Robust fitting of mixtures using the Trimmed Likelihood Estimator

N. Neykov, P. Filzmoser, R. Dimova, P. Neytchev

PII: S0167-9473(06)00501-9
DOI: doi:10.1016/j.csda.2006.12.024
Reference: COMSTA 3572

To appear in: *Computational Statistics & Data Analysis*

Cite this article as: N. Neykov, P. Filzmoser, R. Dimova and P. Neytchev, Robust fitting of mixtures using the Trimmed Likelihood Estimator, *Computational Statistics & Data Analysis*, doi:10.1016/j.csda.2006.12.024

Robust fitting of mixtures using the Trimmed Likelihood Estimator

N. Neykov [a,*], P. Filzmoser [b], R. Dimova [c], P. Neytchev [a]

[a] *National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko chaussee, 1784 Sofia, Bulgaria.*

[b] *Vienna University of Technology, Department of Statistics and Probability Theory, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria*

[c] *Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics, Sofia, Bulgaria*

Abstract

The Maximum Likelihood Estimator (MLE) has commonly been used to estimate the unknown parameters in a finite mixture of distributions. However, the MLE can be very sensitive to outliers in the data. In order to overcome this the Trimmed Likelihood Estimator (TLE) is proposed to estimate mixtures in a robust way. The superiority of this approach in comparison with the MLE is illustrated by examples and simulation studies. Moreover, as a prominent measure of robustness, the Breakdown Point (BDP) of the TLE for the mixture component parameters is characterized. The relationship of the TLE with various other approaches that have incorporated robustness in fitting mixtures and clustering are also discussed in this context.

Key words: Maximum Likelihood Estimator, Trimmed Likelihood Estimator, Breakdown Point, Finite mixtures of distributions, Robust clustering, Outlier detection

1 Introduction

Finite mixtures of distributions have been widely used to model a wide range of heterogeneous data. In most applications the mixture model parameters

* Corresponding author. Tel.: +3592-8072731, Fax: +3592-9733569
Email addresses: Neyko.Neykov@meteo.bg (N. Neykov), P.Filzmoser@tuwien.ac.at (P. Filzmoser), rdimova@fmi.uni-sofia.bg (R. Dimova), Plamen.Neytchev@meteo.bg (P. Neytchev).

are estimated by the MLE via the expectation-maximization (EM) algorithm (see e.g. McLachlan and Peel, 2000). It is well known, however, that the MLE can be very sensitive to outliers in the data. In fact, even a single outlier can completely ruin the MLE which in mixture settings means that at least one of the component parameters estimate can be arbitrarily large. To overcome this, robust parametric alternatives of the MLE have been developed, e.g., Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987). A direct fitting of mixture models to data by these robust estimators is of limited use. The reason is that these robust estimators are designed to fit a parametric model to the majority of the data whereas the remaining data which do not follow the model are considered as outliers. In practice, however, the data could be quite heterogeneous without having a homogeneous part consisting of at least 50% of the data. Fortunately, since the EM algorithm is capable to transfer a complex mixture MLE problem into relatively simple single component MLE problems, some of the ideas of robust estimation have been adapted to mixture models. Details can be found in Campbell (1984), Kharin (1996), Davé and Krishnapuram (1997), Medasani and Krishnapuram (1998), McLachlan and Peel (2000), Hennig (2003), just to name a few. In this way robustness has been adapted to meet the problem with outliers in mixtures of the location-scale family of distributions. Generally speaking, robust fitting of mixtures of distributions outside this family has not been developed yet. Exceptions are Markatou (2000) and Neykov et al. (2004) who discussed fitting mixtures of Poisson regressions based on the weighted MLE and Trimmed Likelihood Estimator (TLE) via simulations.

Thus, after many years of parallel development of fitting mixtures, cluster analysis, outlier detection and robust techniques, the need for a synthesis of some of these methods beyond the location scale family of distributions has become apparent. Such a synthesis can be a flexible and powerful tool for an effective analysis of heterogeneous data. So, the aim of this paper is to make a step toward the achievement of this goal by offering a unified approach based on the TLE methodology. Because the TLE accommodates the classical MLE, the finite mixture methodology based on the MLE can be adapted and further developed. In this paper the superiority of this approach in comparison with the MLE is illustrated.

The paper is organized as follows. In Section 2, the basic properties of the weighted TLE are presented. In Section 3 we briefly discuss the EM algorithm and explain how robustness can be incorporated. Moreover, the TLE software implementation and adjustments to the framework of mixtures with existing software are presented. Comparisons of the MLE and TLE by examples and simulations are presented in Section 4. Finally, in Section 5 the conclusions are given.

