Statistics and Probability Letters 145 (2019) 12–20

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

On the consistency of penalized MLEs for Erlang mixtures

Cuihong Yin [a], X. Sheldon Lin [b,*], Rongtan Huang [c], Haili Yuan [d]

[a] *School of Insurance, Southwestern University of Finance and Economics, Chengdu, China*
[b] *Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada M5S 3G3*
[c] *School of Mathematical Sciences, Xiamen University, Xiamen, China*
[d] *School of Mathematics and Statistics, Wuhan University, Wuhan, China*

ARTICLE INFO

Article history:
Received 13 July 2017
Received in revised form 29 July 2018
Accepted 5 August 2018
Available online 16 August 2018

Keywords:
Erlang mixture
Insurance loss modeling
Penalized maximum likelihood estimates
Consistency

ABSTRACT

In Yin and Lin (2016), a new penalty, termed as iSCAD penalty, is proposed to obtain the maximum likelihood estimates (MLEs) of the weights and the common scale parameter of an Erlang mixture model. In that paper, it is shown through simulation studies and a real data application that the penalty provides an efficient way to determine the MLEs and the order of the mixture. In this paper, we provide a theoretical justification and show that the penalized maximum likelihood estimators of the weights and the scale parameter as well as the order of mixture are all consistent.

1. Introduction

The Erlang mixture model under consideration in this paper is of the following density:

$$h(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)=\sum_{j=1}^{m}\alpha_j g(x;\gamma_j,\theta),\ x>0, \tag{1.1}$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)$ is the mixing distribution and m is the number of components or the order of the mixture. Further, each component density is an Erlang of the form:

$$g(x;\gamma_j,\theta)=\frac{x^{\gamma_j-1}e^{-x/\theta}}{\theta^{\gamma_j}(\gamma_j-1)!},\ x>0, \tag{1.2}$$

with common scale parameter $\theta>0$ and positive integer shape parameter γ_j. To ensure the unique expression of density function (1.1), we assume that $\gamma_1<\gamma_2<\cdots<\gamma_m$ as we did in Yin and Lin (2016). The Erlang mixture model and its multivariate version have been widely used in modeling insurance losses due to its desirable distributional properties. For example, risk measures such as VaR and TVaR can be calculated easily. For more details on the applications, see Lee and Lin (2010), Cossette et al. (2013), Porth et al. (2014), Verbelen et al. (2015), Hashorva and Ratovomirija (2015), Verbelen et al. (2016), and references therein.

As a mixture model, an expectation–maximization (EM) algorithm is naturally used to fit the model to data by estimating the scale parameter and the mixing weights. However, the shape parameter of each of the Erlang components is not estimated. In order to include all possible Erlang distributions for component selection, one must start with a large number of components in an Erlang mixture when running the EM algorithm. Over-fitting could be a concern in this situation. To

[*] Corresponding author.
E-mail address: sheldon@utstat.utoronto.ca (X. Sheldon Lin).

https://doi.org/10.1016/j.spl.2018.08.004
0167-7152/© 2018 Elsevier B.V. All rights reserved.

maintain the goodness of fit and to avoid over-fitting at the same time, an ad hoc method for shape parameter selection and BIC are used. See Lee and Lin (2010) and Verbelen et al. (2015). Several issues arise. First, the ad hoc method requires repeated runs of the EM algorithm, which can be computationally burdensome. Second, the chosen shape parameters are often suboptimal in terms of the order of the mixture. Third, using BIC often results in a poor fit of a model to the sparse right tail of the data, a major shortcoming in insurance loss modeling and risk measure calculation. Last, statistical properties of the corresponding estimators cannot be obtained under the ad hoc approach. Yin and Lin (2016) propose a new thresholding penalty function, termed as iSCAD, to penalize the likelihood when estimating the scale parameter and the mixing weights of the Erlang mixture. This approach is motivated by the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) in regression analysis and the MSCAD introduced in Chen and Khalili (2008) for Gaussian mixtures. The thresholding feature of the proposed penalty ensures the sparsity of the mixture, which allows us to avoid over-fitting and maintain fitting accuracy at the same time. Moreover, the structure of the penalty results in the unbiasedness and continuity in estimation.

In this paper, we turn to the issue of consistency of the estimates including the order estimate when using the iSCAD penalized likelihood for the Erlang mixture model, as the consistency of the order is one of the most important statistical issues for mixture modeling. In the current statistics literature, most research focuses on Gaussian mixtures and a number of methods have been proposed. See Leroux (1992), James et al. (2001), Keribin (2000), Ciuperca et al. (2003), Ahn (2009), Chen et al. (2012) and references therein. However, few research have been done on non-negative non-Gaussian mixtures and few existing results may be directly applicable to the aforedescribed Erlang mixture.

