Robust estimation of mixtures of skew-normal distributions

Stima robusta di misture di normali asimmetriche

L.A. García-Escudero, F. Greselin, A. Mayo-Iscar, and G. McLachlan

Abstract Recently, observed departures from the classical Gaussian mixture model in real datasets motivated the introduction of mixtures of skew t, and remarkably widened the application of model based clustering and classification to great many real datasets. Unfortunately, when data contamination occurs, classical inference for these models could be severely affected. In this paper we introduce robust estimation of mixtures of skew normal, to resist sparse outliers and even pointwise contamination that may arise in data collection. Hence, in each component, the skewed nature of the data is explicitly modeled, while any departure from it is dealt by the robust approach. Some applications on real data show the effectiveness of the proposal.

Sommario *Recentemente, a fronte di dataset reali multimodali con asimmetria e code pesanti, è stato introdotto il modello mistura di t asimmetriche, ampliando considerevolmente il campo di applicazione delle classiche misture di distribuzioni Gaussiane. La stima di questi modelli non è però robusta rispetto a contaminazioni e/o errori che possano accadere nella raccolta dei dati. In questo lavoro si introduce uno stimatore robusto per le misture di normali asimmetriche, in grado di resistere a valori anomali e a contaminazione puntuale. La natura asimmetrica dei dati è esplicitamente modellata in ciascuna componente, mentre la stima robusta consente*

L.A. García-Escudero
Department of Statistics and Operations Research and IMUVA, University of Valladolid, Valladolid, Spain, e-mail: lagarcia@eio.uva.es

F. Greselin
Department of Statistics and Quantitative Methods, Milano-Bicocca University, Milano, Italy e-mail: francesca.greselin@unimib.it

A. Mayo-Iscar
Department of Statistics and Operations Research and IMUVA, University of Valladolid, Valladolid, Spain e-mail: agustinm@eio.uva.es

G. McLachlan
Department of Mathematics, University of Queensland, Brisbane, Australia e-mail: g.mclachlan@uq.edu.au

di gestire ogni allontanamento dal modello. Applicazioni su dati reali documentano l'efficacia della proposta.

Key words: Clustering, Robustness, Trimming, Constrained estimation, Skew data, model-based classification, Finite mixture models.

1 Introduction

Finite mixtures of distributions have been widely used as a powerful tool to model heterogeneous data and to approximate complex probability densities, presenting multimodality, skewness and heavy tails. During the last decade, there has been an increasing interest in finding more flexible methods to accurately represent observed data and to reduce unrealistic assumptions. This very active and stimulating context has seen the appearance of many contributions. Among the available proposals in the literature, mixtures of skew normal can incorporate asymmetry in components (see f.i., [2]). On the other hand, mixtures of t distributions can model heavier tails by down-weighting the contribution of extremal observations, as shown in [4, 7]. Mixtures of skew t may accomodate for both asymmetry and leptokurtosis in the grouped data, and therefore remarkably widened the application of model based clustering and classification (see, for example, [6]).

When dealing with model fitting, the elegant theory of likelihood inference provides estimators with desirable properties such as consistency and efficiency. However, these estimators are not robust and there is usually a trade-off between robustness and efficiency. Hence, due to the possible presence of contaminating data (background noise, pointwise contamination, unexpected minority patterns, etc.) a small fraction of outliers, (located far from the groups and, even, between them) could severely affect the model fitting, and a robust approach is needed. Surely, considering skew t distributions is an interesting proposal to achieve robustness with respect to uniform noise or a few sparse outliers. However, Hennig (2004) noted that they are not effective against gross outliers or pointwise contamination that may arise in data collection, their asymptotic breakdown point being zero.

In view of all these considerations, we introduce here a new proposal. To gain effective protection against all type of outliers we jointly use trimming and constrained estimation along the estimation of mixtures of skew Gaussian distributions. We apply our robust estimation to skew Gaussian components (instead of skew t) because they are more parsimonious in parameters and easier in estimation. Indeed the flexibility inherited by trimming does not require any assumption on the heaviness of the tails. The asymptotic breakdown point of the resulting method is strictly positive, an indication of robustness even against gross outliers. As final remark, due to its properties, our methods is offered as a very general tool for clustering heterogeneous skew populations.