2 The Trimmed Likelihood methodology

Definitions. The Weighted Trimmed Likelihood Estimator (WTLE) is defined in Hadi and Luceño (1997) and in Vandev and Neykov (1998) as

$$\hat{\theta}_{WTLE} := \arg\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta), \tag{1}$$

where $f(y_{\nu(1)};\theta) \le f(y_{\nu(2)};\theta) \le \ldots \le f(y_{\nu(n)};\theta)$ for a fixed θ, $f(y_i;\theta) = -\log\varphi(y_i;\theta)$, $y_i \in I\!R^q$ for $i = 1,\ldots,n$ are i.i.d. observations with probability density $\varphi(y;\theta)$, which depends on an unknown parameter $\theta \in \Theta^p \subset I\!R^p$, $\nu = (\nu(1),\ldots,\nu(n))$ is the corresponding permutation of the indices, which depends on θ, k is the trimming parameter and the weights $w_i \ge 0$ for $i = 1,\ldots,n$ are nondecreasing functions of $f(y_i,\theta)$ such that at least $w_{\nu(k)} > 0$.

The basic idea behind trimming in (1) is the removal of those $n-k$ observations whose values would be highly unlikely to occur if the fitted model was true. The combinatorial nature of the WTLE is emphasized by the representation

$$\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta) = \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f(y_i;\theta) = \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f(y_i;\theta),$$

where I_k is the set of all k–subsets of the set $\{1,\ldots,n\}$. Therefore, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE, and the WTLE is given by the partition with the minimal negative log likelihood.

The WTLE accommodates: (i) the MLE if $k = n$; (ii) the TLE if $w_{\nu(i)} = 1$ for $i = 1,\ldots,k$ and $w_{\nu(i)} = 0$ otherwise; (iii) the Median Likelihood Estimator (MedLE) if $w_{\nu(k)} = 1$ and $w_{\nu(i)} = 0$ for $i \ne k$; If $\varphi(y;\theta)$ is the multivariate normal density function then the MedLE and TLE coincide with the MVE and MCD estimators (Rousseeuw and Leroy, 1987). If $\varphi(y;\theta)$ is the normal regression error density, the MedLE and TLE coincide with the LMS and LTS estimators (Rousseeuw and Leroy, 1987). Details can be found in Vandev and Neykov (1993, 1998). General conditions for the existence of a solution of (1) can be found in Dimova and Neykov (2004) whereas the asymptotic properties are investigated in Cizek (2004). The Breakdown Point (BDP) (i.e. the smallest fraction of contamination that can cause the estimator to take arbitrary large values) of the WTLE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ for some constant d which depends on the density considered, see Müller and Neykov (2003). The choice of d in mixture settings will be discussed in Section 3.

The FAST-TLE algorithm. Computing the WTLE is infeasible for large data sets because of its combinatorial nature. To get an approximative TLE so-

lution an algorithm called FAST-TLE was developed in Neykov and Müller (2003). It reduces to the FAST-LTS and FAST-MCD algorithms considered in Rousseeuw and Van Driessen (1999a,b) in the normal regression or multivariate Gaussian cases, respectively. The basic idea behind the FAST-TLE algorithm consists of carrying out finitely many times a two-step procedure: a trial step followed by a refinement step. In the trial step a subsample of size k^* is selected randomly from the data sample and then the model is fitted to that subsample to get a trial ML estimate. The refinement step is based on the so called concentration procedure: (a) The cases with the k smallest negative log likelihoods based on the current estimate are found, starting with the trial MLE as initial estimator. (Instead of the trial MLE any arbitrarily plausible value can be used.); (b) Fitting the model to these k cases gives an improved fit. Repeating (a) and (b) yields an iterative process. The convergence is always guaranteed after a finite number of steps since there are only finitely many k–subsets out of $\binom{n}{k}$. At the end of this procedure the solution with the lowest trimmed likelihood value is stored. There is no guarantee that this value will be the global minimizer of (1) but one can hope that it would be a close approximation to it. The trial subsample size k^* should be greater than or equal to d which is necessary for the existence of the MLE but the chance to get at least one outlier free subsample is larger if $k^* = d$. Any k within the interval [d, n] can be chosen in the refinement step. A recommendable choice of k is $\lfloor (n+d+1)/2 \rfloor$ because then the BDP of the TLE is maximized according to Müller and Neykov (2003). The algorithm could be accelerated by applying the partitioning and nesting techniques as in Rousseeuw and Van Driessen (1999a,b). We note that if the data set is small all possible subsets with size k can be considered.

3 Finite mixtures and robustness

To make the robust approaches in mixture and cluster settings more understandable we will briefly sketch the MLE within these frameworks based on the EM algorithm. For more details see McLachlan and Peel (2000).

The MLE and EM algorithm. Let (y_i, x_i^T) for $i = 1, \ldots, n$ be a sample of i.i.d. observations such that y_i is coming from a mixture of distributions $\psi_1(y_i; x_i, \theta_1), \ldots, \psi_g(y_i; x_i, \theta_g)$ conditional on $x_i \in I\!R^p$, in proportions $\pi_1, \ldots, \pi_g$ defined by

$$\varphi(y_i; x_i, \Psi) = \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j), \tag{2}$$

where $\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1, \ldots, \theta_g)^T$ is the unknown parameter vector. The proportions satisfy the conditions $\pi_j > 0$ for $j = 1, \ldots, g$, and $\sum_{j=1}^{g} \pi_j = 1$. The MLE of Ψ is defined as a maximum of the log likelihood

$$\log L(\Psi) = \sum_{i=1}^{n} \log\Big\{\sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j)\Big\}. \tag{3}$$

