Penalized Maximum Likelihood Estimator for Finite Skew-Laplace-Normal Mixtures

Weisan Wu
Xinyu Yang

Skew-Laplace-Normal Mixture models is a more flexible framework than the normal mixture models for heterogeneous data with asymmetric behaviors. But it's likelihood function have some bad math properties, such as the unboundedness of likelihood function and the divergency of skewness parameter, it often mislead statistic inference. In this paper, we given a penalizing the likelihood function method to deal with these problem simultaneously, and we given the detail of proof on parameter have strongly consistency. We also give a modified penalized EM-type algorithms to compute penalized estimators.

Keywords: Skew-Laplace-Normal Mixtures, Penalized maximum likelihood estimator, EM-type algorithms, limiting distribution, non-regularity
Tob Regul Sci.™ 2021;7(4-1): 776-787
DOI: doi.org/10.18001/TRS.7.4.1.32

INTRODUCTION

In the past few decades, finite mixture models have been widely studied and applied in many statistic-based fields, such as density estimation, classification, clustering, and pattern recognition. In a large number of the attendant studies, the Gaussian component is most commonly used.

However, the normality assumptions for component densities could be violated in many fields, with data involving asymmetric and skewness properties fairly common. Azzalini[1] proposed skew normal distribution to fit the data. The application of SN was studied by Genton[2]. Furthermore, many authors now use skew normal distribution as the components of mixture models.

In this paper, we considered the Skew-Laplace-Normal (SLN) distribution[3], whose density function has given by

$$f_{SLN}(y;\xi,\omega^2,\alpha) = 2 f_L(y;\xi,\omega)\Phi(\alpha\frac{y-\xi}{\omega}), y \in R \qquad (1)$$

where $\xi \in R^1$ is the local parameter, $\alpha \in R^1$ is the skewness parameter and $\omega^2 \in (0,\infty)$ is the scale parameter. $f_L(y;\xi,\omega)$ shows that the density function of Laplace distribution with $f_L(y;\xi,\omega) = \frac{1}{2\omega}e^{-\frac{|y-\xi|}{\omega}}$. We write $Y \sim SLN(\xi,\omega^2,\alpha)$, if Y is SLN distribution. The random variable Y has scale mixture representation:

$$Y = \xi + \omega\frac{Z}{V}, \qquad (2)$$

where $Z \sim SN(0,1,\alpha)$, V has a distribution with the pdf $f_V(v) = v^{-3}e^{-\frac{1}{2v^2}}, v > 0$, Z and V are independent variables. Using the stochastic representation of Skew-Normal (SN) distributed random variable Z [1], the random variable Y have stochastic representation

Weisan Wu School of Mathematics and Statistics, Baicheng Normal University, Baicheng, China, Xinyu Yang* School of Mathematics and Statistics, Baicheng Normal University, Baicheng, China, *These authors contributed equally to this work as co- corresponding author, *Corresponding author: School of Mathematics and Statistics, Baicheng Normal University, Baicheng, China (E-mail: Weisan Wu, wuws009@outlook.com; Xinyu Yang, xy_yang0612@outlook.com)*

$$Y = \xi + \omega\left(\frac{\alpha |Z_1|}{\sqrt{V^2(V^2+\alpha^2)}} + \frac{Z_2}{\sqrt{V^2+\alpha^2}}\right), \qquad (3)$$

where $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ are independent random variables. Let $\gamma = \sqrt{V^{-2}(V^2+\alpha^2)}\,|Z_1|$. Then, the SLN distribution have hierarchical representation:

$$Y \mid \gamma, v \sim N\left(\xi + \frac{\omega\alpha\gamma}{v^2+\alpha^2}, \frac{\omega^2}{v^2+\alpha^2}\right), \qquad (4)$$

$$\gamma \mid v \sim TN\left(0, \frac{v^2+\alpha^2}{v^2}; (0,\infty)\right), \qquad (5)$$

where $TN(g)$ shows the truncated normal distribution. The joint density function of Y, γ, V is

$$f(y,\gamma,v) = \frac{1}{\pi\omega v^2} e^{-\frac{1}{2v^2}} e^{-\frac{1}{2}\left[\frac{v^2(y-\xi)^2}{\omega^2} + \left(\gamma - \frac{\alpha(y-\xi)}{\omega}\right)^2\right]}. \qquad (6)$$

Taking the integral of equation (6) over γ, we obtain

$$f(y,v) = \sqrt{\frac{2}{\pi}} \frac{1}{\omega v^2} e^{-\frac{1}{2v^2} - \frac{v^2x^2}{2}} \Phi^{-1}(\alpha x), \qquad (7)$$

where $x = \frac{y-\xi}{\omega}$. Further, we need to compute conditional expectation about $V^2 \mid Y = y$, $\gamma \mid Y = y$ and $\gamma^2 \mid Y = y$. Using some basic probability formula we can get the following results:

$$E(V^2 \mid y) = \frac{\omega}{|y-\xi|}, \qquad (8)$$

$$E(\gamma \mid y) = \alpha x + \frac{\Phi(\alpha x)}{\phi(\alpha x)}, \qquad (9)$$

$$E(\gamma^2 \mid y) = 1 + \alpha x E(\gamma \mid y). \qquad (10)$$

Given the density function (1) and a finite number p as the components, the density function of Skew-Laplace-Normal Mixture (SLNM) model is

$$f_{SLNM}(y;G) = \sum_{k=1}^{p} \pi_k f_{SLN}(y;\theta_k) = \int f_{SLN}(y;\theta)dG(\theta), \quad (11)$$

where $\pi_k, \theta_k = (\xi_k, \omega_k^2, \alpha_k), \theta = (\theta_1, \ldots, \theta_p) \in \Theta$, Θ is the parametric space are the mixing proportion and component parameters respectively.

For convenience, we use the notation G for all parameters in SLNM models, and also write its cumulative distribution function $G(\theta) = \sum_{k=1}^{p} \pi_k I(\theta_k \le \theta)$, $I(\cdot)$ is the indicator function.