2 Finite Mixtures of Canonical Fundamental Skew Normal

We consider here the location-scale variant of the Canonical Fundamental Skew Normal (CFUSN)[1], whose parameters allow to separately govern location, scale, correlation, and skewness. The model arises from a $p+q$ multivariate normal r.v. $(\mathbf{U},\mathbf{V})$, such that

$$\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} \sim \mathcal{N}_{q+p}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{I}_q & 0\\ 0 & \Sigma\end{bmatrix}\right)$$

where Σ is a positive definite scale matrix and 0 is a vector of zeros with appropriate dimension. Then, given a $p \times q$ matrix Δ and a p-dimensional vector μ, we arrive at a stochastic representation for $\mathbf{Y}$, obtained via a convolution, i.e.

$$\mathbf{Y} = \mu + \Delta|\mathbf{U}| + \mathbf{V},$$

which follows the CFUSN distribution, whose density is given by

$$f(\mathbf{y};\mu,\Sigma,\Delta) = 2^q \phi_p(\mathbf{y};\mu,\Omega)\Phi_q\big(\Delta^T\Omega^{-1}(\mathbf{y}-\mu);0,\Lambda\big), \qquad (1)$$

where $\Omega = \Sigma + \Delta\Delta^T$ and $\Lambda = \mathbf{I}_q - \Delta^T\Omega^{-1}\Delta$. As usual, $\phi_p(\mathbf{y};\mu,\Sigma)$ denotes the p-dimensional density of the multivariate Gaussian with mean μ and scale Σ evaluated at $\mathbf{y}$, while $\Phi_q(\cdot)$ denotes the cumulative distribution function. The probability density function for a g-component mixture model of CFUSNs can be written as

$$\sum_{h=1}^{g} \pi_h f(\mathbf{y};\mu_h,\Sigma_h,\Delta_h), \qquad \pi_h \geq 0, \qquad \sum_{h=1}^{g} \pi_h = 1, \qquad (2)$$

where $f(\mathbf{y};\mu_h,\Sigma_h,\Delta_h)$ denotes the h^{th} skew normal component with location parameter μ_h, scale matrix Σ_h and skew parameter Δ_h, given in (1). We denote the unknown parameter by $\theta = (\theta_1,\dots,\theta_g)$, with $\theta_h = (\pi_h,\mu_h,\Sigma_h,\Delta_h)$ related to component h, π_h being the group weights, and adopt the acronym FM-CFUSN for (2).

3 Robust estimation for FM-CFUSN

Aiming at achieving robustness and obtaining good breakdown properties for the ML estimators, a constructive way to obtain a robust estimation is given by providing a feasible EM algorithm for model fitting, where we incorporate impartial trimming, just before the E-step, and constrained estimation along the M-step. The key idea in *trimming* is that a small portion of observations, which are highly unlikely to occur under the current fitted model, is discarded from contributing to the mixture estimates. In the maximization, therefore, we consider the following *trimmed* log-likelihood function [3, 8]

$$\ell_{trim} = \sum_{j=1}^{n} \zeta(\mathbf{y}_j) \log \left[\sum_{h=1}^{g} \phi_p(\mathbf{y}_j; \mu_h, \Omega_h) \Phi_q(\Delta_h^T \Omega_h^{-1}(\mathbf{y}_j - \mu_h); 0, \Lambda_h) \pi_h \right]. \quad (3)$$

By $\zeta(\cdot)$ we denote a 0-1 trimming indicator function that indicates whether observation $\mathbf{y}_j$ is trimmed off: $\zeta(\mathbf{y}_j)$=0, or not: $\zeta(\mathbf{y}_j)$=1. A fixed fraction α of observations, whose contributions to the likelihood are lower than their α-quantile, will be unassigned by setting $\sum_{j=1}^{n} \zeta(\mathbf{y}_j) = [n(1-\alpha)]$ just before each E-step, in such a way that they do not influence the parameter estimation (by $[\cdot]$ we denote the integer part of the argument). Hence α denotes the *trimming level.*