Under certain assumptions on $\psi_j(y_i; x_i, \theta_j)$ for $j = 1, \ldots, g$ the MLE of Ψ exists and belongs to a compact set. However, the resulting MLE is not reasonable if these assumptions are violated. Usually (3) is not maximized directly. The EM algorithm is a standard technique to obtain the MLE of Ψ. It is assumed that each observation (y_i, x_i^T) is associated with an unobserved state $z_i = (z_{i1}, z_{i2}, \ldots, z_{ig})^T$ for $i = 1, \ldots, n$, where z_{ij} is one or zero, depending on whether y_i does or does not belong to the jth component. Treating (y_i, x_i^T, z_i^T) as a complete observation, its likelihood is given by $P(y_i, x_i, z_i) = P(y_i, x_i|z_i)P(z_i) = \prod_{j=1}^{g} \psi_j(y_i; x_i, \theta_j)^{z_{ij}} \pi_j^{z_{ij}}$. Therefore the complete-data log-likelihood is defined by

$$\log L_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij}\{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{4}$$

Considering the z_{ij} as missing the EM algorithm proceeds iteratively in two steps, called the E-step and M-step for expectation and maximization respectively. The E-step on the $(l+1)$th iteration involves the calculation of the conditional expectation of the complete-data log-likelihood, given the observed data (y_i, x_i^T) and using the current estimate $\Psi^{(l)}$ of Ψ,

$$Q(\Psi; \Psi^{(l)}) = \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; x_i, \Psi^{(l)})\{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}, \tag{5}$$

where $\tau_j(y_i; x_i, \Psi^{(l)}) = \pi_j^{(l)} \psi_j(y_i; x_i, \theta_j^{(l)}) / \sum_{h=1}^{g} \pi_h^{(l)} \psi_h(y_i; x_i, \theta_h^{(l)})$ is the current estimated posterior probability that y_i belongs to the jth mixture component. The function $Q(\Psi; \Psi^{(l)})$ minorizes $\log L(\Psi)$, i.e., $Q(\Psi; \Psi^{(l)}) \leq \log L(\Psi)$ and $Q(\Psi^{(l)}; \Psi^{(l)}) = \log L(\Psi^{(l)})$. The M-step in the $(l+1)$th iteration maximizes $Q(\Psi; \Psi^{(l)})$ with respect to Ψ. This yields a new parameter estimate $\Psi^{(l+1)}$. These two steps are alternately repeated until convergence occurs.

The maximization problem can be simplified as (5) can be seen to consist of two parts. The first depends only on the parameters $\pi_1, \ldots, \pi_{g-1}$ whereas the second part depends only on $\theta_1, \ldots, \theta_g$. Consequently, the prior probabilities π_j are updated by

$$\pi_j^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_j(y_i; x_i, \Psi^{(l)}) \tag{6}$$

and the expression for θ_j is maximized,

$$\max_{\theta_1,\ldots,\theta_g} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j), \tag{7}$$

considering the posterior probabilities $\tau_j(y_i; x_i, \Psi^{(l)})$ as the prior weights. Under the assumption that θ_j (for $j = 1, \ldots, g$) are distinct a priori, expression (7) is maximized for each component separately,

$$\max_{\theta_j} \sum_{i=1}^{n} \tau_j(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j), \quad \text{for } j = 1, \ldots, g. \tag{8}$$

In case that θ_j are non-distinct, many techniques exist to reformulate (7) by single summations, see McLachlan and Peel (2000).

The classification EM algorithm. This approach consists of assigning the observation (y_i, x_i^T) to the hth component if $\tau_h(y_i; x_i, \Psi^{(l)}) \geq \tau_j(y_i; x_i, \Psi^{(l)})$ for $j = 1, \ldots, g$. In case of equal estimated posterior probabilities an observation is assigned arbitrarily to one of the components. Hence instead of (8) the following expression is maximized

$$\max_{\theta_j} \sum_{i=1}^{n_j} \log \psi_j(y_i; x_i, \theta_j) \qquad \text{for } j = 1, \ldots, g, \tag{9}$$

where n_j is the jth cluster sample size and $n_1 + n_2 + \ldots + n_g = n$. This is a k–means-type algorithm which converges in a finite number of iterations. The resulting estimates are neither MLE nor consistent, see McLachlan and Peel (2000). However, they could be used as starting values in the EM algorithm.

The expressions (8) and (9) are standard MLE problems. In this way the EM algorithm decomposes complex MLE problems into more simple ones that can be solved by widely available software packages.