The parameter space of G can be written as

$$\Gamma = \{G = (\pi_1, \ldots, \pi_p, \xi_1, \ldots, \xi_p, \omega_1^2, \ldots, \omega_p^2, \alpha_1, \ldots, \alpha_p) : 0 \le \pi_k \le 1, \sum_{k=1}^{p} \pi_k = 1, -\infty < \xi_k, \alpha_k < +\infty, \omega_k \ge 0,$$
$$k = 1, \ldots, p\}$$

Therefore, we can get G 's log-likelihood function

$$l_n(G) = \sum_{i=1}^{n} \log f_{SLNM}(y_i;G) = \sum_{i=1}^{n} \log \sum_{k=1}^{p} \frac{\pi_k}{\omega_k} e^{-\frac{|y_i-\xi_k|}{\omega_k}} \Phi\left(\frac{\alpha_k}{\omega_k}(y_i - \xi_k)\right), \quad (12)$$

So, when we fixed $\xi_1 = y_1$, for any parameter α, let $\omega_1 \to 0$, the first undesirable mathematical property, that is $l_n(G) \to +\infty$. The second undesirable property is shape parameter α on the bound of MLE. The divergence of skew parameter will lead to the classical statistical theories do not hold.

Although the probability of a divergent MLE goes rapidly to 0 as $n \to +\infty$, while for small or moderate sample sizes, it can be non-negligible. Meanwhile, the divergent estimator involves an enormous amount of computational workload and a bad inferential process.

Suppose observations $y_1, y_2, \ldots, y_n$ from the LSNM model, it's log-likelihood function is given by

$$l_n(G) = \sum_{i=1}^{n} \log f_{SLNM}(y_i; G) = \sum_{i=1}^{n} \log\{\sum_{k=1}^{p} \pi_k f_{SLN}(y;\theta_k)\}, \quad (13)$$

Given the sample $y_1, y_2, \ldots, y_n$, let the location parameter of first component be $\xi_1 = y_1$. Obviously, for any skew parameter α_1, when $\omega_1 \to 0$, we have

$$f_{SLNM}(y_1;\theta_1) = 2 f_L(y_1 - \xi_1)\Phi(\alpha_1\omega_1^{-1}(y_1 - \xi_1)) = \omega_1^{-1} \to \infty. \quad (14)$$

Let the other component parameters be $\xi_k = \xi_0, \omega_k^2 = 1, \alpha_k = \alpha_0$ for $k = 2, 3, \ldots, p$, where $\xi_0 \neq y_i$ for any $i = 2, 3, \ldots, n$. It is clear that the component density $f_{SLNM}(y_i;\theta_k)$ is bounded by

$$f_{SLNM}(y_i;\theta_k) = 2 f_L(y_i - \xi_k)\Phi(\alpha_k\omega_k^{-1}(y_i - \xi_k)) \le 2, \quad (15)$$

for $k = 2, 3, \ldots, p$, $i = 2, 3, \ldots, n$.

Let $\theta_0 = (\xi_0, 1, \alpha_0)$, and the mixing proportion be $\pi_1 = 0.5$, and $\pi_k = \dfrac{0.5}{p-1}$ for $k = 2, 3, \ldots, p$. we have

$$\begin{aligned} l_n(G) &= \log\{\sum_{k=1}^{p} \pi_k f_{SLNM}(y_1;\theta_k)\} + \sum_{i=2}^{n} \log\{\sum_{k=1}^{p} \pi_k f_{SLNM}(y_i;\theta_k)\} \\ &\ge \log \pi_1 f_{SLNM}(y_1;\theta_1) + \sum_{i=2}^{n} \log\{\sum_{k=2}^{p} \pi_k f_{SLNM}(y_i;\theta_0)\} \\ &= n\log 0.5 + \log f_{SLNM}(y_1;\theta_1) + \sum_{i=2}^{n} \log f_{SLNM}(y_i;\theta_0) \to \infty, \end{aligned}$$

as $\omega_1 \to 0$. Therefore, the MLE of G is inconsistent.

In this paper, to overcome both likelihood function's undesirable properties, we introduce a penalty function in SLNM models, it defined as

$$\begin{aligned} pl_n(G) &= l_n(G) + p_n(G), \\ p_n(G) &= p_n(\omega) + p_n(\alpha) = \sum_{k=1}^{p} \tilde{p}_n(\omega_k) + \sum_{k=1}^{p} \tilde{p}_n(\alpha_k). \end{aligned} \quad (16)$$

Then PMLE of G would be obtained by $\tilde{G} = \arg\max_G pl_n(G)$. With reasonable penalties, we need to choose proper penalized term over the filed Γ. For example, we need to prove finite loglikelihood function when ω goes to 0 or infinity, and α tends to infinity.

The rest of this article is organized as follows.

In Section 2, we give some basis theories that including basis lemmas and penalties.

In Section 3, we give some important theory of strong consistency of the proposed PMLE.

in Section 4, we give the penalized EM algorithms. Detail proofs of theorem are given in the Appendix.

PRELIMINARIES

Technical Lemmas

We need to give a bound on the number of samples included in a very small area of the location parameters, that is an important technique to strong consistency theory.

Without proofs, we give the following Lemma[4]

Lemma 1. Let $Y_1, \ldots, Y_n$ be a simple random sample from F with density function $f_{SLN}(y)$. Suppose $f_{SLN}(y)$ is a continuous function and $M = \sup_y f_{SLN}(y) < \infty$. Let

$F_n(y)=\frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y)$. Thus, as $n\to\infty$,

$$\sup_{y\in R^1}\{F_n(y+\varepsilon)-F_n(y)\}\le 2M\varepsilon+10n^{-1}\log n, \qquad (17)$$

holds uniformly for all $\varepsilon>0$ almost surely.