Furthermore - and this will be our second step - we implement a *constrained ML estimation* for the Σ_h matrices in the components of the mixture. The ML estimates $\hat{\theta}$ based on a set of i.i.d. observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is now rephrased into

$$\hat{\theta} = \arg\max_{\theta} \ell_{trim}(\theta|\mathbf{y}), \quad \text{for } \theta \in \Theta, \quad (4)$$

where Θ denotes the parameter space. Also in this case - the same happens for the non-robust case - the defining problem is ill-posed because the log-likelihood tends to ∞ when either $\mu_h = \mathbf{y}_j$ and $|\Sigma_h| \to 0$. As a trivial consequence, the EM algorithm can be trapped into non-interesting local maximizers, called "spurious" solutions.

For this reason, we set a constraint on the maximization of ℓ_{trim}, by imposing

$$\lambda_{l,h} \leq c\,\lambda_{m,k} \quad \text{for} \quad 1 \leq l \neq m \leq p \quad \text{and} \quad 1 \leq h \neq k \leq g \quad (5)$$

where $\{\lambda_{l,h}\}_{l=1,\ldots,p}$ are the eigenvalues of Σ_h, for $h = 1, \ldots, g$ and $1 \leq c < +\infty$.

We will denote by Θ_c the constrained parameter space under requirement (5).

4 Applications to real data

We consider here the Australian Institute of Sports (AIS) dataset, consisting of $p = 11$ physical and hematological measurements on 202 athletes (100 females and 102 males) in different sports, and available within the R package *sn*. Our purpose is to provide a model for the entire dataset, and since the group labels (athletes gender) are provided in advance, the aim is to classify athletes by this feature. By applying the robust FM-CFUSN, with $p = 11$, $q = 1$, 50 starting values and stopping the EM after a maximum of 100 iterations, we got the results shown in Figure 1 (left panel). After the robust estimation, also the 20 trimmed observations can be classified, by using the Bayes' rule and assigning each unit to the component with maximum a posteriori probability, yielding finally to 4 misclassified units (Figure 1, right panel). Notice that this is a very encouraging result when compared to similar approaches available in the literature (see also [5], where a detailed analysis has been done), as the use of an ordinary normal mixture model yields 8 misclassifications, and the t-version of the FM-CFUSN model yields 4 misclassifications.

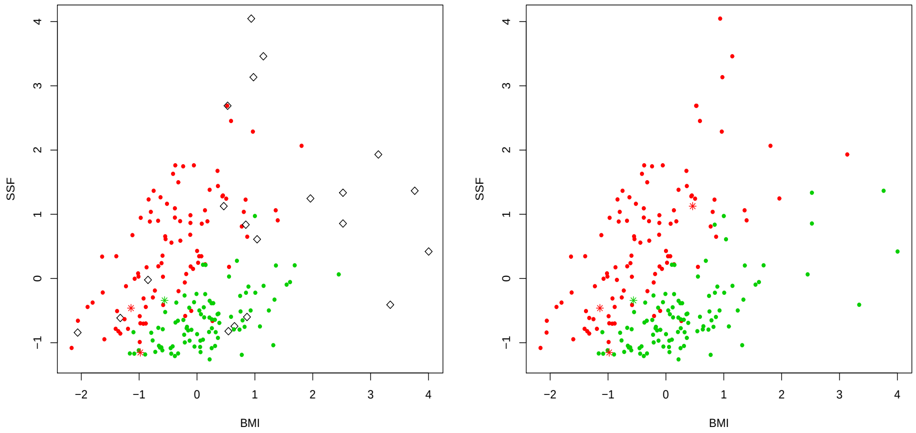

Fig. 1 Classification of the AIS dataset (left panel) by using robust FM-CFUSN (c=16, α=0.1, female data in red, male in green, represented as filled circles when right classification takes place, otherwise as stars; trimmed units are denoted by diamonds). Bivariate plots refer to variables weight/height2 (BMI) and sum of skin folds (SSF)