The Breakdown Point of the WTLE in mixture settings. As a consequence of the EM algorithm, the BDP of the WTLE in mixture settings can be characterized via the BDP of the trimmed version of (5), the trimmed conditional expectation of the complete-data negative log-likelihood estimator

$$\min_{\Psi} \min_{I \in I_k} \sum_{i \in I} \sum_{j=1}^{g} -\tau_j(y_i; x_i, \Psi^{(l)}) \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{10}$$

Here only the BDP of the WTLE for the parameters θ_j for $j = 1, \ldots, g$ will be treated because the BDP for $\pi_1, \ldots, \pi_g$ needs special consideration. Therefore

the fullness index d of the set $F_\theta = \{\sum_{j=1}^{g} -\log \psi_j(y_i; x_i, \theta_j) \text{ for } i = 1, \ldots, n\}$ has to be characterized using the d–fullness technique of Vandev and Neykov (1993), and Müller and Neykov (2003). It can be proved easily that the fullness index of F_θ is equal to $d = \sum_{j=1}^{g} d_j$ under the assumption that θ_j are distinct a priori and the sets $F_{\theta_j} = \{-\log \psi_j(y_i; x_i, \theta_j) \text{ for } i = 1, \ldots, n\}$ are d_j–full for $j = 1, \ldots, g$. Derivation of the fullness index of any of the sets F_{θ_j} is a routine task. Consequently, there always exists a solution of the optimization problem (10) if k^* and k are within the interval $[\sum_{j=1}^{g} d_j, n]$. If k satisfies $\left\lfloor (n + \sum_{j=1}^{g} d_j)/2 \right\rfloor \leq k \leq \left\lfloor (n + \sum_{j=1}^{g} d_j + 1)/2 \right\rfloor$ the BDP of the WTLE is maximized and equal to $\frac{1}{n} \left\lfloor (n - \sum_{j=1}^{g} d_j)/2 \right\rfloor$. Generally, the fullness index of F_θ is less than the above in case of non-distinct parameters. The fullness indices d_j are equal if $\psi_j(y_i; x_i, \theta_j)$ for $j = 1, \ldots, g$ belong to the same distribution family, e.g., $d_j = p+1$ in the p-variate normal case (Vandev and Neykov, 1993). Therefore the BDP of the WTLE in mixtures of p-variate heteroscedastic normals is equal to $\frac{1}{n} \lfloor (n - g(p+1))/2 \rfloor$. The index of fullness of a mixture of p-variate homoscedastic normals is $g + p$ and thus the BDP of the WTLE in this setting is equal to $\frac{1}{n} \lfloor (n - g - p)/2 \rfloor$. The WTLE reduces to the weighted MCD estimator in both cases if $g = 1$ whereas the BDPs coincide with the BDP of the MCD estimator which is equal to $\frac{1}{n} \lfloor (n - p - 1)/2 \rfloor$. The same holds for mixtures of multiple normal and Poisson regressions with intercept and rank p of the covariates matrix. If the data are not in general position (which is often the case with mixtures of GLMs) this number should be much larger, at least $g(\mathtt{N}(X) + 2)$, see Müller and Neykov (2003) for the definition of $\mathtt{N}(X)$.

Robust fitting of mixtures. If one is able to perform all k–subsets MLE fits of n cases for the mixture model (2) then the WTLE could be found. As this is infeasible for large n the FAST-TLE algorithm can be used to get an approximation. The FAST-TLE algorithm is a general approach for robust estimation and thus any MLE procedure for fitting mixtures can be used. However, the usage of the EM algorithm has a number of conceptual advantages. For instance, fitting mixtures of p-variate normals by the FAST-TLE using the classification EM algorithm reduces to the cluster analysis estimation techniques described by Garcia-Escudero et al. (2003), Gallegos and Ritter (2005), and Hardin and Rocke (2004) under the restriction that the covariance matrices are spherical, homoscedastic and heteroscedastic, respectively. FAST-TLE fitting mixture of normal regressions using the classification EM algorithm would coincide with carrying out cluster-wise regression by the FAST-LTS algorithm of Rousseeuw and Van Driessen (1999a).

Generally, other techniques for robust fitting of mixtures or clustering can be derived by replacing the g standard MLE problems in (8) or (9) by appropriate g robust estimation problems. This idea was adapted by Campbell (1984) in robustly fitting mixtures of normals involving the M-estimators (Huber, 1981)

of location and scale. The usage of M-estimators for the cluster-wise multiple linear regression case is discussed by Hennig (2003).

Software adjustments of the FAST-TLE to mixtures. Since the trial and refinement steps are standard MLE procedures, the FAST-TLE algorithm can be easily implemented using widely available software. We illustrate this in the framework of mixtures of linear regression models, multivariate heteroscedastic normals, and Poisson regressions using the program FlexMix of Leisch (2004). FlexMix was developed in R (`http:/www.R-project.org`) as a computational engine for fitting arbitrary finite mixture models, in particular, mixtures of GLMs and model-based cluster analysis by using the EM algorithm.

In the mixture setting with g components, the trial sample size k^* must be at least $g(p+1)$ to overcome the degenerated case of unbounded likelihood. Thus we recommend a larger trial sample size to increase the chance to allocate at least $p+1$ cases to each mixture component. If this is not the case, any program would fail to get an estimate that could serve as a trial estimate. If this happens a new random subsample of k^* observations has to be drawn and supplied to the software estimation procedure. This trial and error process continues until a trial estimate is derived. The refinement subsample size k has to be $\lfloor (n+g(p+1))/2 \rfloor$ to ensure the highest BDP of the TLE. If the expected percentage α of outliers in the data is a priory known, a recommendable choice of k is $\lfloor n(1-\alpha) \rfloor$ to increase the efficiency of the TLE.

