, \Phi^{-1}_{a,\sigma^2}(\frac{n}{n+1})$ is called a (a, σ^2)-**Normal standard dataset** (NSD) with n points.

Assume for the moment that at least $n - 1$ points come from a $\mathcal{N}(0, 1)$ distribution. (Denote $c_0 = \sigma_0$ in this particular setup.) c_0 should be chosen so that it is advantageous to isolate an α_n-outlier as its own cluster, but not a non-outlier. This, of course, depends on the non-outlying data. As "calibration benchmark", form a dataset with n points by adding an α_n-outlier to a (0,1)-NSD with $n - 1$ points. Choose c_0 so that BIC(1) =BIC(2) (this can easily be seen to be uniquely possible). For c_0 small enough, the 2-components solution will consist of one component matching approximately the ML-estimator for the NSD and one component fitting only the outlier. Resulting values are given in Table 4.

The interpretation is as follows: Based on $\sigma_0 = c_0$, a dataset consisting of an $(n - 1)$-point NSD and an α_n-non-outlier will be estimated as homogeneous, while there will be more then one cluster if the nth point is an outlier. The same holds for an $n - 1$-point (α, σ^2)-NSD and $\sigma_0 = c_0\sigma$. I suggest the use of $\sigma_0 = c_0\sigma_{max}$, where σ^2_{max} is the largest variance such that a data subset with this variance can be considered as "cluster" with respect to the given application. This may not look like an advantage, because the need to specify a lower bound σ_0 is only replaced by the need to specify an upper bound σ_{max}. But the upper bound has a clear interpretation which does not refer to an unknown underlying truth. At least if the mixture model is used as a tool for cluster analysis, points of a cluster should belong together in some sense, and with respect to a particular application, it can usually be said that points above a certain variation can no longer be considered as "belonging together".

A dataset to analyze will usually not have the form "NSD plus outlier", of course. The clusters in the data will usually be smaller than $n - 1$ points, and they will have a variance smaller than σ^2_{max}. Assume now that there is a homogeneous data subset of $n_1 < n$ points with variance $\sigma^2 \leq \sigma^2_{max}$. The question arises if an α_{n_1}-outlier, non-outlier, respectively, will be isolated from the cluster in the presence of other clusters elsewhere. σ_0 is calculated on the base of the BIC penalty for 1 vs. 2 clusters with n points. That is, the difference in penalty is $3 \log n$. Table 4 also gives the c_0-values computed with an NSD of size $n_1 = n/2 - 1$ plus $\alpha_{n/2}$-outlier and of size $n_1 = n/5 - 1$ plus $\alpha_{n/5}$-outlier, but again with penalty difference $3 \log n$ to show which restriction constant would be needed to isolate at least $\alpha_{n/2}$-outliers, $\alpha_{n/5}$-outliers, respectively, from the homogeneous subset of size n_1 under the assumption that the parameters for the rest of the data remain unaffected. The values coincide satisfactorily with the values computed for n, so that these values look reasonable as well for small homogeneous subsets.

With a variance smaller than σ_{max}, an α-outlier with $\alpha > \alpha_n$ is needed to be isolated from a cluster with a variance smaller than σ_{max}, i.e., the broad tendency is that components with larger variances are preferred over one-point-components.

n	20	50	100	200	1000
c_0	2.10e-2	4.99e-3	1.66e-3	5.51e-4	4.34e-5
$n_1 = n/2 - 1$	9	24	49	99	499
c_0	2.15e-2	5.25e-3	1.76e-3	5.87e-4	4.57e-5
$n_1 = n/5 - 1$	3	9	19	39	199
c_0	2.25e-2	5.44e-3	1.88e-3	6.35e-4	4.93e-5

Table 1. Minimum scale restriction factor c_0 for Normal mixtures. Note that $\log c_0$ is almost exactly linear in $\log n$, so that further values can easily be obtained by interpolation.

Although the argumentation is only valid for NMML with estimated s, I tentatively suggest to apply the resulting values also for the other methods, because the derivation of analogous strategies for them rises certain difficulties.

5 Examples

Consider a dataset of 50 points, namely a (0,1)-NSD with 25 points combined with a (5,1)-NSD with 25 points. Let $\sigma_{max} = 5 \Rightarrow \sigma_0 = 0.025$, $b = 0.0117$ for NMI. For NMML, t_νMML with $\nu \geq 1$, NMN and NMI, always components corresponding almost exactly to the two NSDs are optimal under $s = 2$ fixed. How large must an additional outlier be chosen so that the 50 original points fall into only one cluster and the second mixture component fits only the outlier? For NMML, breakdown begins with an additional point at about 15.2 (13.3; values in parentheses are for $\sigma_0 = 0.001$ to demonstrate the dependence of the robustness on σ_0). For t_3MML, the outlier must lie at about 800 (350), t_1MML needs the outlier at about $3.8e6$ ($8e5$), and NMN breaks down with an additional point at $3.5e7$ ($1.5e6$). The lower breakdown bound (8) of NMI evaluates to $\frac{2}{52}$. The original components are joined by three outliers at 9.8 (while NMN can be broken down by fewer outliers, it would need three outliers to be placed not until 70 to join the origninal components). If the (5,1)-NSD is replaced by a (50,1)-NSD, the lower breakdown bound of NMI is $\frac{7}{57}$ and experimentally 11 outliers at 100, say, are needed for breakdown. Turning back to the combination of the (0,1)-NSD and the (5,1)-NSD, for $\sigma_0 = 0.001$, the lower breakdown bound reduces to $\frac{1}{52}$, and two outliers at 9.8 suffice to join the original components.