In this paper, we examine the consistency of the estimators of the weight parameter and common scale parameter, as well as the order estimator, when using the iSCAD penalized likelihood. In Section 2 we introduce the iSCAD penalty, the corresponding penalized likelihood and the estimators obtained from the maximum penalized likelihood. Main results and their proofs are given in Section 3 in which we show that the estimators are consistent.

2. The iSCAD penalty

Yin and Lin (2016) propose a new penalty function termed as iSCAD, which penalizes individually the weights of an Erlang mixture. For each weight π_j, $j = 1, \ldots, m$, the iSCAD penalty function is defined as

$$\begin{aligned} P_\lambda(\pi_j) = {} & \lambda\{\log \frac{a\lambda+\varepsilon}{\varepsilon} + \frac{a^2\lambda^2}{2} - \frac{a\lambda}{a\lambda+\varepsilon}\} I(\pi_j > a\lambda) \\ & + \lambda\{\log \frac{\pi_j+\varepsilon}{\varepsilon} - \frac{\pi_j^2}{2} + (a\lambda - \frac{1}{a\lambda+\varepsilon})\pi_j\} I(\pi_j \le a\lambda), \end{aligned} \tag{2.1}$$

where λ is a tuning parameter that is a function of n with condition $\lambda \to 0$, as $n \to \infty$. $a = \frac{m}{m-\lambda} > 1$ is to ensure that the estimator $\hat{\pi}_j$ of π_j is continuous and parameter $\varepsilon = \lambda^{3/2}$ is to ensure that the range of π_j includes 0. These conditions are motivated by the conditions in Theorem 4 of Leroux (1992) to ensure to not overestimate the order of the Erlang mixture. The tuning parameter will ensure the sparsity of the mixture as it serves as a lower bound of the mixing weights. This property is crucial to avoid over-fitting and maintain fitting accuracy at the same time. Moreover, the structure of the iSCAD penalty and its derivative will result in the unbiasedness and continuity in estimation of the mixing distribution when an EM algorithm is used.

Insurance loss/claim data are mostly left truncated with known truncation points (in the form of a deductible or retention limit). See the data sets in Beirlant et al. (2006) and Verbelen et al. (2015). Suppose l to be a truncation point. Then, the probability density function of a left-truncated Erlang mixture is

$$\begin{aligned} h(x;\boldsymbol{\phi}) &= \frac{h(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}{\overline{H}(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)} = \sum_{j=1}^{m} \alpha_j \frac{g(x;\gamma_j,\theta)}{\overline{H}(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)} \\ &= \sum_{j=1}^{m} \alpha_j \frac{\overline{G}(l;\gamma_j,\theta)}{\overline{H}(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)} \frac{g(x;\gamma_j,\theta)}{\overline{G}(l;\gamma_j,\theta)} = \sum_{j=1}^{m} \pi_j\, g_\theta(x;l,\gamma_j), \end{aligned} \tag{2.2}$$

where $\boldsymbol{\phi} = (\pi_1, \ldots, \pi_m, \theta)$. There, $\overline{H}(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)$ and $\overline{G}(x;\gamma_j,\theta)$ are the survival functions of $h(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)$ and $g(x;\gamma_j,\theta)$, respectively,

$$g_\theta(x;l,\gamma_j) = \frac{g(x;\gamma_j,\theta)}{\overline{G}(l;\gamma_j,\theta)},$$

and

$$\pi_j = \alpha_j \frac{\overline{G}(l;\gamma_j,\theta)}{\overline{H}(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}. \tag{2.3}$$

Further, let $G_\theta(x;l,\gamma_j)$ be the cumulative distribution function of $g_\theta(x;l,\gamma_j)$.

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random sample of size n from a sparse Erlang mixture with density (2.2). Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be its realization. The penalized log-likelihood function with iSCAD penalty is defined as

$$L_n(\boldsymbol{\phi}) = \log f_n(x_1, \ldots, x_n; \boldsymbol{\phi}) = \sum_{i=1}^{n} \log h(x_i; \boldsymbol{\phi}) - n \sum_{j=1}^{m} P_\lambda(\pi_j). \tag{2.4}$$

That is, $f_n(x_1, \ldots, x_n; \boldsymbol{\phi}) = \prod_{i=1}^{n} h(x_i; \boldsymbol{\phi}) \exp\{-n\sum_{j=1}^{m} P_\lambda(\pi_j)\}$, and in particular,

$$f_1(x; \boldsymbol{\phi}) = h(x; \boldsymbol{\phi}) \exp\{-\sum_{j=1}^{m} P_\lambda(\pi_j)\}. \tag{2.5}$$