Most of the software procedures for fitting mixtures, in particular the FlexMix program, maximize the expression (8) or (9) according to the user specified weight option. For instance, if the hard weighting option is specified then the classification EM algorithm is performed by FlexMix. We recommend this option within the trial step only. Hence depending on the weight option various algorithms can be designed.

As a final remark we note that in the refinement steps the negative log likelihoods $-\log \varphi(y_i; x_i, \Psi)$ defined by (2) are evaluated at the current estimate $\hat{\Psi}$ and then sorted in ascending order to get the indices of those k cases with the smallest negative log-likelihoods, starting with the trial estimate Ψ^* of Ψ at the first iteration of the refinement step. In practice, we need 4 or 5 refinement steps at most to reach convergence.

4 Examples

In the examples below we compare MLE and FAST-TLE approaches using the program FlexMix as a computational MLE and FAST-TLE procedure. Sometimes FlexMix returns less components than initially specified. This is

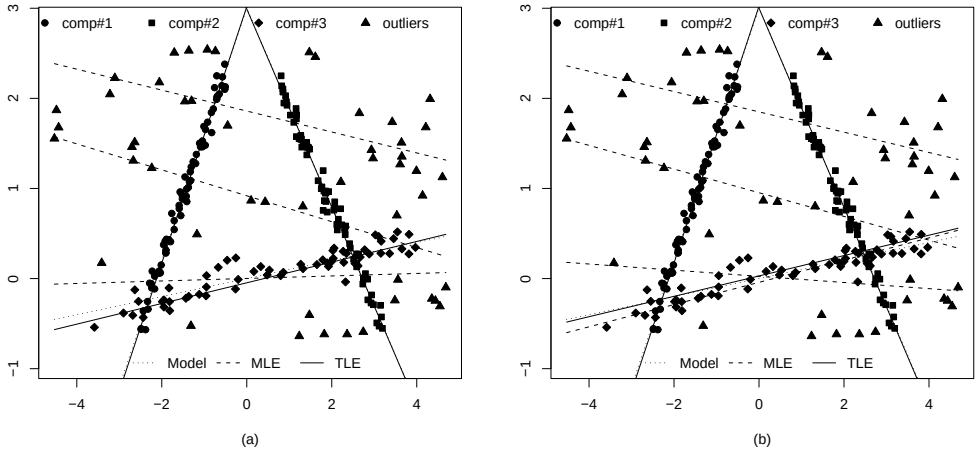

Fig. 1. Mixture of three regressions: true model (dotted lines), fits based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 20% trimming and 3 components, (b) 40% trimming and 4 components.

because FlexMix allows a removal of components containing less observations than a user specified percentage to overcome numerical instabilities. Since the true number of mixture components is unknown in practice, FlexMix is always run with various numbers of components. The Bayesian Information Criterion (BIC) based on the MLE and FAST-TLE can then be used to determine the number of mixture components. In this way we can assess the quality of the fits as in our examples the number of components and their parameters are known. A fit is considered as successful if *all* components are correctly estimated even if some non-sense fits occur additionally. Correct estimation means that at least 95% of the observations that are assigned to a particular component are indeed generated from this model.

Mixture of three regression lines with noise

In this example we consider a mixture of three simple normal linear regressions with additional noise. The regression lines were generated according to the models $y_{1i} = 3+1.4x_i+\epsilon_i$ (70 data points), $y_{2i} = 3-1.1x_i+\epsilon_i$ (70 data points), and $y_{3i} = 0.1x_i + \epsilon_i$ (60 data points), where x_i is uniformly distributed in the intervals [-3,-1] and [1,3], respectively, and ϵ_i is a standard normal distribution with $\sigma = 0.1$. To these data we added 50 outliers uniformly distributed in the area $[-4.5, 4.5] \times [-0.8, 2.8]$. The points that follow the models are marked by rhombs, squares and bullets whereas the outliers are marked by triangles. The plots in Figure 1 are typical results of the MLE and FAST-TLE fits. The dotted, dashed and solid lines correspond to the true models, MLE and FAST-TLE fits, respectively. Starting with an increasing percentage of trimming from 20 to 45 and number of components from 2 to 5 the FAST-TLE algorithm converged to the correct two components mixture model in almost all trials whereas the MLE failed.

Mixture of three bivariate normal models with noise

By this example the behavior of the FAST-TLE is studied for the simulated data set discussed in McLachlan and Peel (2000). This data consists of 100 observations generated from a 3-component bivariate normal mixture model with equal mixing proportions and component parameters, respectively as

$$\mu_1 = (0 \quad 3)^T, \quad \mu_2 = (3 \quad 0)^T, \quad \mu_3 = (-3 \quad 0)^T,$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & .5 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & .5 \end{pmatrix}.$$

Fifty outliers, generated from a uniform distribution over the range -10 to 10 on each variate are added to the original data. Thus a sample of 150 observations is obtained. McLachlan and Peel (2000) model this data by a mixture of t–distributions and reduce the influence of the outliers.