Note that NMN is “practically” robust in the sense that it can cope with more than one large outlier, as long as they are below $3.5e7$ and scattered enough. For example, if 7 points $1e3, 5e3, 1e4, 5e4, 1e5, 5e5, 1e6$ are added to the original 50 points, all 7 outliers are classified as noise ($\sigma_0 = 0.025$; the same holds for NMI). To a certain extent this also applies to t_1MML. The seven additional outliers given above lead to breakdown, while outliers at $(100, 200, 500, 1000, 2000, 5000, 10000)$ do still not join the original components.

With estimated s, (10) gives a lower breakdown bound of $\frac{2}{52}$ for NMML and NMN and $\frac{3}{53}$ for t_1MML at the original 50 points ($s = 2$ is estimated correctly by all methods). These bounds are rather conservative. Empirically, 13 points equally spaced between 1.8 and 3.2 lead to breakdown by $\hat{s} = 1$ for NMML and NMN. t_1MML is a bit more stable: the mentioned 13 additional “inliers” lead to the estimation of $\hat{s} = 3$. Extreme outliers always get their own new mixture components. It is interesting that the breakdown point can be driven above $\frac{1}{2}$ by enlarging the separation between the components. For a (0,0.001)-NSD of 25 points and a (100000,0.001)-NSD of 25 points, NMML's lower breakdown bound is $\frac{58}{108}$. Empirically a breakdown point larger than 0.9 can be reached by much less separation.

Consider as a last example a (0,1)-NSD of 45 points combined with a (5,1)-NSD of 5 points. For fixed $s = 2$, NMN needs an outlier at $2e6$ to join the original two components corresponding to the NSD. t_1MML interprets the (5,1)-NSD as extreme points belonging to the (0,1)-NSD and isolates outliers down to 7.5 as one-point-components. While this setup may seem to be less stable than the constellation with two clusters of 25 points each, NMML joins an outlier up to 40 with the (5,1)-NSD and NMI breaks down with at least 3 outliers at 11 (compared to 9.8 above) at a lower breakdown bound of $\frac{2}{52}$.

For estimated s, NMML needs 12 points between the components to join them (at a lower breakdown bound of $\frac{2}{52}$), while NMN and t_1MML estimate the original 50 points as only one regular component, while the (5,1)-NSD is estimated as noise, belonging to the only component, respectively, so that there is no second mixture component which could break down.

Note that the results of this section have been computed by using the EM-algorithm (see, e.g., McLachlan and Peel (2000)) several times with initial configurations chosen by use of prior information about the generation of the data. Not all of the likelihood maxima will be reproduced by default applications of available software.

6 Conclusion

A finite-sample-addition breakdown point for estimators of the parameters of mixture models has been defined for a known and unknown number of mixture components. It has been shown that the ability to estimate the number of mixture components is crucial to attain a satisfactory breakdown point for ML estimators. For fixed s, a better breakdown behaviour can be attained by adding an improper uniform density to the likelihood. Note that the robustness behaviour for fixed s is relevant in practice, because even if the number of components is estimated, there is usually an upper bound on s for computational reasons. For example, for a dataset of 1000 points, one will often estimate s under the restriction $s \leq 10$, say, while there may be much more than 10 outliers. Therefore, NMI, NMN, or t_1MML are recommended in spite of the breakdown robustness of the simple NMML under estimated s. However, NMI, NMN and t_νMML may not recognize mixture components supported by too few points.

Breakdown and robustness in mixture models and cluster analysis do not only depend on the method, but also on the internal stability of the clustering of the dataset.

References

BANFIELD, J. D. and RAFTERY, A. E. (1993): Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics, 49, 803–821.*

DAVIES, P. L. and GATHER, U. (1993): The identification of multiple outliers. *Journal of the American Statistical Association, 88, 782–801.*

DONOHO, D. L. and HUBER, P. J. (1983): The notion of breakdown point. In P. J. Bickel, K. Doksum, and J. L. Hodges jr. (Eds.): *A Festschrift for Erich L. Lehmann*, Wadsworth, Belmont, CA, 157–184.

FRALEY, C. and RAFTERY, A. E. (1998): How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *Computer Journal, 41, 578–588.*

GALLEGOS, M. T. (2003): Clustering in the Presence of Outliers. In: M. Schwaiger, O. Opitz (Eds.): *Exploratory Data Analysis in Empirical Research.* Springer, Berlin, 58–66.

GARCIA-ESCUDERO, L. A. and GORDALIZA, A. (1999): Robustness Properties of k Means and Trimmed k Means. *Journal of the American Statistical Association, 94, 956–969.*

HAMPEL, F. R. (1971): A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics, 42, 1887–1896.*

HENNIG, C. (2002): *Breakdown points for Maximum Likelihood-estimators of location-scale mixtures*, Research Report No. 105, Seminar für Statistik, ETH-Zürich, `ftp://ftp.stat.math.ethz.ch/Research-Reports/105.html`. Submitted.

KHARIN, Y. (1996): *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers, Dordrecht.

MCLACHLAN, G. J. and PEEL, D. (2000): *Finite Mixture Models*, Wiley, New York.

PEEL, D. and MCLACHLAN, G. J. (2000): Robust mixture modeling using the t distribution. *Statistics and Computing, 10, 335–344.*

REDNER, R. A. and WALKER, H. F. (1984): Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review, 26, 195–239.*

SCHWARZ, G. (1978): Estimating the dimension of a model, *Annals of Statistics, 6, 461–464.*