Note that the density functions of multivariate normal model and LSNM model satisfy the following inequality

$$f_{SLNM}(y;\theta)=2f_L(y-\xi;\omega)\Phi(\alpha\omega^{-1}(y-\xi))\le 2f_L(y-\xi;\omega), \quad (18)$$

Suppose that $y_1, y_2, \ldots, y_n$ be a random sample from LSNM model, by the above inequality and the strong law of large numbers, as $n\to\infty$, we have

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i\in G)-\frac{1}{n}\sum_{i=1}^{n}2I(x_i\in G)\xrightarrow{a.s.}\int_G f_{SLNM}(y;\theta)dy-\int_G 2f_L(x-\xi;\omega)dx\le 0$$

Then, we can obtain the upper bound of $\sum_{i=1}^{n} I(y_i\in G)$. Let $\varepsilon=|\omega\log\omega^2|$, where $\omega>0$. With a slight alteration, we list the result of SLNM as follows:

Lemma 2. Let observations $Y_1,\ldots,Y_n$ from a finite mixture of SLN distributions with density function $f(y;G_0)$, except for a zero-probability event not depending on ω, when $n\to\infty$, almost surly we have

$$\sup_{x\in R^1}\sum_{i=1}^{n} I(|Y_i-\xi|\le|\omega\log\omega^2|)\le 4Mn|\omega\log\omega^2|+10\log n$$

Choice of Penalties

Lemma 1 and Lemma 2 give some basis conditions of the penalties to ensure the consistency of the proposed PMLE:

Condition1. $\forall\omega>0, p_n(\omega)=o(n)$, $\sup_{\omega>0}\max\{0,p_n(\omega)\}=o(n)$,

Condition2. $p_n(\omega)<(\log n)^2\log\omega$, if $\omega<n^{-1}\log n$ and n is a large number,

Condition3. $p_n(\alpha)$ have maximum value at the point $\alpha=0$, and $p_n(\alpha)\to-\infty$ as $\alpha_k^2\to+\infty$. In addition, $p_n(\alpha)=0$ at $\alpha=0$,

Condition4. $p_n(\omega)$ and $p_n(\alpha)$ are differential functions to ω and α respectively, and as $n\to+\infty$, $p_n'(\omega)=o(n^{\frac{1}{2}})$ and $p_n'(\alpha)=o(n^{\frac{1}{2}})$.

many of functions can be as a penalty function which easily to satisfied condition 1-4. Condition 1 gives the upper-bound and lower-bound of $p_n(\omega)$, while condition 2 makes $p_n(\omega)$ sufficiently severe to prevent $\omega^2\to+\infty$,condition 3 limits the $p_n(\alpha)$ avoid the skewness parameter not divergence, condition 4 shows that limiting distribution of the penalized MLE must be exist. Here, we use sample variance s_n^2 to construct penalized term $p_n(\omega)$, like most papers we recommend the penalty functions have next form

$$p_n(\omega_k)=-a_n\left(\frac{s_n^2}{\omega^2}+\log\frac{\omega^2}{s_n^2}\right), \qquad (19)$$

$$p_n(\alpha_k)=-b_n(\alpha_k^2-\log(1+\alpha_k^2)), \qquad (20)$$

where a_n and b_n are positive numbers, it can be taken different number to tunning penalized terms $p_n(\omega_k)$ and $p_n(\alpha_k)$.

CONSISTENCY OF THE PENALIZED MLE

Consistency of the Penalized MLE When $p=p_0$

We denote conditional expectation under the true mixing distribution G_0 as $K_0=E_{G_0}\{\log f(Y;G_0)\}$. Suppose that ε_0 and $\bar{\alpha}$ are positive constants. Given p, M and K_0, there exists $\varepsilon_0\to 0$ satisfying following two inequalities:

$$4pM\varepsilon_0\log^2\varepsilon_0\le 1, -\log\varepsilon_0\le p(K_0-2).$$

Then, we also pick a parameter $\bar{\alpha}$ such that $\bar{\alpha}>\max_k\{|\alpha_{0k}|\}, k=1,\ldots,p$,where α_{0k} is the k-th element of α_0, and it is the element of G_0. Note that we need to select ε_0 and $\bar{\alpha}$ are depend on G_0.

We defined the follow regions based on the ε_0 and $\bar{\alpha}$

$$\Gamma_{\omega} = \{G \in \Gamma : \min\{\omega_k\} \le \varepsilon_0, k = 1, \ldots, p\},$$
$$\Gamma_{\alpha} = \{G \in \Gamma : \max\{|\alpha_k| \ge \bar{\alpha}, k = 1, \ldots, p\},$$
$$\Gamma^* = \Gamma - \Gamma_{\omega} \cup \Gamma_{\alpha}.$$

The parameter of the SLNM distribution has at least one component will be close to zero, if it falls in the region Γ_{ω}. The penalty $p_n(\omega)$ will counter it such that PMLE with is with a diminishing probability. Similarly, PMLE will not fall in the region Γ_{α} because of at least one $|\alpha_k|$ diverges to infinity in this region.

Firstly, we will give the theory of the consistency on SLNM parameters. Without loss of generality, we need to sort the component deviations by ascending order, that is $\omega_{(1)} \le \ldots \le \omega_{(p)}$, and the corresponding mixing proportion and other parameters, like $\pi_{(k)}$ and $\theta_{(k)} = (\xi_{(k)}, \omega^2_{(k)}, \alpha_{(k)})$, $k \in \{1, \ldots, p\}$, Then, for index $\tau \in \{1, \ldots, p\}$, we divided the parameter space Γ_{ω} to several parts as follows

$$\Gamma^{\tau}_{\omega} = \{G \in \Gamma_{\omega} : \omega_{(1)} \le \ldots \le \omega_{(\tau)} < \tau_0 < \omega_{(\tau+1)} \le \ldots \le \omega_{(p)}\}.$$

In particular, when $\tau = p$,

$$\Gamma^{p}_{\omega} = \{G \in \Gamma_{\omega} : \omega_{(1)} \le \ldots \le \omega_{(p)} < \tau_0 < \varepsilon_0\}.$$

Theorem 1 Let $f(y; G_0)$ is the density function for true distribution. Take the penalized likelihood $pl_n(G)$ has form like equations (19) and (20), and it's satisfied condition 1-4. Then for any $G \in \Gamma^{p}_{\omega}$, as $n \to +\infty$, we almost surely have

$$\sup_{\Gamma^{p}_{\omega}} pl_n(G) - pl_n(G_0) \to -\infty, \qquad (21)$$

To save space , the detail of Theorem's proof was list in Appendix. For the spaces Γ^{τ}_{ω} with $1 \le \tau \le p-1$, we can obtain the similar results as follows like Theorem 1.

