COMPSTAT'2004 Symposium

MIXTURE OF GLMS AND THE TRIMMED LIKELIHOOD METHODOLOGY

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

Key words: Trimmed likelihood estimator, robust estimator, mixture of GLMs, FlexMix, R.

COMPSTAT 2004 section: Robustness.

Abstract: The Maximum Likelihood Estimator (MLE) has been widely used to estimate the unknown parameters in the finite mixture of Generalized Linear Models (GLMs). However, the MLE can be very sensitive to outliers in the data. In this paper we consider an approach based on the Trimmed Likelihood Estimator (TLE) to estimate mixtures of GLMs in a robust way. The superiority of this approach in comparison with the MLE is illustrated through a simulation study.

1 Introduction

Finite mixture of distributions have been widely used to model a wide range of heterogeneous data, e.g., [15] or [29]. In most applications the mixture model parameters are estimated by the MLE. It is well known, however, that the MLE can be very sensitive to outliers in the data. In fact, even a single outlier can ruin completely the MLE. To overcome this, robust parametric alternatives of the MLE have been developed, e.g., [6], [7], [8], [11], [4], [14], [19], [21], [28]. Few of these alternatives have been used in fitting finite mixtures of GLMs. For instance mixtures of Poissons and normals based on the weighted MLE technique are discussed in [13]. To reduce the outliers influence on parameter estimates of a mixture of two normals, the Median of the negative Likelihood Estimator (MedLE) is recommended in [24], the Breakdown Point (BP) properties of which were studied in [25] and [26].

An indirect technique for the detection of interesting multiple structures in data by means of redescending M-estimators is suggested in [16] tracing all possible solutions to the M-estimating equations. This approach is extended further in [10]. Another way of doing robust estimation in mixtures of location-scale models has been the replacement of the classical multivariate location and scatter with their robust counterparts based on M and MCD estimates, as in [1], [5] and [20]. Mixtures of t-distributions are recommended in [15], but this approach is not resistant against leverage points. Thus robustness has been adapted to meet some problems with outliers in clustering and the clusterwise regression, a particular case of mixtures of GLMs. Generally speaking, robust fitting of mixtures has not been well developed yet.

Thus, after many years of parallel development of cluster analysis, outlier detection and robust techniques, the need for a synthesis between all of them has become apparent. It was demonstrated in [9] and [20] that 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 trimmed likelihood methodology for fitting finite mixtures of distributions. The superiority of this approach in comparison with the MLE is illustrated through a simulation study in the mixtures framework of GLMs.

2 Trimmed likelihood methodology

The Weighted Trimmed Likelihood estimator is defined in [7] and [27] as

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

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

The basic idea behind the trimming in this estimator is in the removal of those $n-k$ observations whose values would be highly unlikely to occur if the fitted model was true. Due to the representation $\min\limits_{\theta\in\Theta^p} \sum\limits_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta) = \min\limits_{\theta\in\Theta^p} \min\limits_{I\in I_k} \sum\limits_{i\in I} w_i f(y_i;\theta) = \min\limits_{I\in I_k} \min\limits_{\theta\in\Theta^p} \sum\limits_{i\in I} w_i f(y_i;\theta)$, where I_k is the set of all k–subsets of the set $\{1,\dots,n\}$, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE. The WTL_k estimator is given by that partition with that MLE fit for which the negative log likelihood is minima.

The WTL_k estimator reduces to: (i) the MLE if $k = n$; (ii) the TLE if $w_{\nu(i)} = 1$ for $i = 1,\dots,k$ and $w_{\nu(i)} = 0$ otherwise, the MedLE if $w_{\nu(k)} = 1$ and $w_{\nu(i)} = 0$ otherwise, e.g., [19]. If $\varphi(y,\theta)$ is the multivariate normal density function then the MedLE and TLE coincide with MVE and MCD estimators of [21], if $\varphi(y,\theta)$ is the normal regression error density the MedLE and TLE coincide with the LMS and LTS estimators of [21]. Detailes can be found in [26] and [27].