Yin and Lin (2016) applied an EM algorithm to (2.4) and obtained penalized maximum likelihood estimators for the mixing weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ and the common scale parameter θ as described in the following. Suppose that the kth iteration in the EM algorithm gives the estimates $\boldsymbol{\phi}^{(k)} = (\pi_1^{(k)}, \ldots, \pi_m^{(k)}, \theta^{(k)})$. The values of the parameters in the next iteration are obtained by the following formulas

$$\pi_j^{(k+1)} = \bar{q}_j^{(k)} I(\bar{q}_j^{(k)} > a\lambda) + \frac{m}{\lambda}(\bar{q}_j^{(k)} - \lambda)_+ I(\bar{q}_j^{(k)} \leq a\lambda), \tag{2.6}$$

where $\bar{q}_j^{(k)} \triangleq \frac{\sum_{i=1}^{n} q(j|x_i, \boldsymbol{\phi}^{(k)})}{n}$. Here, $q(j \mid x_i, \boldsymbol{\phi}^{(k)})$ is probability of the observation x_i coming from the jth component:

$$q(j \mid x_i, \boldsymbol{\phi}^{(k)}) = \frac{\pi_j^{(k)} g_{\theta^{(k)}}(x_i; l, \gamma_j)}{\sum_{j=1}^{m} \pi_j^{(k)} g_{\theta^{(k)}}(x_i; l, \gamma_j)}.$$

The updated estimate of the scale parameter θ is given by

$$\theta^{(k+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i - t^{(k)}}{\sum_{j=1}^{m} \gamma_j \bar{q}_j^{(k)}},$$

where

$$t^{(k)} = \sum_{j=1}^{m} \bar{q}_j^{(k)} \left. \frac{l^{\gamma_j} e^{-l/\theta}}{\theta^{\gamma_j - 1}(\gamma_j - 1)!\, \overline{G}(l; \gamma_j, \theta)} \right|_{\theta = \theta^{(k)}}.$$

Let $\hat{\boldsymbol{\phi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_m, \hat{\theta})$ be the estimated values in the final EM iteration. The estimate of the order of the mixture is

$$\hat{m} = \#\{\hat{\pi}_j | \hat{\pi}_j \neq 0, j = 1, \ldots, m\}. \tag{2.7}$$

For notational cleanness, we rename the shape parameters $\{\gamma_j | \hat{\pi}_j \neq 0, j = 1, \ldots, m\}$ in the increasing order as $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{\hat{m}})$ and the corresponding mixing weights $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_{\hat{m}})$. Finally, from (2.3) the estimates of the corresponding original weight parameters $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\hat{m}})$ are obtained as

$$\tilde{\alpha}_j = c \frac{\tilde{\pi}_j}{\overline{G}(l; \tilde{\gamma}_j, \hat{\theta})}. \tag{2.8}$$

where c is a normalizing constant such that $\sum_{j=1}^{\hat{m}} \tilde{\alpha}_j = 1$.

3. Consistency results and their proofs

In this section we show that the estimators obtained in the proposed penalized MLE are consistent beginning with several lemmas. Recall that the density of the Erlang distribution with left truncated point l and shape parameter γ_j is

$$g_\theta(x; l, \gamma_j) = \frac{g(x; \gamma_j, \theta)}{\overline{G}(l; \gamma_j, \theta)}, \quad x > l. \tag{3.1}$$

Let the range of the scale parameter space be $\Theta = \{\theta : c_1 \leq \theta \leq c_2$, where c_1 and c_2 are positive constants$\}$ with the true parameter $\theta_0 \in \Theta$.

The first lemma concerns boundedness properties of the truncated Erlang (3.1).

Lemma 3.1. *For $\theta \in \Theta$ and sufficiently small $\rho_\theta > 0$, define $g_{\theta,\rho_\theta}(x; l, \gamma_j) = \sup_{\theta'}\{g_{\theta'}(x; l, \gamma_j) : |\theta - \theta'| \leq \rho_\theta,\ \theta' \in \Theta\}$ and $g^*_{\theta,\rho_\theta}(x; l, \gamma_j) = \max(1, g_{\theta,\rho_\theta}(x; l, \gamma_j))$. Then,*

$$\int_l^\infty \log g^*_{\theta,\rho_\theta}(x; l, \gamma_j) dG_{\theta_0}(x; l, \gamma_j) < \infty$$

and

$$\int_l^{\infty} |\log g_\theta(x; l, \gamma_j)| dG_{\theta_0}(x; l, \gamma_j) < \infty.$$

Proof. It is well known that the mode of Erlang distribution (1.2) is $\theta \cdot (\gamma_j - 1)$ when $\gamma_j \geq 1$. If $\gamma_j > 1$, the density of left-truncated Erlang (3.1) is bounded as

$$g_\theta(x; l, \gamma_j) = \frac{g(x; \gamma_j, \theta)}{\bar{G}(l; \gamma_j, \theta)} = \frac{x^{\gamma_j-1}e^{-x/\theta}}{\bar{G}(l; \gamma_j, \theta)\theta^{\gamma_j}(\gamma_j-1)!} \leq \frac{(\gamma_j-1)^{\gamma_j-1}e^{-(\gamma_j-1)}}{\bar{G}(l; \gamma_j, \theta)\theta(\gamma_j-1)!}.$$