Theorem 2
At the same assume as the Theorem 1, for $G \in \Gamma^{\tau}_{\omega}$, $1 \le \tau \le p-1$, as $n \to +\infty$ and almost surely

$$\sup_{\Gamma^{\tau}_{\omega}} pl_n(G) - pl_n(G_0) \to -\infty, \qquad (22)$$

Note that $\Gamma_{\omega} = \cup^{p}_{\tau=1} \Gamma^{\tau}_{\omega}$. From Theorem 1 and Theorem 2, we found that PMLE of G is exclude in Γ_{ω} except for a zero probability event. Next, we consider the region $\Gamma^{C}_{\omega} \cup \Gamma_{\alpha}$.

Theorem 3 At the same assume as the Theorem 1, for $G \in \Gamma^{\tau}_{\omega}$, $1 \le \tau \le p-1$, as $n \to +\infty$ and almost surely

$$\sup_{\Gamma^{C}_{\omega} \cup \Gamma_{\alpha}} pl_n(G) - pl_n(G_0) \to -\infty, \qquad (23)$$

From these theorems, we know the penalized MLE of $\hat{G}$ not falls in Γ_{ω} or Γ_{α}. Therefore, it suffices to show that $\hat{G}$ will be falls in complementary of $\Gamma_{\omega} \cup \Gamma_{\alpha}$, that is the set Γ^*. The conclusion of consistency of $\hat{G}$ was present as theorem 4.

Theorem 4 Under the same assumption as in Theorem 1, $\hat{G}$ is a mixing distribution which satisfied

$$pl_n(\hat{G}) - pl_n(G_0) \ge c > -\infty,$$

Then, as $n \to \infty$,, we almost surely have $G \to G_0$.

Rewrite $\Gamma^* = \{G \in \Gamma : \min_k\{\omega_k\} \ge \varepsilon_0, \max_k\{|\alpha_k|\} \ge \bar{\alpha}\}$, $\hat{G} \in \Gamma^*$, it means that there is a positive bound to component scale parameter and the absolute value of skewness parameters. Because of the set Γ^* is regular, the consistency is then covered by the technique[5].

We know the PMLE $\hat{G}$ is strongly consistent because of $pl_n(\hat{G}) - pl_n(G_0) \geq 0$. Otherwise, for $p = p_0$, the limitation points of all elements in $\hat{G}$ are in the G_0 almost surely. Nextly, let $S_n(G) = \frac{\partial l_n(G)}{\partial G}$ and $S_n'(G) = \frac{\partial^2 l_n(G)}{\partial G \partial G^T}$ be the score vector and second derivative matrix of $l_n(G)$. Because the SLNM model had regularity at the point G_0, the Fisher information matrix $I(G_0) = -E\{S_n'(G_0)\} = E\{S_n^T(G_0) S_n(G_0)\}$ is a positive definitely matrix. Using the classical asymptotic theory and condition 4, we have

$$\hat{G} - G_0 = -\{S_n'(G_0)\}^{-1} S_n(G_0) + o_p(n^{-\frac{1}{2}}).$$

So, the penalized estimator of SLNM model have the asymptotic normality and efficiency.

Theorem 5 Under the same conditions as in Theorem 1 and Condition 4, as $n \to \infty$, $\sqrt{n}(\hat{G} - G_0) \to N(0, I^{-1}(G_0))$ in distribution.

Consistency of the Penalized MLE When $p > p_0$

In most instances, people often want to know a finite upper bound of the mixture order p rather than the exact p_0 [6]. In our cases, we use some distance to measure the difference between the mixing distribution G and G_0,

$$D(G, G_0) = \int_{\Theta} | G(\theta), G_0(\theta) | e^{-|\theta|} d\theta, \quad (24)$$

where $\theta = (\xi, \omega, \alpha) \in \Theta$, $|\theta| = |\xi| + |\omega| + |\alpha|$ and $d\theta = d\xi d\omega d\alpha$. The distance is bounded with the inequalities $0 < D(G, G_0) < \int_{\Theta} e^{-|\theta|} d\theta$ and can implied $\hat{G} \to G_0$ in distribution from $D(\hat{G}, G_0) \to 0$. Use this distance we can get the following theorem.

Theorem 6 Assume the same conditions as in Theorem 1, let $p > p_0$, for any mixing distribution G with p components satisfying

$$pl_n(G) - pl_n(G_0) \geq c > -\infty,$$

Then as $n \to \infty$, almost surely we have $G \to G_0$.

THE PENALIZED EM ALGORITHMS

Consider the complete data $(Y, Z) = \{Y_i, Z_i\}_{i=1}^n$, where $Z_i = (Z_{1i}, \cdots, Z_{pi})$ is the latent component-indicators vector, it follows a multinomial distribution with discrete probabilities $\pi_1, \cdots, \pi_p$. For convenient, we write it as $Z_i \sim M(1; \pi_1, \cdots, \pi_p)$, and $Z_1, \cdots, Z_n$ are mutually independent. Based on the component-indicators, for each $Y_i, i = 1, \cdots, n$, we have a hierarchical representation for SLNM as follows

$$Y_i | \gamma_i, v_i, Z_{ki} = 1 \sim N(\xi_k + \frac{\omega_k \alpha_k \gamma_i}{v_i^2 + \alpha_k^2}, \frac{\omega_k^2}{v_i^2 + \alpha_k^2}),$$

$$\gamma_i | v_i, Z_{ki} = 1 \sim TN(0, \frac{v_i^2 + \alpha_k^2}{v_i^2}; (0, \infty)),$$

$$v_i | Z_{ki} = 1 \sim f(v_i) = v_i^{-3} e^{-\frac{1}{2v_i^2}},$$

$$Z_i \sim M(1; \pi_1, \cdots, \pi_p).$$

Let $y = (y_1, \cdots, y_n), \gamma = (\gamma_1, \cdots, \gamma_n), v = (v_1, \cdots, v_n), z = (z_1, \cdots, z_n)$, and let (y, γ, v, z) be the complete data. From the hierarchical representation given in above equation, the complete data log-likelihood function of G is as follows:

$$\mathrm{l}(G;y,\gamma,v,z)=\sum_{i=1}^{n}\sum_{k=1}^{p}z_{ki}\{\log\pi_k-\frac{1}{2}\log\omega_k^2-2\log v_i-\frac{1}{2v_i^2}$$
$$-\frac{1}{2}(v_i^2\frac{(y_i-\xi_k)^2}{\omega_k^2}+\gamma_i^2-2\eta_k(y_i-\xi_k)\gamma_i+\eta_k^2(y_i-\xi_k)^2)\}$$