The original observations, the outliers, the 3 components MLE and FAST-TLE fits with 15%, 25%, 35% and 45% trimming are presented in Figure 2 (a)–(d). The original observations, i.e., data that follow the models are marked by rhombs, squares and bullets whereas the outliers are marked by triangles. The dotted contours of the ellipses on the plots correspond to the true models whereas the solid and dashed contours of the 99% confidence ellipses correspond to the FAST-TLE and MLE fits, respectively. For the robust fits we can see that a lower or higher trimming percentage than the true contamination level still allows the correct estimation of the ellipsoid centers while the covariances are overestimated or underestimated due to the too high or low trimming percentage. The fits are excellent if the specified trimming percentage is close to the true percentage of contamination. The classical MLE fits are poor when using 3 or even more components.

Generally, in real applications the number of mixture components is unknown and the BIC is widely used for model assessment. The trimmed analog of BIC is defined by $\text{TBIC} = -2\log(TL_k(\tilde{\Psi})) + m\log(k)$, where $TL_k(\tilde{\Psi})$ is the maximized trimmed likelihood, k is the trimming parameter, and m is the number of parameters in the mixture model. Obviously, TBIC reduces to BIC if $k = n$. To get an impression of the empirical distribution of these quantities for this example a limited Monte Carlo simulation study was conducted for a range of different situations. We fit the simulated three bivariate mixtures of normals with 1 to 5 components and vary the trimming percentage from 0% to 45% in steps of 5%. The experiment was independently replicated 500 times for any combination. The resulting TBIC median values (rounded) are presented in Table 1. The smallest values for each column are marked in *italics*. One can see that these values stabilize in a model with 3 components which is the correct model. A two-phase regression fit of the 3rd row values

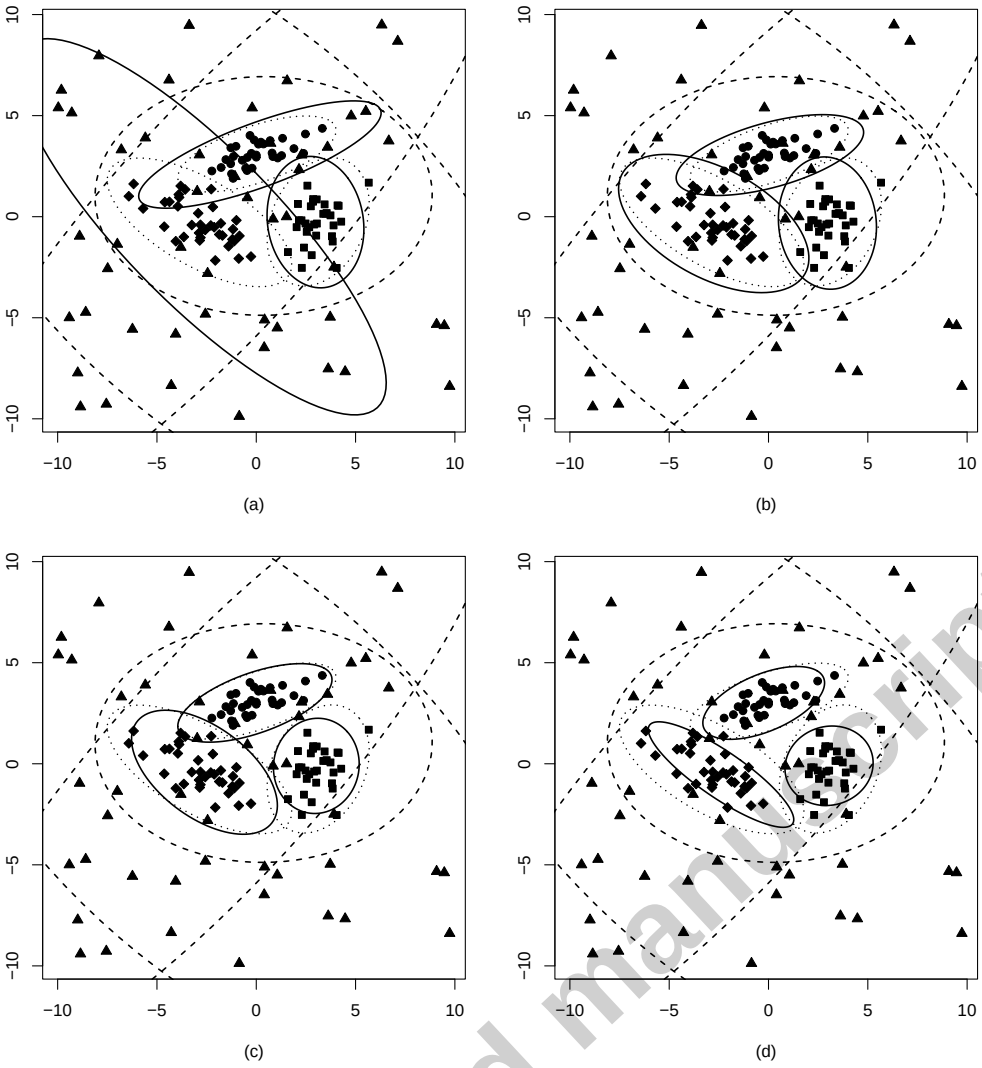

Fig. 2. Data set of McLachlan and Peel (2000) with mixtures of 3 normals with noise: true model (dotted lines) and fits of a three component normal mixture based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 15%, (b) 25%, (c) 35%, and (d) 45% of trimming.