If $\gamma_j = 1$ and $x \geq l$, the density of left-truncated Erlang (3.1) is bounded as

$$g_\theta(x; l, \gamma_j = 1) = \frac{e^{-x/\theta}}{\bar{G}(l; \gamma_j = 1, \theta)\theta} \leq \frac{e^{-l/\theta}}{\bar{G}(l; \gamma_j = 1, \theta)\theta}.$$

Define $B(\theta) = \max\{\frac{(\gamma_j-1)^{\gamma_j-1}e^{-(\gamma_j-1)}}{\bar{G}(l;\gamma_j,\theta)\theta(\gamma_j-1)!}, \frac{e^{-l/\theta}}{\bar{G}(l;\gamma_j=1,\theta)\theta}, 1\}$. Obviously, $g_\theta(x; l, \gamma_j) \leq B(\theta)$. Thus, we have

$$\begin{aligned}
&\int_l^{+\infty} \log g^*_{\theta,\rho_\theta}(x; l, \gamma_j) dG_{\theta_0}(x; l, \gamma_j)\\
&= \int_l^{+\infty} \log(\max\{1, g_{\theta,\rho_\theta}(x; l, \gamma_j)\}) dG_{\theta_0}(x; l, \gamma_j)\\
&\leq \int_{\{x|g_{\theta,\rho_\theta}(x;l,\gamma_j)>1, x\geq l\}} \log \sup_{|\theta'-\theta|\leq\rho_\theta} B(\theta') dG_{\theta_0}(x; l, \gamma_j) < \infty.
\end{aligned}$$

To show $\int_l^{+\infty} |\log g_\theta(x; l, \gamma_j)| dG_{\theta_0}(x; l, \gamma_j) < \infty$, we have the following

$$\begin{aligned}
&\int_l^{+\infty} |\log g_\theta(x; l, \gamma_j)| dG_{\theta_0}(x; l, \gamma_j)\\
&= \int_l^{+\infty} |(\gamma_j - 1)\log x - \frac{x}{\theta}| dG_{\theta_0}(x; l, \gamma_j) + C_1(j)\\
&= \int_l^{1} |(\gamma_j - 1)\log x - \frac{x}{\theta}| dG_{\theta_0}(x; l, \gamma_j) + \int_1^{+\infty} |(\gamma_j - 1)\log x - \frac{x}{\theta}| dG_{\theta_0}(x; l, \gamma_j) + C_1(j)\\
&\leq (\gamma_j - 1 + \frac{1}{\theta}) \int_1^{+\infty} x dG_{\theta_0}(x; l, \gamma_j) + C_2(j) < \infty,
\end{aligned}$$

where $C_1(j) = |\log(\theta^{\gamma_j}(\gamma_j - 1)!\bar{G}(l; \gamma_j, \theta))|$ and $C_2(j) = C_1(j) + (|(\gamma_j - 1)\log l| + \frac{1}{\theta})(G_{\theta_0}(1; l, \gamma_j) - G_{\theta_0}(l; l, \gamma_j))$ are both constants, $j = 1, \ldots, m$. □

In the next lemma, we show the properties in Lemma 3.1 can be extended to the Erlang mixture model. The proof is similar to that in Redner (1981). Denote the parameter space of the mixture as

$$\varPhi = \{\boldsymbol{\phi} = (\pi_1, \ldots, \pi_m, \theta) : \sum_{j=1}^{m} \pi_j = 1, \pi_j \geq 0, 0 < c_1 \leq \theta \leq c_2.\}$$

with the true parameters being $\boldsymbol{\phi}_0 = (\pi_{0,1}, \ldots, \pi_{0,m}, \theta_0) \in \varPhi$, in which there are only m_0 non-zero weights. For any $\boldsymbol{\phi}', \boldsymbol{\phi} \in \varPhi$, we define the distance between $\boldsymbol{\phi}'$ and $\boldsymbol{\phi}$ as $|\boldsymbol{\phi}' - \boldsymbol{\phi}| = \sum_{j=1}^{m} \arctan|\pi_j' - \pi_j| + |\theta' - \theta|$.

Lemma 3.2. *Recall that the mixture model (2.2) with the left truncated point $l > 0$ is given by*

$$h(x; \boldsymbol{\phi}) = \sum_{j=1}^{m} \pi_j g_\theta(x; l, \gamma_j), \ x > l,$$

with pre-given l and $\gamma_j, j = 1, \ldots, m$. Let $H(x; \boldsymbol{\phi})$ be the cumulative distribution function.