General conditions for the existence of a solution of (1) can be found in [3] whereas the consistency is proved in [2]. In the GLMs framework, the BP properties of (1) are studied in [17]. For the particular cases of normal, logistic and log-linear regression models it is proved that the BP of the WTL_k estimator is $\frac{1}{n}\min\{n-k+1, k-\mathcal{N}(X)\}$, where $\mathcal{N}(X) := \max_{0\ne\beta\in R^p} \mathrm{card}\left\{i\in\{1,\dots,n\};\ x_i^\top\beta = 0\right\}$ is the maximum number of carriers $x_i \in R^p$ lying in a subspace, $X := (x_i^\top)$ is the carriers data matrix and $x_i^\top\beta$ is the so called linear predictor. If x_i are linearly independent then $\mathcal{N}(X) = p-1$. The BP can be exemplified by the range of the values of k.

For increasing k the estimator will possess a BP point less than the highest possible but it will be more efficient at the same time.

Computing the WTL_k estimator is infeasible for large data sets because of its combinatorial and nonlinear optimization nature. To get approximate TLE an algorithm called FAST-TLE was developed in [18]. It reduces to the FAST-LTS/LMS/LTA or FAST-MCD/MVE algorithms considered in [9], [22] and [23] in the normal regression or multivariate Gaussian cases. 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. The cases with the k smallest negative log likelihoods from the trial fit are found. Fitting the model to these k cases gives an improved fit. Repeating the improvement step 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 $\binom{n}{k}$ in all. At the end of this procedure the solution with the lowest value of (1) 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 $\mathcal{N}(X)+1$ which is needed for the existence of the MLE but the chance to get at least one outlier free subsample is larger if $k^* = \mathcal{N}(X)+1$. Any k within the interval [N(X)+1, n] can be chosen in the refinement step. A recommendable choice of k is $\lfloor (n+\mathcal{N}(X)+1)/2 \rfloor$ because then the BP of the TLE is maximized (see, [17]). The algorithm could be accelerated by applying the partitioning and nesting techniques as in [22] or [23]. We note that if the data set is small all possible subsets with size k can be considered.

3 Finite mixture of GLMs

Now a short reminder to mixtures will be given. Details can be found in [15]. Let (y_i, x_i) for $i = 1, \ldots, n$ be a sample of i.i.d. observations such that y_i is coming from a mixture of $\psi_1(y_i; x_i, \theta_1), \ldots, \psi_g(y_i; x_i, \theta_g)$ distributions, conditional on the variables $x_i \in R^p$, 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 log likelihood is given by $\log L(\Psi) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j)\}$. The EM algorithm is a standard technique to obtain the MLE of Ψ. It

consists in maximizing the complete data log likelihood given by

$$\log L_c(\Psi) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij}\{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}, \tag{3}$$

where z_{ij} denote the component-indicator variables, depending on whether y_i does or does not belong to the jth component. The algorithm proceeds iteratively, alternating the E and M steps. In the $(l+1)$th iteration of the E-step the posterior probabilities for each observation are computed as $\hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)}) = \pi_j^{(l)}\psi_j(y_i; x_i, \theta_j^{(l)})/\sum_{j=1}^{g}\pi_j^{(l)}\psi_j(y_i; x_i, \theta_j^{(l)})$. In the $(l+1)$th M-step iteration the prior probabilities are computed by $\pi_j^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)})$ and then the function is maximized

$$\max_{\theta_1,\ldots,\theta_g} \sum_{j=1}^{g}\sum_{i=1}^{n} \hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j). \tag{4}$$

For mixtures of GLMs, $\theta_j = h(x^T\beta_j)$, $j = 1, \ldots, g$, the function h is appropriately chosen and under the assumption that the parameters $\beta_1, \ldots, \beta_g$ have no elements in common a priori (4) is maximized for each component separately using the posterior probabilities as weights (see, [15]).