where $\eta_k=\dfrac{\alpha_k}{\omega_k}$ is the re-parameterized parameter. This function can be maximized to obtain the estimator of G, and the expectation of the complete data log-likelihood function given the observed data y_i as follows:

$$E(\mathrm{l}(\Theta;y,\gamma,v,Z\mid y_i)=\sum_{i=1}^{n}\sum_{k=1}^{p}E(Z_{ki}\mid y_i)\{\log\pi_k-\frac{1}{2}\log\omega_k^2-2E(\log v_i\mid y_i)$$
$$-E(\frac{1}{2v_i^2}\mid y_i)-\frac{1}{2}(E(v_i^2\mid y_i)\frac{(y_i-\xi_k)^2}{\omega_k^2})$$
$$-\frac{1}{2}(E(\gamma_i^2\mid y_i)-2\eta_k(y_i-\xi_k)E(\gamma_i\mid y_i)+\eta_k^2(y_i-\xi_k)^2)\}$$

The ECM algorithm proceeds as follows[7,8,9]

E-step: Compute the conditional expectations

$$\hat{z}_{ki}^{(t)}=E(Z_{ki}\mid y_i,\hat{G}^{(t)})=\frac{\hat{\pi}_k^{(t)}f_k(y_i,\hat{G}^{(t)})}{f(y_i,\hat{G}^{(t)})},$$

$$\hat{v}_{ki}^{(t)}=E(V_i^2\mid y_i,\hat{G}^{(t)})=\frac{\hat{\omega}_k^{(t)}}{|y_i-\hat{\xi}_k^{(t)}|},$$

$$\hat{\gamma}_{1ki}^{(t)}=E(\gamma_i\mid y_i,\hat{G}^{(t)})=\hat{\alpha}_k^{(t)}\hat{x}_{ki}^{(t)}+\frac{\phi(\hat{\alpha}_k^{(t)}\hat{x}_{ki}^{(t)})}{\Phi(\hat{\alpha}_k^{(t)}\hat{x}_{ki}^{(t)})},$$

$$\hat{\gamma}_{2ki}^{(t)}=E(\gamma_i^2\mid y_i,\hat{G}^{(t)})=1+\hat{\alpha}_k^{(t)}\hat{x}_{ki}^{(t)}\hat{\gamma}_{1ki}^{(t)},$$

where $\hat{x}_{ki}^{(t)}=(y_i-\hat{\xi}_k^{(t)})/\hat{\omega}_k^{(t)}$. Then, we can get the objective function $Q(G;\hat{G}^{(t)})$ as follows:

$$Q(G;\hat{G}^{(t)})=E(\mathrm{l}_C(G))+p_n(G)$$
$$=\sum_{i=1}^{n}\sum_{k=1}^{p}\hat{z}_{ki}^{(t)}\{\log\pi_k-\frac{1}{2}\log\omega_k^2-\hat{v}_{ki}^{(t)}\frac{(y_i-\xi_k)^2}{\omega_k^2}-\frac{1}{2}(\hat{\gamma}_{2ki}^{(t)}$$
$$-2\eta_k(y_i-\xi_k)\hat{\gamma}_{1ki}^{(t)}+\eta_k^2(y_i-\xi_k)^2)\}+\sum_{k=1}^{p}p_n(\omega_k)+\sum_{k=1}^{p}p_n(\eta_k)$$

where, we replaced $p_n(\alpha)$ by $p_n(\eta)=-b_n(\omega^2\eta^2-\log(1+\omega^2\eta^2))$.

CM-Step: Maximize $Q(G;\hat{G}^{(t)})$ with respect to G under the restriction with $\sum_{k=1}^{p}\pi_k=1$.

1. Update $\pi_k^{(t)}$ by $\pi_k^{(t+1)}=\dfrac{1}{n}\sum_{i=1}^{n}\hat{z}_{ki}^{(t)}$,

2. Update $\xi_k^{(t)}$ by.

$$\xi_k^{(t+1)}=\frac{2\sum_{i=1}^{n}\hat{z}_{ki}^{(t)}\hat{v}_{ki}^{(t)}y_i+\hat{\alpha}_k^{2(t)}\sum_{i=1}^{n}\hat{z}_{ki}^{(t)}y_i-\hat{\omega}_k^{(t)}\hat{\alpha}_k^{(t)}\sum_{i=1}^{n}\hat{z}_{ki}^{(t)}\hat{\gamma}_{1ki}^{(t)}}{\sum_{i=1}^{n}\hat{z}_{ki}^{(t)}(2\hat{v}_{ki}^{(t)}+\hat{\alpha}_k^{2(t)})},$$

3. Fix $\xi_k=\xi_k^{(t+1)}$, with the definition of $p_n(\omega)$, obtain $\hat{\omega}_k^{2(t+1)}$ by setting

$$\hat{\omega}_k^{2(t+1)} = \frac{2\sum_{i=1}^{n} \hat{z}_{ki}^{(t)} \hat{v}_{ki}^{(t)} (y_i - \xi_k^{(t+1)})^2 + 2s_n^2}{\sum_{i=1}^{n} \hat{z}_{ki}^{(t)} + 2s_n^2},$$

4. Fix $\xi_k = \xi_k^{(t+1)}$ and $\hat{\omega}_k^{\,2} = \hat{\omega}_k^{\,2(t+1)}$, with equivalent transformation of

$$\sum_{k=1}^{p} p_n(\eta_k) = \sum_{k=1}^{p} [2\omega_k^2 \eta_k^2 - \log(1 + \omega_k^2 \eta_k^2)]$$ and $\eta_k^{(t+1)}$ is solution of

$$\eta_k^{(t+1)} \sum_{i=1}^{n} \hat{z}_{ki}^{(t)} (y_i - \xi_k^{(t+1)})^2 - \sum_{i=1}^{n} \hat{z}_{ki}^{(t)} (y_i - \xi_k^{(t+1)}) \hat{\gamma}_{1ki}^{(t)} + \frac{2\hat{\omega}_k^{4(t+1)} \hat{\eta}_k^{3(t)}}{1 + \hat{\omega}_k^{2(t+1)} \hat{\eta}_k^{2(t)}} = 0.$$

$\hat{\alpha}_k^{(t+1)}$ can be obtained as follows:

$$\hat{\alpha}_k^{(t+1)} = \hat{\omega}_k^{(t+1)} \hat{\eta}_k^{(t+1)}.$$

DISCUSSION

In this paper, we consider a penalty term SLNM model parameter estimation problem, in view of the likelihood function unbounded model and skewness parameter divergence, we add penalized term in the log-likelihood function respectively to the scale parameter and skew parameter. It makes the scale parameter to converge to 0 and the log-likelihood function when the skewness parameter divergent bounded. For parameter estimation, we present an improved EM algorithm, which can effectively estimate local parameter, scale parameter, skewness parameter and mixing proportion.