For each $\boldsymbol{\phi} \in \varPhi$ and sufficiently small $\rho_{\boldsymbol{\phi}} > 0$,

$$\int_l^{\infty} \log h^*(x; \boldsymbol{\phi}, \rho_{\boldsymbol{\phi}}) dH(x; \boldsymbol{\phi}_0) < \infty$$

and

$$\int_l^{\infty} |\log h(x; \boldsymbol{\phi})| dH(x; \boldsymbol{\phi}_0) < \infty,$$

where $h(x; \boldsymbol{\phi}, \rho_{\boldsymbol{\phi}}) = \sup_{\boldsymbol{\phi}'}\{h(x; \boldsymbol{\phi}') : |\boldsymbol{\phi}' - \boldsymbol{\phi}| \leq \rho_{\boldsymbol{\phi}}, \boldsymbol{\phi}' \in \varPhi\}$, $h^*(x; \boldsymbol{\phi}, \rho_{\boldsymbol{\phi}}) = \max(1, h(x; \boldsymbol{\phi}, \rho_{\boldsymbol{\phi}}))$.

Proof. The first boundedness property holds because

$$\begin{aligned}
&\int_l^\infty \log h^*(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})dH(x;\boldsymbol{\phi}_0)\\
&=\int_l^\infty \log\max\{1,\sup_{|\boldsymbol{\phi}'-\boldsymbol{\phi}|\le\rho_{\boldsymbol{\phi}}}\sum_{j=1}^m \pi_j' g_{\theta'}(x;l,\gamma_j)\}dH(x;\boldsymbol{\phi}_0)\\
&\le\sum_{j=1}^m\int_l^\infty \log\max\{1,\sup_{|\boldsymbol{\phi}'-\boldsymbol{\phi}|\le\rho_{\boldsymbol{\phi}}} g_{\theta'}(x;l,\gamma_j)\}dH(x;\boldsymbol{\phi}_0)\\
&=\sum_{j=1}^m\int_l^\infty \log g^*_{\theta,\rho_{\boldsymbol{\phi}}}(x;l,\gamma_j)dH(x;\boldsymbol{\phi}_0)<\infty.
\end{aligned}$$

To show the second boundedness property, let $A_1=\{x\in[l,\infty)|\sum_{j=1}^m\pi_j g_\theta(x;l,\gamma_j)\ge 1\}$ and $A_2=[l,\infty)\backslash A_1$. Then we have

$$\begin{aligned}
&\int_l^\infty |\log\sum_{j=1}^m \pi_j g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)\\
&=\int_{A_1}|\log\sum_{j=1}^m \pi_j g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)+\int_{A_2}|\log\sum_{j=1}^m \pi_j g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)\\
&\le\sum_{j=1}^m\int_{A_1}|\log g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)+\sum_{j=1}^m\int_{A_2}|\log g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)\\
&\le\sum_{j=1}^m\int_l^\infty|\log g_\theta(x;l,\gamma_j)|dH(x;\boldsymbol{\phi}_0)<\infty.\quad\square
\end{aligned}$$

The next lemma shows that the order of the mixture model cannot be underestimated using an idea in Theorem 4 of Leroux (1992). Divide the parameter space Φ into $\Phi^*=\{\boldsymbol{\phi}\in\Phi : m^*<m_0\}$ and $\Phi^{**}=\{\boldsymbol{\phi}\in\Phi : m^*\ge m_0\}$, where m^* is the number of non-zero weights.

Lemma 3.3. *For each $\boldsymbol{\phi}\in\Phi^*$, $L_n(\boldsymbol{\phi})<L_n(\boldsymbol{\phi}_0)$.*

Proof. It follows from Lemma 2.4 of Newey and McFadden (1994) that

$$\sup_{\boldsymbol{\phi}\in\Phi^*}\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi})}{n}\xrightarrow{P}\int h(\boldsymbol{x};\boldsymbol{\phi}_0)\log h(\boldsymbol{x};\boldsymbol{\phi})d\boldsymbol{x}\ as\ n\to\infty.$$

Because of the law of large number, we also have

$$\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi}_0)}{n}\xrightarrow{P}\int h(\boldsymbol{x};\boldsymbol{\phi}_0)\log h(\boldsymbol{x};\boldsymbol{\phi}_0)d\boldsymbol{x}\ as\ n\to\infty.$$

Thus, for $n\to\infty$,

$$\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi}_0)}{n}-\sup_{\boldsymbol{\phi}\in\Phi^*}\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi})}{n}\xrightarrow{P}\int h(\boldsymbol{x};\boldsymbol{\phi}_0)\log(h(\boldsymbol{x};\boldsymbol{\phi}_0)/h(\boldsymbol{x};\boldsymbol{\phi}))d\boldsymbol{x}.\tag{3.2}$$