against the trimming percentages detects a change point between 25% and 30% trimming which could be interpreted as a data contamination estimate. We note that the true contamination level in this data set is slightly higher, however, a part of the noise observations conforms the mixture model. From this and other similar studies we could conclude that the TBIC might be used to assess robustly the number of mixture components and the percentage of contamination in the data.

Mixture of two Poisson regression models with noise

In this example we consider two Poisson regression models with equal mixing proportions and with additional noise. For each Poisson regression model 100

Table 1
Simulation experiment for the data of McLachlan and Peel (2000): resulting TBIC median values (rounded) based on different numbers of components (rows) and different trimming percentages (columns).

	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
1	1672	1510	1382	1253	1119	1003	915	837	749	650
2	1654	1494	1338	1202	1054	920	822	734	643	559
3	1585	1436	1313	1190	*1047*	*902*	*795*	*709*	*620*	*538*
4	1595	*1429*	*1304*	*1178*	1040	908	807	720	631	549
5	*1594*	1430	1309	1184	1051	922	822	736	647	566

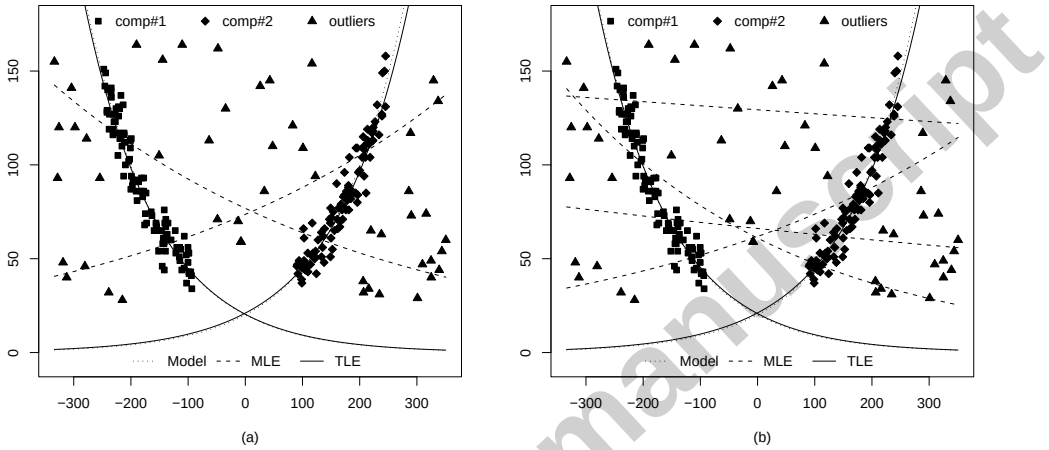

Fig. 3. Mixture of two Poisson regression components: true model (dotted lines), fits based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 20% trimming and 2 components, and (b) 40% trimming and 4 components.

data points are generated according to the Poisson distribution with means $\log \lambda_1 = 3 - 0.008x$ and $\log \lambda_2 = 3 + 0.008x$, where x is uniformly distributed in the intervals [-225,-25] and [25,225], respectively. For the noise we generated 50 points from a uniform distribution over the range of each variate. The plots in Figure 3 are typical results of the MLE and FAST-TLE fits for a simulated data set. The points that follow the models are marked by squares and rhombs whereas the outliers are marked by triangles. The dotted, dashed and solid lines correspond to the true models, MLE and TLE fits, respectively. Starting with an increasing number of components from 2 to 5 the FAST-TLE algorithm converged to the correct two components mixture model in most of the trials whereas the MLE failed, see Figure 3.

In order to get more insight we generated 100 independent data samples according to the above model. Each data set was fitted by a mixture model with 2, 3, 4 and 5 components and with 20% trimming. Similar to the previous ex-

amples, the estimated number of components as returned by FlexMix can be smaller than initially specified. For each considered model we count how often a model with a certain number of components is returned among all simulated data sets. The results for the MLE and FAST-TLE are reported in Table 2. The number of specified components is presented by the rows in the table, and the number of returned components by the columns. Additionally to the frequencies we provide the number of successful fits (number below in *italics*), i.e., both Poisson regression components of the mixture model were correctly estimated. For the MLE method we see that the chance for successful fits increases only for a larger required number components. Overall, the method has severe problems in estimating the models since only 37 out of 400 fits were successful. For FAST-TLE the increase of the initial number of components has almost no effect, since a model with 2 components is optimal in more than 90% of the fits. Moreover, these models are almost always successful fits. In total, 392 out of the 400 experiments were successful.