The approach developed herein is arguably valid for regaining the consistency and efficiency, and has the advantage of placing no additional constraints on the parameter space. This methodology could be extensively to other classes of finite skew mixture models, such as the finite mixture of Skew-Normal-t distributions and the mixture of Skew-Normal-Cauchy models.

APPENDIX

Proof of Theorem 1:

Define sets $A_{(k)} = \{i : |y_i - \xi_{(k)}| < |\omega_{(k)} \log \omega_{(k)}^2|\}, k = 1, \ldots, p$. For any set S, denote $n(S)$ be the number of elements in S. Furthermore, we defined $\ell_n(G; S) = \sum_{i \in S} f(y_i; \Psi)$. For $G \in \Gamma_\omega^p$ and some small enough ε_0, the mixture density $f_{SLNM}(y_i; G) \le \frac{1}{\omega_{(k)}}$ for any $i \in A_{(k)}$. Since $n(\cap_{t=1}^{k-1} A_{(k)}^C \cap A_{(k)}) \le n(A_{(k)})$, use the bound of $n(A_{(k)})$ in lemma 2, we almost surely have

$$\begin{aligned} \ell_n(G; \cap_{t=1}^{k-1} A_{(k)}^C \cap A_{(k)}) &\le \sum_{i \in A_{(k)}} \log f_{SLNM}(y_i; G) \\ &\le n(A_{(k)}) \log \omega_{(k)} \\ &\le 4Mn\omega_{(k)} \log^2 \omega_{(k)} - 10 \log \omega_{(k)} \log n \end{aligned}$$

Next, we adding penalty function $p_n(\omega_{(k)})$ to the likelihood function, it satisfying conditions 1-2, above formula can be written as

$$\ell_n(G; \cap_{t=1}^{k-1} A_{(k)}^C \cap A_{(k)}) + p_n(\omega_{(k)})$$

$$\le 4Mn\omega_{(k)}\log^2\omega_{(k)} - (10\log n - \log^2 n)\log\omega_{(k)}$$
$$\le 4Mn\omega_{(k)}\log^2\omega_{(k)}$$
$$\le 4Mn\varepsilon_0\log^2\varepsilon_0$$

For any $i \in \bigcap_{t=1}^{p} A_{(t)}^C$, since $|y_i - \xi_{(k)}| > |\omega_{(k)}\log\omega_{(k)}^2|$, it is easy to show

$$f_{SLNM}(y_i;G) \le \log\{\sum_{k=1}^{p}\frac{\pi_{(k)}}{\omega_{(k)}}e^{-\frac{|y_i-\xi_{(k)}|}{\omega_{(k)}}}\}$$
$$\le \log\frac{1}{\omega_{(k)}}e^{\log\omega_{(k)}^2} = \log\omega_{(k)} < \log\varepsilon_0 < 0$$

By $4pM\varepsilon_0\log^2\varepsilon_0 \le 1, -4pM\varepsilon_0\log\varepsilon_0 \le \frac{p-1}{p}$ holds for small enough ε_0, we can implied that

$$n(\bigcap_{t=1}^{k-1}A_{(t)}^C) \ge n - \sum_{t=1}^{p}n(A_{(t)}) \ge \frac{n}{p},$$

Hence, the summation log-likelihood contributions of samples in $n(\bigcap_{t=1}^{k-1}A_{(t)}^C)$ are bounded by

$$l_n(G;\bigcap_{t=1}^{k-1}A_{(t)}^C) = n(\bigcap_{t=1}^{k-1}A_{(t)}^C)\log f_{SLNM}(y_i;G) \le -\frac{n}{p}\log\varepsilon_0,$$

Thus, for $G \in \Gamma_\omega^p$ and some sufficiently small ε_0, use above results and condition 3, the penalized log-likelihood function has the upper bound as

$$pl_n(G) = \sum_{t=1}^{p}\{l_n(G;\bigcap_{t=1}^{k-1}A_{(t)}^C \cap A_{(k)}) + p_n(\omega_{(k)})\} + l_n(G;\bigcap_{t=1}^{p}A_{(t)}^C) + o(n)$$
$$\le 4pMn\varepsilon_0\log^2\varepsilon_0 - \frac{n}{p}\log\varepsilon_0 + o(n)$$
$$\le n + n(K_0 - 2) + o(n)$$
$$\le n(K_0 - 1) + o(n)$$

By the strong law of large numbers, we have almost surely $\frac{1}{n}pl_n(G_0) \to K_0$. Consequently, as $n \to \infty$, almost surely we have,

$$\sup_{\Gamma_\omega^p} pl_n(G) - pl_n(G_0) \le -n + o(n) \to -\infty .$$

Proof of Theorem 2

Let $\bar{\Gamma}_\omega^\tau$ be a compaction Γ_ω^τ allowing $\omega_{(1)} = \cdots = \omega_{(\tau)} = 0$. For $G \in \bar{\Gamma}_\omega^\tau$, define the following continuous functions

$$g(y;G) = \sum_{k=1}^{\tau}\pi_{(k)}e^{-\frac{|y-\xi_{(k)}|}{2\varepsilon_0}} + \sum_{k=\tau+1}^{p}\pi_{(k)}f_{SLNM}(y;\theta_{(k)}),$$