Since $\boldsymbol{\phi}\in\Phi^*$, $m^*<m_0$ and hence $\boldsymbol{\phi}\ne\boldsymbol{\phi}_0$. The identifiability property of the mixture model implies that $h(\boldsymbol{x};\boldsymbol{\phi})\ne h(\boldsymbol{x};\boldsymbol{\phi}_0)$, and the properties of Kullback–Leibler divergence lead to

$$\int h(\boldsymbol{x};\boldsymbol{\phi}_0)\log(h(\boldsymbol{x};\boldsymbol{\phi}_0)/h(\boldsymbol{x};\boldsymbol{\phi}))d\boldsymbol{x}>0.$$

As a result,

$$\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi}_0)}{n}-\frac{\sum_{i=1}^n\log h(X_i;\boldsymbol{\phi})}{n}>0,\tag{3.3}$$

for each $\boldsymbol{\phi}\in\Phi^*$ and for large n.

For any weight parameter π_j of $\boldsymbol{\phi}\in\Phi^*$, the definition of the iSCAD penalty implies that when $\pi_j=0$, $P_{\lambda_n}(\pi_j)=0$, and when $\pi_j>0$,

$$\lim_{n\to\infty}P_{\lambda_n}(\pi_j)=\lim_{n\to\infty}\lambda_n\{\log\frac{a\lambda_n+\varepsilon_n}{\varepsilon_n}+\frac{a^2\lambda_n^2}{2}-\frac{a\lambda_n}{a\lambda_n+\varepsilon_n}\}$$

$$= \lim_{n\to\infty}[\lambda_n(\log(a\lambda_n+\varepsilon_n))] - \lim_{n\to\infty}(\lambda_n \log \varepsilon_n)$$
$$+ \lim_{n\to\infty}\lambda_n(\frac{a^2\lambda_n^2}{2} - \frac{a\lambda_n}{a\lambda_n+\varepsilon_n}) = 0.$$

We thus have

$$\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}) - \sum_{j=1}^m P_{\lambda_n}(\pi_j) \xrightarrow{P} 0, \ \mathit{as}\ n\to\infty. \tag{3.4}$$

Combining (3.3) and (3.4), we have, for large n,

$$\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}) - \sum_{j=1}^m P_{\lambda_n}(\pi_j) < \frac{\sum_{i=1}^n \log h(X_i;\boldsymbol{\phi}_0)}{n} - \frac{\sum_{i=1}^n \log h(X_i;\boldsymbol{\phi})}{n}. \tag{3.5}$$

Equivalently,

$$L_n(\boldsymbol{\phi}) = \sum_{i=1}^n \log h(X_i;\boldsymbol{\phi}) - n\sum_{j=1}^m P_{\lambda_n}(\pi_j) < \sum_{i=1}^n \log h(X_i;\boldsymbol{\phi}_0) - n\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}) = L_n(\boldsymbol{\phi}_0). \tag{3.6}$$

The lemma is therefore proved. □

The next lemma concerns the expected penalized log-likelihood and we show that it is maximized at $\boldsymbol{\phi}_0$.

Lemma 3.4. *For each $\boldsymbol{\phi} \in \Phi^{**}$, $\boldsymbol{\phi} \neq \boldsymbol{\phi}_0$, we have*

$$L(\boldsymbol{\phi}) = \mathbb{E}_0 \log f_1(X;\boldsymbol{\phi}) < \mathbb{E}_0 \log f_1(X;\boldsymbol{\phi}_0) = L(\boldsymbol{\phi}_0).$$

Proof. To prove, we re-express the iSCAD penalty function (2.1) as

$$P_{\lambda_n}(\pi_j) = A\cdot I(\pi_j > a\lambda_n) + B(\pi_j)\cdot I(\pi_j \le a\lambda_n),$$

where $A = \lambda_n\{\log\frac{a\lambda_n+\varepsilon_n}{\varepsilon_n} + \frac{a^2\lambda_n^2}{2} - \frac{a\lambda_n}{a\lambda_n+\varepsilon_n}\}$ that does not contain π_j and $B(\pi_j) = \lambda_n\{\log\frac{\pi_j+\varepsilon_n}{\varepsilon_n} - \frac{\pi_j^2}{2} + (a\lambda_n - \frac{1}{a\lambda_n+\varepsilon_n})\pi_j\}$. For sufficient large n, if $\pi_j > 0$, $P_{\lambda_n}(\pi_j) = A$. Thus, $\sum_{j=1}^m P_{\lambda_n}(\pi_j) = m^*A$. Similarly, we have $\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}) = m_0A$. For each $\phi \in \Phi^{**}$ we have

$$\frac{\sum_{j=1}^m P_{\lambda_n}(\pi_j)}{\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j})} = \frac{m^*}{m_0} \ge 1. \tag{3.7}$$

That is,

$$\sum_{j=1}^m P_{\lambda_n}(\pi_j) \ge \sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}). \tag{3.8}$$