4 Adjustments of the FAST-TLE to mixture of GLMs

The FAST-TLE algorithm can be easily implemented using the environment of software packages such as GLIM, S-PLUS, R, SAS, STATISTICA, etc., since the trial and refinement steps are based on a standard MLE procedure. In the following we illustrate this in the framework of mixtures of GLMs using the program FlexMix as a computational engine for fitting mixtures of GLMs models and model-based cluster analysis in R, described in [12]. The trial and refinement sample sizes k^* and k depend not only on the sample size but also on the number of mixture components and model parameters. As the linear predictor of a standard GLMs consists of an intercept and p carriers the number of the unknown parameters is $p+1$, hence in a mixture with g components this number is $g(p+1)$. Therefore the trial sample size k^* in a g components mixture of GLMs with random carriers has to be at least $g(p+1)$ to ensure the estimability in each component, otherwise $g(\mathcal{N}(X)+2)$. We recommend a larger trial subsample size in order to increase the chance for each component to contain at least $(p+1)$ cases. We also recommend a larger refinement sample size, say 80% or 90% of the data size, as in mixtures the majority of the data have to accommodate several heterogeneous components. Any prior information about the data structure could be useful at this stage. According to the FAST-TLE algorithm a trial MLE, $\widetilde{\Psi}$, is found by maximizing (3) over the trail subsample with size k^* instead of n. In the refinement step we are evaluating (2) at $\widetilde{\Psi}$ for $i = 1, \ldots, n$

and then sorting $f(y_i; x_i, \widetilde{\Psi}) = -\log\varphi(y_i; x_i, \widetilde{\Psi})$ in ascending order to get the indices of the k smallest cases. The improved fit $\widehat{\Psi}$ is then obtained by maximizing (3) over these k cases.

5 Examples

Two artificially generated data sets, the mixture of normal and Poisson regression models, respectively, are shown on the upper two plots of Figure 1. On the left-hand upper plot, the points 1-40 are generated according to $x \sim N(2,1)$, $y = 2 + x + \varepsilon$, $\varepsilon \sim N(0, 0.1)$, whereas the points 41-80 are their mirror pattern, and the points 81-100 are outliers uniformly distributed in the area $[x_{min}, x_{max}]\times[y_{min}, y_{max}]$. On the right-hand upper plot the points 1-48 are generated according to $x \sim U(20, 200)$, $\quad \mathrm{E}y = \log\lambda = 3 + 0.01x$, $\quad y \sim Poisson(\lambda)$, whereas the points 49-96 are their mirror pattern, and the points 97-100 are outliers. In both plots, the points that follow the models are marked by triangles and rhombs whereas the outliers are marked by bullets. The lines on the upper two plots, and their dotted analogs on the other 4 plots correspond to the true models. The continuous lines in the middle and bottom plots correspond to the fits. The upper plots heading values correspond to the negative log likelihood sums based on the whole samples and "good" subsamples evaluated at the true model parameters. The left heading values on the remaining plots correspond to the negative log likelihood and TLE minima based on the whole sample cases. The right heading values correspond to the negative log likelihood sums of the "good" cases evaluated at the MLE and TLE based on the whole samples.

The middle two plots give an impression about the MLE fits due to the program FlexMix starting and ending with a mixture of 4 components. The results of the mixture of Poisson models will be discussed only because of the similarity with the normal case. In fact we performed 4 experiments over the same data set in order to guarantee the reliability of the estimation procedure because of the internal random mechanism of the EM algorithm, respectively the FlexMix program, as regards the initial classification of the data. Each one of these experiments consists of 250 FlexMix runs starting respectively with 2, 3, 4 and 5 specified mixture components to assess the quality of the fits. As a result of these fits 4×250 plots were produced and examined. Only 12 times the mixture components were "correctly identified" which means: (i) on the background of 5 specified components the true components were 8 times nicely fitted in those 250 runs, however, 3 nonsense structures were also identified at the same time; (ii) on the background of 4 specified components the true components were 4 times nicely fitted in those 250 runs, however, 2 nonsense structures were also identified; (iii) in the remaining 500 trials neither a single nor 2 or 3 components of the mixture fit was satisfactory.