where $f_{SLNM}(y;\theta_{(k)})$ is the density function of k-th component. Because $\omega_{(p)} \ge \cdots \ge \omega_{(\tau+1)} \ge \varepsilon_0$, $g_\tau(y;G)$ is bounded over $\bar{\Gamma}_\omega^\tau$. Therefore, $G \in \bar{\Gamma}_\omega^\tau$, when $\varepsilon_0 \le 1$, we have

$$\log E_{G_0}\{g_\tau(Y;G) / f_{SLNM}(Y;G_0)\}$$

$$
\begin{aligned}
&= \log \int [\sum_{k=1}^{\tau} \pi_k e^{-\frac{|y-\xi_0|}{2\varepsilon_0}} + \sum_{k=\tau+1}^{p} \pi_k f_{SLNM}(y;\theta_k)]dy \\
&= \log[\int \sum_{k=1}^{\tau} \pi_k [e^{-\frac{|y-\xi_0|}{2\varepsilon_0}} + f_{SLNM}(y;\theta_k)]dy + \int \sum_{k=\tau+1}^{p} \pi_k f_{SLNM}(y;\theta_k)dy] \\
&= \log[\int \sum_{k=1}^{\tau} \pi_k 2\varepsilon_0 [\frac{1}{2\varepsilon_0} e^{-\frac{|y-\xi_0|}{2\varepsilon_0}} - \frac{1}{2\varepsilon_0} f_{SLNM}(y;\theta_k)]dy + 1] \\
&\le \log[\sum_{k=1}^{\tau} \pi_k (2\varepsilon_0 - 1) + 1] \\
&= -\Delta_\tau(\varepsilon_0) < 0
\end{aligned}
$$

It is also clearly that $\Delta_\tau(\varepsilon_0)$ is a non-increasing function and $\lim_{\varepsilon_0 \to 0} \Delta_\tau(\varepsilon_0) \in (0,\infty)$. Hence, the inequality $8\tau M \varepsilon_0 \log^2 \varepsilon_0 < \Delta_\tau(\varepsilon_0)$ holds for small enough ε_0.

Define $l_n^\tau(G) = \sum_{i=1}^{n} \log g_\tau(y_i;G)$ on $\bar{\Gamma}_\omega^\tau$, by the strong law of large numbers and the upper bound of Jensen's inequality, we have almost surely

$$
\sup_{G \in \bar{\Gamma}_\omega^\tau} \frac{1}{n}\{l_n^\tau(G) - l_n(G_0)\} \to E_{G_0} \log\{g_\tau(Y;G)/f_{SLNM}(Y;G_0)\} = -\Delta_\tau(\varepsilon_0).
$$

For $G \in \bar{\Gamma}_\omega^\tau$ and $\tau \in \{1, \ldots, p-1\}$, we have $f_{SLNM}(y_i;G) \le \frac{1}{\omega_{(k)}} g_\tau(y_i;G)$ for all $i \in A_{(k)}$. For the rest observations, since $|y_i - \xi_{(k)}| \ge \omega_{(k)} \log \omega_{(k)}^2$, and if $\omega_{(k)}$ is small enough that $\omega_{(k)}^{-1} = e^{-\log \omega_{(k)}} < e^{\frac{1}{2}\log \omega_{(k)}^2}$, thus

$$
\begin{aligned}
f_{SLNM}(y;\theta_{(k)}) &\le \frac{1}{\omega_{(k)}} e^{-\frac{|y-\xi_{(k)}|}{\omega_{(k)}}} < e^{\frac{1}{2}\log \omega_{(k)}^2} e^{-\frac{|y-\xi_{(k)}|}{\omega_{(k)}}} \\
&= e^{\frac{|y-\xi_{(k)}|}{2\omega_{(k)}} - \frac{|y-\xi_{(k)}|}{\omega_{(k)}}} = e^{\frac{|y-\xi_{(k)}|}{2\omega_{(k)}}} \le e^{-\frac{|y-\xi_{(k)}|}{\omega_{(k)}}},
\end{aligned}
$$

holds with $\omega_{(k)} \le \varepsilon_0$, which implies $f_{SLNM}(y_i;G) \le g_\tau(y_i;G)$.

So, the log-likelihood contribution of y_i have upper bounds as

$$
\log f_{SLNM}(y_i;G) \le \begin{cases} -\log \omega_{(k)} + \log g_\tau(y_i;G), i \in A_{(k)}, \\ \log g_\tau(y_i;G), otherwise. \end{cases}
$$

So, in the log-likelihood function, we have

$$
\begin{aligned}
l_n(G) &= \sum_{i \in A_{(k)}} \log f_{SLNM}(y_i;G) \\
&\le \sum_{i \in A_{(k)}} \log g_\tau(y_i;G) - \sum_{i \in A_{(k)}} \log \omega_{(k)} \\
&\le l_n^\tau(G) - \sum_{k=1}^{\tau} n(A_{(k)}) \log \omega_{(k)}
\end{aligned}
$$

With these results , we can show that

$$
\begin{aligned}
&\sup_{\Gamma_\omega^\tau} l_n^\tau(G) - l_n(G_0) \\
&\le \sup_{\Gamma_\omega^\tau}\{l_n^\tau(G) - l_n(G_0)\} + \sup_{\Gamma_\omega^\tau} \sum_{k=1}^{\tau}\{-n(A_{(k)}) \log \omega_{(k)} + p_n(\omega_{(k)}) + o(n)\} \\
&\le -n\Delta_\tau(\varepsilon_0) + 4\tau M n \varepsilon_0 \log^2 \varepsilon_0 + o(n) \\
&\le -n\Delta_\tau(\varepsilon_0)/2 + o(n)
\end{aligned}
$$

for the chosen ε_0. Note that $\Delta_\tau(\varepsilon_0) > 0$, thus $\tau \in \{1, \ldots, p-1\}$ almost sure have $\sup_{\Gamma_\omega^\tau}\{\ell_n^\tau(G) - \ell_n(G_0)\} \to \infty$ as $n \to \infty$.

Proof of Theorem 3

When $G \in \Gamma_\omega^C \cap \Gamma_\alpha$, since the component deviances have a positive lower bound and divergent skew parameters do not lead to density function goes to infinite, then $f_{SLNM}(y;G)$ is bounded over $\Gamma_\omega^C \cap \Gamma_\alpha$.