It follows from (2.5) and Lemma 3.2 that

$$\mathbb{E}_0|\log f_1(X;\boldsymbol{\phi}_0)| \le \mathbb{E}_0|\log h(X;\boldsymbol{\phi}_0)| + \sum_{k=1}^m P_{\lambda_n}(\pi_{0,k}) < \infty. \tag{3.9}$$

Let $U = \log f_1(X;\boldsymbol{\phi}) - \log f_1(X;\boldsymbol{\phi}_0)$. Then, $U \neq 0$ due to the identifiability of the mixture model (see Teicher, 1963), and

$$\mathbb{E}_0[e^U] = \mathbb{E}_0[\frac{f_1(X;\boldsymbol{\phi})}{f_1(X;\boldsymbol{\phi}_0)}] = \exp\{-(\sum_{j=1}^m P_{\lambda_n}(\pi_j) - \sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}))\} \le 1.$$

By Jensen's inequality,

$$\mathbb{E}_0 U < \log \mathbb{E}_0[e^U] \le 0.$$

The lemma is proved. □

The next lemma shows the convergence of the expected log-likelihood. The proof is similar to that Lemma 2 of Wald (1949).

Lemma 3.5. *For each $\boldsymbol{\phi} \in \Phi^{**}$ and sufficiently small $\rho_{\boldsymbol{\phi}} > 0$, let $f_1(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}}) = \sup_{\phi'}\{f_1(x;\boldsymbol{\phi}') : |\boldsymbol{\phi}-\boldsymbol{\phi}'| \le \rho_{\boldsymbol{\phi}}, \boldsymbol{\phi}' \in \Phi^{**}\}$, $f_1^*(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}}) = \max(1, f_1(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}}))$ and $f_1^*(x;\boldsymbol{\phi}) = \max(1, f_1(x;\boldsymbol{\phi}))$. Then,*

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0} \mathbb{E}_0 \log f_1(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}}) = \mathbb{E}_0 \log f_1(X;\boldsymbol{\phi}).$$

Proof. It follows from the definition that $\log f_1^*(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})$ is increasing in $\rho_{\boldsymbol{\phi}}$. The continuity of $f_1(X;\boldsymbol{\phi})$ at each $\boldsymbol{\phi}\in\Phi^{**}$ leads to

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0}\log f_1^*(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\log f_1^*(X;\boldsymbol{\phi}).$$

By the monotone convergence theorem, we have

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0}\mathbb{E}_0\log f_1^*(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\mathbb{E}_0\log f_1^*(X;\boldsymbol{\phi}). \tag{3.10}$$

Now, denote $f_1^{**}(x;\boldsymbol{\phi})=\min(1,f_1(x;\boldsymbol{\phi}))$ and $f_1^{**}(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\min(1,f_1(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}}))$. Clearly,

$$|\log f_1^{**}(x;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})|\le|\log f_1^{**}(x;\boldsymbol{\phi})|$$

(3.9) implies $\mathbb{E}_0|\log f_1^{**}(X;\boldsymbol{\phi})|<\infty$. The continuity of $f_1(X;\boldsymbol{\phi})$ at each $\boldsymbol{\phi}\in\Phi^{**}$ implies

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0}\log f_1^{**}(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\log f_1^{**}(X;\boldsymbol{\phi}),$$

and hence

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0}\mathbb{E}_0\log f_1^{**}(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\mathbb{E}_0\log f_1^{**}(X;\boldsymbol{\phi}) \tag{3.11}$$

by the dominated convergence theorem.

Since $\mathbb{E}_0\log f_1(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\mathbb{E}_0\log f_1^*(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})+\mathbb{E}_0\log f_1^{**}(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})$ and $\mathbb{E}_0\log f_1(X;\boldsymbol{\phi})=\mathbb{E}_0\log f_1^*(X;\boldsymbol{\phi})+\mathbb{E}_0\log f_1^{**}(X;\boldsymbol{\phi})$, it follows from (3.10) and (3.11) that Lemma 3.5 holds. □

We now present a main consistency result. Again, the proof is motivated by Wald (1949).

Theorem 3.1. *Let Ω be any closed subset of the parameter space Φ^{**}, and the true parameter $\boldsymbol{\phi}_0\notin\Omega$. Then, we have*

$$P\left(\lim_{n\to\infty}\frac{\sup_{\boldsymbol{\phi}\in\Omega}f_1(X_1;\boldsymbol{\phi})f_1(X_2;\boldsymbol{\phi})\cdots f_1(X_n;\boldsymbol{\phi})}{f_1(X_1;\boldsymbol{\phi}_0)f_1(X_2;\boldsymbol{\phi}_0)\cdots f_1(X_n;\boldsymbol{\phi}_0)}=0\right)=1.$$