The bottom plots give an impression about the TLE fits due to the FAST-TLE algorithm using the FlexMix program with $k^* = 0.1n$ and $k = 0.8n$ in

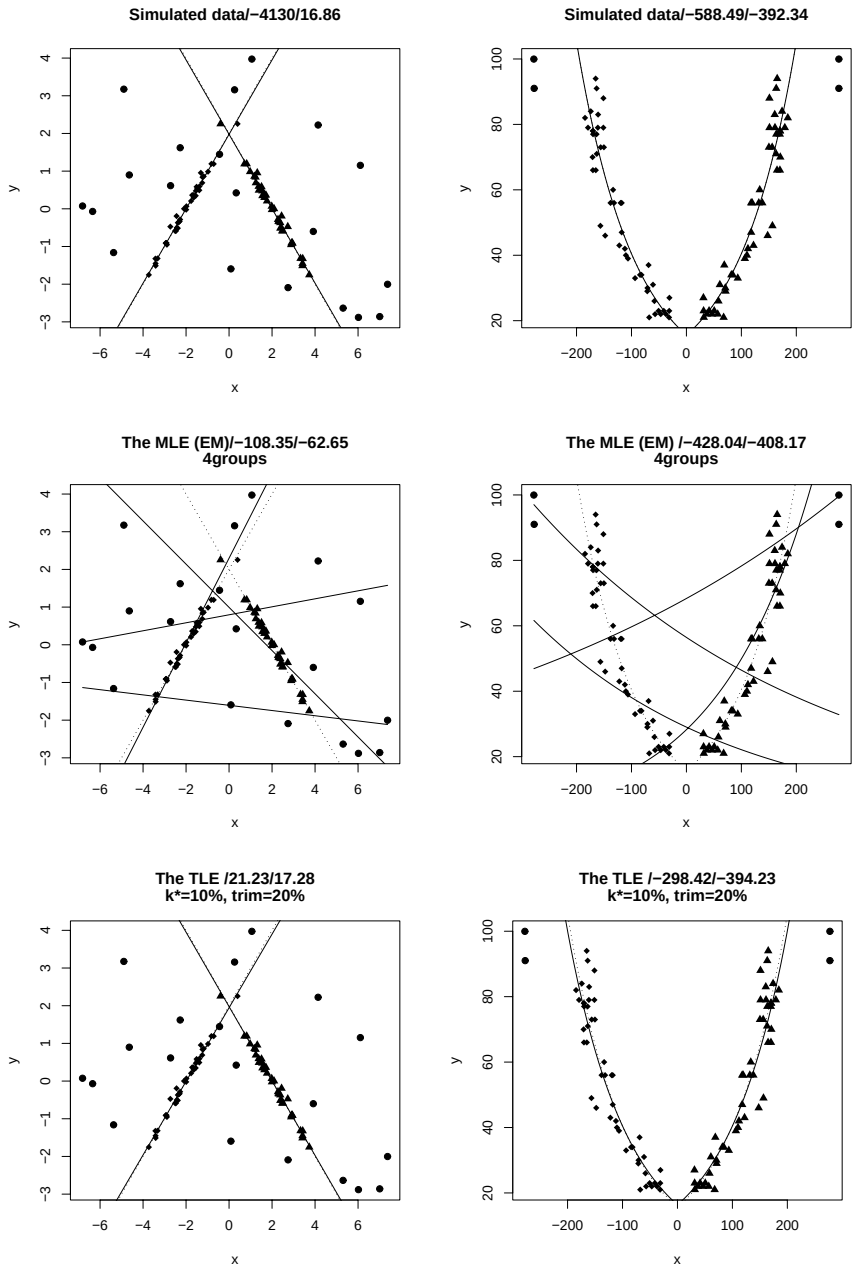

Figure 1: The artificially generated data sets based on mixtures of two normal and two Poisson regression models with outliers are given on the left-hand and right-hand upper plots, respectively. The MLE and TLE fits are given on the middle and bottom plots, respectively.

both mixture types starting with mixtures of 2 components in 500 runs. The true structures were correctly identified in all runs within 30 repetitions of the procedure. Similar results were obtained with $k^* = 0.1n,\ k = 0.7n$. The MLE and TLE behavior was studied over many simulated mixtures of GLMs data with similar designs. The results were similar to the presented here.

References

[1] Campbell N.A. (1984). *Mixture models and atypical values.* Math. Geology **16**,465 – 477.

