Journal of Statistical Planning and Inference 164 (2015) 27–38

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Outlier detection and robust mixture modeling using nonconvex penalized likelihood

Chun Yu [a], Kun Chen [b], Weixin Yao [c,*]

[a] School of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013, China
[b] Department of Statistics, University of Connecticut, Storrs, CT 06269, United States
[c] Department of Statistics, University of California, Riverside, CA 92521, United States

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form 9 March 2015
Accepted 10 March 2015
Available online 19 March 2015

Keywords:
EM algorithm
Mixture models
Outlier detection
Penalized likelihood

ABSTRACT

Finite mixture models are widely used in a variety of statistical applications. However, the classical normal mixture model with maximum likelihood estimation is prone to the presence of only a few severe outliers. We propose a robust mixture modeling approach using a mean-shift formulation coupled with nonconvex sparsity-inducing penalization, to conduct simultaneous outlier detection and robust parameter estimation. An efficient iterative thresholding-embedded EM algorithm is developed to maximize the penalized log-likelihood. The efficacy of our proposed approach is demonstrated via simulation studies and a real application on Acidity data analysis.

1. Introduction

Nowadays finite mixture distributions are increasingly important in modeling a variety of random phenomena (see Everitt and Hand, 1981; Titterington et al., 1985; McLachlan and Basford, 1988; Lindsay, 1995; Böhning, 1999). The m-component finite normal mixture distribution has probability density

$$f(y;\boldsymbol{\theta}) = \sum_{i=1}^{m} \pi_i \phi(y;\mu_i,\sigma_i^2), \tag{1.1}$$

where $\boldsymbol{\theta} = (\pi_1,\mu_1,\sigma_1;\ldots;\pi_m,\mu_m,\sigma_m)^T$ collects all the unknown parameters, $\phi(\cdot;\mu,\sigma^2)$ denotes the density function of $N(\mu,\sigma^2)$, and π_j is the proportion of jth subpopulation with $\sum_{j=1}^{m}\pi_j = 1$. Given observations $(y_1,\ldots,y_n)$ from model (1.1), the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is given by,

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \pi_j \phi(y_i;\mu_j,\sigma_j^2)\right\}, \tag{1.2}$$

which does not have an explicit form and is usually calculated by the EM algorithm (Dempster et al., 1977).

The MLE based on the normality assumption possesses many desirable properties such as asymptotic efficiency, however, the method is not robust and both parameter estimation and inference can fail miserably in the presence of outliers. Our focus in this paper is hence on robust estimation and accurate outlier detection in mixture model. For the simpler problem

* Corresponding author.
E-mail addresses: chuckyu0106@126.com (C. Yu), kun.chen@uconn.edu (K. Chen), weixin.yao@ucr.edu (W. Yao).

http://dx.doi.org/10.1016/j.jspi.2015.03.003
0378-3758/

of estimating of a single location, many robust methods have been proposed, including the M-estimator (Huber, 1981), the least median of squares (LMS) estimator (Siegel, 1982), the least trimmed squares (LTS) estimator (Rousseeuw, 1983), the S-estimates (Rousseeuw and Yohai, 1984), the MM-estimator (Yohai, 1987), and the weighted least squares estimator (REWLSE) (Gervini and Yohai, 2002). In contrast, there is much less research on robust estimation of the mixture model, in part because it is not straightforward to replace the log-likelihood in (1.2) by a robust criterion similar to M-estimation. Peel and McLachlan (2000) proposed a robust mixture modeling using t distribution. Markatou (2000) proposed using a weighted likelihood for each data point to robustify the estimation procedure for mixture models. Fujisawa and Eguchi (2005) proposed a robust estimation method in normal mixture model using a modified likelihood function. Neykov et al. (2007) proposed robust fitting of mixtures using the trimmed likelihood. Other related robust methods on mixture models include Hennig (2002, 2003), Shen et al. (2004), Bai et al. (2012), Bashir and Carter (2012), Yao et al. (2014), and Song et al. (2014).

We propose a new robust mixture modeling approach based on a mean-shift model formulation coupled with penalization, which achieves simultaneous outlier detection and robust parameter estimation. A case-specific mean-shift parameter vector is added to the mean structure of the mixture model, and it is assumed to be sparse for capturing the rare but possibly severe outlying effects caused by the potential outliers. When the mixture components are assumed to have equal variances, our method directly extends the robust linear regression approaches proposed by She and Owen (2011) and Lee et al. (2012). However, even in this case the optimization of the penalized mixture log-likelihood is not trivial, especially for the SCAD penalty (Fan and Li, 2001). For the general case of unequal component variances, the variance heterogeneity of different components complicates the declaration and detection of the outliers, and we thus propose a general scale-free and case-specific mean-shift formulation to solve the general problem.

2. Robust mixture model via mean-shift penalization

In this section, we introduce the proposed robust mixture modeling approach via mean-shift penalization (RMM). To focus on the main idea, we restrict our attention on the normal mixture model. The proposed approach can be readily extended to other mixture models, such as gamma mixture and logistic mixture. Due to the inherent difference between the case of equal component variances and the case of unequal component variances, we shall discuss two cases separately.

2.1. RMM for equal component variances

Assume the mixture components have equal variances, i.e., $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$. The proposed robust mixture model with a mean-shift parameterization is to assume that the observations $(y_1, \ldots, y_n)$ come from the following mixture density

$$f(y_i; \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2), \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\theta} = (\pi_1, \mu_1, \ldots, \pi_m, \mu_m, \sigma)^T$, and γ_i is the mean-shift parameter for the ith observation. Apparently, without any constraints, the addition of the mean-shift parameters results in an overly parameterized model. The key here is to assume that the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ is sparse, i.e., γ_i is zero when the ith data point is a normal observation with any of the m mixture components, and only when the ith observation is an outlier, γ_i becomes nonzero to capture the outlying effect. Therefore, the sparse estimation of $\boldsymbol{\gamma}$ provides a direct way to accommodate and identify outliers.

Due to the sparsity assumption of $\boldsymbol{\gamma}$, we propose to maximize the following penalized log-likelihood criterion to conduct model estimation and outlier detection,

$$pl_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \sum_{i=1}^{n} \frac{1}{w_i} P_\lambda(|\gamma_i|) \tag{2.2}$$

where $l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2)\right\}$, w_is are some prespecified weights, $P_\lambda(\cdot)$ is some penalty function used to induce the sparsity in $\boldsymbol{\gamma}$, and λ is a tuning parameter controlling the number of outliers, i.e., the number of nonzero γ_i. In practice, w_is can be chosen to reflect any available prior information about how likely that y_is are outliers; to focus on the main idea, here we mainly consider $w_1 = w_2 = \cdots = w_n = w$, and discuss the choice of w for different penalty functions.

Some commonly used sparsity-inducing penalty functions include the ℓ_1 penalty (Donoho and Johnstone, 1994a; Tibshirani, 1996a,b)

$$P_\lambda(\gamma) = \lambda|\gamma|, \tag{2.3}$$

the ℓ_0 penalty (Antoniadis, 1997)

$$P_\lambda(\gamma) = \frac{\lambda^2}{2} I(\gamma \neq 0), \tag{2.4}$$