A second application has been developed on annual financial data of 66 American firms, considering the Ratio of Retained Earnings (RE) to total assets, and ratio of earnings before interest an taxes (EBiT) to total assets. The purpose is to classify firms who filed for bankruptcy. The bivariate sample is plotted in Figure 2, where bimodality and skewness are apparent, thus we fit a two component mixture to the data. We set $\alpha = 0.10$ and c=16. After estimating the model without the contribution of the 7 trimmed units (which are apparently located far from the cores of the components), we classified them as well, arriving at only 4 misclassified firms. This compares to 2 misclassifications with using the t-version of the FM-CFUSN model.

A third application has been done on real world natural images from the Berkeley's image segmentation dataset, where the aim is to segment pixels into background and foreground. We also applied our method to perform automated high-dimensional flow cytometric data analysis on real data. All results show that our method provides an effective approach for asymmetric, heavy tailed data in the mixture components. Simulated results show that the estimation is able to resist to noise as well as to the more dangerous pointwise contamination.

In conclusion, even if further study should be devoted along the lines of the present proposal, we introduced a very general robust tool for clustering heterogeneous skew populations, by using the parsimonious and well-known skew Gaussian model and by flexibly dealing with any departure from the skewed components' cores via the trimming approach.

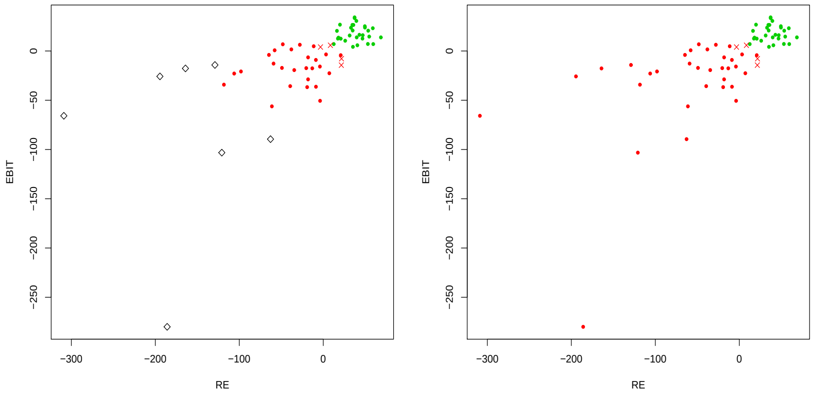

Fig. 2 Classification of the Bankruptcy dataset by using robust FM-CFUSN (bankrupted firms in red, solvent firms in green, represented as filled circles when right classification takes place, otherwise as crosses; In the left panel trimmed units are denoted by diamonds, while in the right panel also trimmed observations have been classified)

References

[1] Arellano-Valle, R. B., Genton, M. G.: On fundamental skew distributions. J. Multiv. Anal., **96** (1), 93–116 (2005).

[2] Branco, M. D., Dey, D. K.: A general class of multivariate skew-elliptical distributions. J. Multiv. Anal., **79** (1), 99-113 (2001).

[3] García-Escudero, L., Gordaliza, A., Mayo-Iscar, A.: A constrained robust proposal for mixture modeling avoiding spurious solutions. Adv. Data Anal. Classif., **8** (1), 27–43 (2014).

[4] Greselin, F., Ingrassia, S.: Constrained monotone EM algorithms for mixtures of multivariate *t* distributions. Stat. Comp., **20** (1), 9–22 (2010).

[5] Lee, S. X., McLachlan, G. J.: Model-based clustering and classification with non-normal mixture distributions. Statistical Methods & Appl., 22 (4), 427–454 (2013).

[6] Lin, T. I.: Robust mixture modeling using multivariate skew t distribution. Stat. Comp., **20**, 343–356 (2010).

[7] McLachlan, G. J., Peel, D.: Robust cluster analysis via mixtures of multivariate *t*-distributions. In: Advances in pattern recognition. Springer Berlin Heidelberg (1998).

[8] Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using the trimmed likelihood estimator. Comp. Stat. & Data Anal., **52** (1), 299–308 (2007).