Table 2
Simulation results for the mixture of two Poisson regressions. Models with 2, 3, 4, and 5 components were fitted for 100 simulated data sets. Out of 400 fits, 37 were successful for MLE and 392 were correctly estimated by FAST-TLE.

	MLE returned components					FAST-TLE returned components				
started	2	3	4	5	Total	2	3	4	5	Total
2	100				100	100				100
	1				*1*	*98*				*98*
3		100			100	93	7			100
		2			*2*	*92*	*7*			*99*
4		94	6		100	96	4	0		100
		4	*4*		*8*	*94*	*4*			*98*
5		19	15	66	100	94	6		0	100
		3	*7*	*16*	*26*	*91*	*6*			*97*
Total	100	213	21	66	400	383	7	0	0	400
	1	*9*	*11*	*16*	*37*	*375*	*17*			*392*

5 Summary and conclusions

The TLE methodology can be used for robustly fitting mixture models. We have demonstrated by examples and simulations that in presence of outliers the TLE gives very reliable estimates comparable to the mixture model generating parameters. Applying the FAST-TLE algorithm to mixtures boils down to

carrying out the classical MLE on subsamples. Procedures for mixture models based on the MLE are widely available and thus the method is easy to implement. Software in R is available from the authors upon request. The TBIC is a useful indicator for determining the number of components and contamination level in the data. If the trimming percentage is chosen too large, some of the observations that follow the model will be trimmed and incorrectly identified as outliers. Therefore an additional inspection of the FAST-TLE posterior weights can be helpful in distinguishing these observations from real outliers. The TLE will lead to greater computational effort, but having in mind the growing power of modern-day processors and memory, one can afford this.

Acknowledgment

The authors would like to thank anonymous referees for many interesting and helpful comments that led to a clearer presentation of the material.

References

Campbell N.A., 1984. Mixture models and atypical values. Math. Geology 16, 465 – 477.

Cizek, P., 2004. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. (Discussion Paper No. 130), Tilburg University, Center for Economic Research.

Davé, R., Krishnapuram, R., 1997. Robust clustering methods: a unified view. IEEE Transactions on Fuzzy Systems 5, 270 – 293.

Dimova, R., Neykov, N.M., 2004. Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel. pp. 83–91.

Gallegos, M.T., Ritter, G., 2005. A robust methods for cluster analysis. The Ann. Statist. 33, 347 – 380.

Garcia-Escudero, L.A., Gordaliza, A., Matran, C., 2003. Trimming tools in exploratory data analysis. Journal of Computational and Graphical Statistics, 12, 434 – 449.

Hadi, A.S., Luceño, A., 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput. Statist. and Data Analysis 25, 251 – 272.

Hampel, F.R., Ronchetti, E. M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust statistics. The approach based on influence functions. Wiley, New York.

Hardin, J., Rocke, D.M., 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput. Statist. and Data Analysis 44, 625 – 638.

Hennig, C., 2003. Clusters, outliers, and regression: fixed point clusters. J. of Multivariate Analysis 86, 183 – 212.

Huber, P., 1981. Robust statistics. John Wiley & Sons, New York.

Leisch, F., 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. J. of Statist. Soft. 11, http://www.jstatsoft.org/

Kharin, Yu. S., 1996. Robustness in Statistical Pattern Recognition. Kluwer Academic Publishers, Dordrecht, London.

Markatou, M., 2000. Mixture models, robustness, and the weighted likelihood methodology. Biometrics 56, 483 – 486.

Medasani, S., Krishnapuram, R., 1998. Robust mixture decomposition via maximization of trimmed log-likelihood with application to image database organization. In: Proceedings of the North American Fuzzy Information Society Workshop, Pensacola, August 1998, pp. 237 – 241.

McLachlan, G.J., Peel, D., 2000. Finite mixture models. Wiley, New York.

Müller, C.H., Neykov, N.M., 2003. Breakdown points of the trimmed likelihood and related estimators in generalized linear models. J. Statist. Plann. Inference 116, 503 – 519.

Neykov, N.M., Müller, C.H., 2003. Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in robust statistics. Physica-Verlag, Heidelberg, pp. 277 – 286.

Neykov, N.M., Filzmoser, P., Dimova, R., Neytchev, P.N., 2004. Mixture of generalized linear models and the Trimmed Likelihood methodology. In: Antoch (Ed.), Proceedings in Computational Statistics. Physica-Verlag, pp. 1585 – 1592.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust regression and outlier detection. Wiley, New York.

Rousseeuw, P.J., Van Driessen, K., 1999a. Computing least trimmed of squares regression for large data sets. Estadistica 54, 163 – 190.

Rousseeuw, P.J., Van Driessen, K., 1999b. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212 – 223.

Vandev, D.L., Neykov, N.M., 1993. Robust maximum likelihood in the Gaussian case. In: Ronchetti, E., Stahel, W.A. (Eds.), New directions in data analysis and robustness. Birkhäuser Verlag, Basel, pp. 259 – 264.

Vandev, D.L., Neykov, N.M., 1998. About regression estimators with high breakdown point, Statistics 32, 111 – 129.