Through the Jensen's inequality, we have $E_{G_0}\log\{f_{SLNM}(Y;G)/f_{SLNM}(Y;G_0)\} < 0$ for any $G \in \Gamma_\omega^C \cap \Gamma_\alpha$. We have also selected ρ large enough so that $G \notin \Gamma_\omega^C \cap \Gamma_\alpha$. Furthermore, it is easy to show that[10], when $n \to \infty$,

$$\sup_{\Gamma_\omega^C \cap \Gamma_\alpha} \sum_{i=1}^{n} \log\{f_{SLNM}(Y;G)/f_{SLNM}(Y;G_0)\} \to -\Delta(\rho) < 0,$$

Note that $\Delta(\rho)$ is greater than zero and is an increasing function of ρ. With the upper bound in $-\Delta(\rho)$ and the conditions C1-C3, we get

$$\begin{aligned} &\sup_{\Gamma_\omega^C \cap \Gamma_\alpha} \ell_n^\tau(G) - \ell_n(G_0) \\ &= \sup_{\Gamma_\omega^C \cap \Gamma_\alpha} \sum_{i=1}^{n} \log\{f_{SLNM}(Y;G)/f_{SLNM}(Y;G_0)\} + \sup_{\Gamma_\omega^C \cap \Gamma_\alpha} p_n(G) - p_n(G_0) \\ &\le -\frac{\Delta(\eta_0)}{2} n + o(n) \end{aligned}$$

Thus, we have $\sup_{\Gamma_\omega^C \cap \Gamma_\alpha} \ell_n^\tau(G) - \ell_n(G_0) \to -\infty$ almost surely as $n \to \infty$

Proof of Theorem 5

With the defined distance and for any $\kappa > 0$, we need to defined a new parameter space $\Omega_{(\kappa)} = \{G : G \in \Gamma, D(G, G_0) \ge \kappa\}$. Clearly, $G_0 \notin \Omega_{(\kappa)}$ when $\kappa > 0$.

For $G \in \Gamma_\omega^p \cap \Omega_{(\kappa)}$, it is easy to show the derivations of Theorem 1 are still applicable by replacing $G \in \Gamma_\omega^p$ with $G \in \Gamma_\omega^p \cap \Omega_{(\kappa)}$. Hence, we can quickly get $\sup_{G \in \Gamma_\omega^p \cap \Omega_{(\kappa)}} p\ell_n(G) - p\ell_n(G_0) \to -\infty$ as $n \to \infty$, and claim that $\tilde{G} \notin \Gamma_\omega^p \cap \Omega_{(\kappa)}$ with probability one.

Since $G_0 \notin \Omega_{(\kappa)}$, for $G \in \Gamma_\omega^\tau \cap \Omega_{(\kappa)}$ where $1 \le \tau \le p-1$ and $G \in \Gamma_\omega^p \cap \Gamma_\alpha \cap \Omega_{(\kappa)}$, the corresponding inequalities $E_{G_0}\log\{g_\tau(Y;G)/f_{SLNM}(Y;G_0)\} < 0$ and $E_{G_0}\log\{f_{SLNM}(Y;G)/f_{SLNM}(Y;G_0)\} < 0$ still holds respectively. Thus, we can get

$$\sup_{\Gamma_\omega^\tau \cap \Omega_{(\kappa)}} \frac{1}{n}\{\ell_n^\tau(G) - \ell_n(G_0)\} \to -\Delta_\tau(\varepsilon_0) < 0,$$

$$\sup_{\Gamma_\omega^\tau \cap \Gamma_\alpha \cap \Omega_{(\kappa)}} \sum_{i=1}^{n} \log\{f_{SLNM}(Y;G)/f_{SLNM}(Y;G_0)\} \to -\Delta(\rho) < 0.$$

for the properly selected ε_0, ρ and well defined $g_\tau(y;G)$. Using these results, with $n \to \infty$, we similarly get $\sup_{\Gamma_\omega^\tau \cap \Omega_{(\kappa)}} p\ell_\tau(G) - p\ell_n(G_0) \to -\infty$ for $1 \le \tau \le p-1$ and

$$\sup_{\Gamma_\omega^\tau \cap \Gamma_\alpha \cap \Omega_{(\kappa)}} \{p\ell_\tau(G) - p\ell_n(G_0)\} \to -\infty.$$

From above conclusions, we know that the PMLE $\tilde{G}$ must fall in set $\Gamma^* \cup \Omega_{(\kappa)}^C$ with probability 1. For any κ, we must be get $D(\tilde{G}, G_0) \to 0$ through $\tilde{G} \in \Omega_{(\kappa)}^C$. Meanwhile,

$\tilde{G} \in \Gamma^*$ also implies $D(\tilde{G}, G_0) \to 0$. Thus, the strong consistency of the penalized MLE is proved under the case $p > p_0$.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, author- ship, and/or publication of this article.

REFERENCES

1. Azzalini A. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 1985; 12: 171-178.
2. Genton MG. Skew-Elliptical Distributions and Their Applications. New York: Chapman and Hall/CRC 2004.
3. Dogru FZ., Arslan O. Parameter estimation for mixtures of skew laplace normal distributions and application in mixture regression modeling, Communications in Statistics Theory and Methods 2017; 0:1-18.
4. Jin LB., Xu WL., Zhu LX. Penalized maximum likelihood estimator for skew normal mixtures arXiv 2016;81-100.
5. Kiefer J., Wolfowitz J. Consistency on the maximum likelihood estimator in the presence of infinitely many incidental parameter, Annal s of Mathematical Statistics 1956; 27:887-906.
6. Chen JH., Tan XM., Zhang RC. Inference for normal mixtures in mean and variance, Statistica Sinica 2008; 18 :443-465.
7. Dempster AP., Laird NM., Rubin D B. Maximum likelihood from incomplete data via the em algorithm (with discussion), Journal of the Royal Statistical Society 1977; 39:1-38.
8. Liu CH., Rubin DB. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence, Biometrika 1994; 81:633-648.
9. Meng XL., Rubin DB. Maximum likelihood estimation via the ecm algorithm: A general framework, Biometrika 1993; 80:267-278.
10. Wald A. Note on the consistency of the maximum likelihood estimate, Annal s of Mathematical Statistics 1949; 20:595-601.