[2] Cizek P.(2002). *Robust estimation in nonlinear regression and limited dependent variable models.* http://econpapers.hhs.se/paper/wpawuwpem/0203003.htm.

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

[4] Field C., Smith B. (1994). *Robust estimation - a weighted maximum likelihood approach.* Int. Statist. Rev. **62**, 405 – 424.

[5] Gallegos M.T. (2000). *A robust method for clustering analysis.* TR MIP-0013, Fakultät für Mathematik und Informatik, Universität Passau.

[6] Green P.J. (1984). *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.* J. Roy. Statist. Soc. Ser. B **46**, 149 – 192.

[7] Hadi A.S., Luceño A. (1997). *Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms.* Computational Statistics and Data Analysis **25**, 251 – 272.

[8] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust statistics. The approach based on influence functions.* Wiley, NY.

[9] Hawkins D.M., Olive D.J. (2002). *Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm* (with discussions). J. Amer. Statist. Assoc. **97**, 136 – 159.

[10] Hennig C. (2003). *Fixed point clusters.* J. of Classification **19**, 249 – 276.

[11] Huber P. (1981). *Robust statistics.* John Wiley & Sons, New York.

[12] Leisch F. (2003). *FlexMix.* Reference manual: `http://cran.R-project.org/doc/packages/flexmix.pdf.`

[13] Markatou M. (2000). *Mixture models, robustness, and the weighted likelihood methodology.* Biometrics **56**, 483 – 486.

[14] Markatou M., Basu A., Lindsay B. (1997). *Weighted likelihood estimating equations: The discrete case with applications to logistic regression.* J. Statist. Plann. Inference **57**, 215 – 232.

[15] McLachlan G.J., Peel D. (2000). *Finite mixture models.* Wiley, NY.

[16] Morgenthaler S. (1990). *Fitting redescending M-estimators in regression.* In: Robust regression, H. D. Lawrence and S. Arthur (eds.), 105 – 128.

[17] 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.

[18] Neykov N.M., Müller C.H. (2002). *Breakdown point and computation of trimmed likelihood estimators in generalized linear models.* In: Developments in robust statistics, R. Dutter, P. Filzmoser, U. Gather, P. Rousseeuw (eds.), Physica-Verlag, Heidelberg, 277 – 286.

[19] Neykov N.M., Neytchev P.N. (1990). *A robust alternative of the maximum likelihood estimator.* In: Short communications of COMPSTAT'90, Dubrovnik, 99 – 100.

[20] Rocke D.M., Woodruff D.L. (2002). *Computational connections between robust multivariate analysis and clustering.* In: Proc. of COMPSTAT, 255 – 260.

[21] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* Wiley, New York.

[22] Rousseeuw P.J., Van Driessen K. (1999). *Computing LTS regression for large data sets.* Technical report, University of Antwerp, (submitted).

[23] Rousseeuw P.J., Van Driessen K. (1999). *A fast algorithm for the MCD estimator.* Technometrics **41**, 212 – 223.

[24] Tibshirani R., Knight K. (1999). *Bootstrap bumping.* J. Comp. and Graph. Statist. **8**, 671 – 686.

[25] Vandev D.L. (1993). *A note on breakdown point of the least median squares and least trimmed squares.* Statistics and Probability Letters **16**, 117 – 119.

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

[27] Vandev D.L., Neykov N.M. (1998). *About regression estimators with high breakdown point.* Statistics **32**, 111 – 129.

[28] Windham M.P. (1995). *Robustifying model fitting.* J. Roy. Statist. Soc. Ser. B **57**, 599 – 609.

[29] Wedel M., Kamakura W.A. (1998). *Market segmentation: Conceptual and methodological foundations.* Dordrecht: Kluwer Academic Press.

Acknowledgement: The authors thank the referees for valuable comments. The authors would like to thank F. Leisch for the interesting discussions concerning mixtures and the program FlexMix.

Address: N. Neykov, R. Dimova, P. Neytchev, National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko chaussee, Sofia 1784, Bulgaria
P. Filzmoser, Dept. of Statistics & Probability Theory, Vienna, University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria

E-mail: Neyko.Neykov@meteo.bg, P.Filzmoser@tuwien.ac.at