Proof. It follows from Lemmas 3.4 and 3.5 that for each $\boldsymbol{\phi}\in\Omega$, there is a positive and sufficient small value $\rho_{\boldsymbol{\phi}}>0$ such that

$$\mathbb{E}_0\log f_1(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})<\mathbb{E}_0\log f_1(X;\boldsymbol{\phi}_0), \tag{3.12}$$

and

$$\lim_{\rho_{\boldsymbol{\phi}}\to 0}\mathbb{E}_0\log f_1(X;\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})=\mathbb{E}_0\log f_1(X;\boldsymbol{\phi})<\mathbb{E}_0\log f_1(X;\boldsymbol{\phi}_0). \tag{3.13}$$

Since Ω is compact, there exist a finite number of points $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_h$ in Ω such that $\Omega\subset\bigcup_{k=1}^h S(\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k})$, where $S(\boldsymbol{\phi},\rho_{\boldsymbol{\phi}})$ denotes a sphere with center $\boldsymbol{\phi}$ and radius $\rho_{\boldsymbol{\phi}}$. Then,

$$0\le\sup_{\boldsymbol{\phi}\in\Omega}\prod_{i=1}^n f_1(x_i;\boldsymbol{\phi})\le\sum_{k=1}^h\prod_{i=1}^n f_1(x_i;\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k}).$$

Thus, Theorem 3.1 will be proved if we can show that

$$P\left(\lim_{n\to\infty}\frac{f_1(X_1;\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k})f_1(X_2;\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k})\cdots f_1(X_n;\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k})}{f_1(X_1;\boldsymbol{\phi}_0)f_1(X_2;\boldsymbol{\phi}_0)\cdots f_1(X_n;\boldsymbol{\phi}_0)}=0\right)=1,\ k=1,2,\ldots,h.$$

Rewrite the above equations as

$$P\{\lim_{n\to\infty}\sum_{i=1}^n[\log f_1(X_i;\boldsymbol{\phi}_k,\rho_{\boldsymbol{\phi}_k})-\log f_1(X_i;\boldsymbol{\phi}_0)]=-\infty\}=1,\ k=1,\ldots,h.$$

Obviously, these equations hold due to (3.12) and the law of large number. This completes the proof. □

To show the consistency of the estimates, we give another lemma.

Lemma 3.6. *Let $\bar{\boldsymbol{\phi}}_n\in\Phi^{**}$ be a function of the observations $x_1,x_2,\ldots,x_n$ such that*

$$\frac{f_1(x_1;\bar{\boldsymbol{\phi}}_n)f_1(x_2;\bar{\boldsymbol{\phi}}_n)\cdots f_1(x_n;\bar{\boldsymbol{\phi}}_n)}{f_1(x_1;\boldsymbol{\phi}_0)f_1(x_2;\boldsymbol{\phi}_0)\cdots f_1(x_n;\boldsymbol{\phi}_0)}\ge c>0,\ \textit{for large } n. \tag{3.14}$$

Then

$$P\{\lim_{n\to\infty}\bar{\boldsymbol{\phi}}_n=\boldsymbol{\phi}_0\}=1.$$

Proof. It is sufficient to prove that for sufficiently small ρ_{ϕ_0}, the probability that all limit points $\bar{\boldsymbol{\phi}}$ of the sequence $\bar{\boldsymbol{\phi}}_n$ satisfy the inequality $|\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}_0| < \rho_{\phi_0}$ is one. If there exists a limit point $\bar{\boldsymbol{\phi}}$ of the sequence $\bar{\boldsymbol{\phi}}_n$ such that $|\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}_0| \geq \rho_{\phi_0}$, then

$$\sup_{|\boldsymbol{\phi}-\boldsymbol{\phi}_0|\geq\rho_{\phi_0}} \prod_{i=1}^{n} f_1(x_i; \boldsymbol{\phi}) \geq \prod_{i=1}^{n} f_1(x_i; \bar{\boldsymbol{\phi}}_n) \textit{ for large } n.$$

It follows from (3.14) that

$$\frac{\sup_{|\boldsymbol{\phi}-\boldsymbol{\phi}_0|\geq\rho_{\phi_0}} \prod_{i=1}^{n} f_1(x_i; \boldsymbol{\phi})}{f_1(x_1; \boldsymbol{\phi}_0) f_1(x_2; \boldsymbol{\phi}_0) \cdots f_1(x_n; \boldsymbol{\phi}_0)} \geq c > 0.$$

According to Theorem 3.1, this is an event with probability zero. Hence, the probability that all limit points $\bar{\boldsymbol{\phi}}$ of the sequence $\bar{\boldsymbol{\phi}}_n$ satisfy the inequality $|\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}_0| < \rho_{\phi_0}$ is always one. □

We now show that the estimates of all the parameters are consistent.

Theorem 3.2. *The penalized maximum likelihood estimators* $\hat{\boldsymbol{\phi}} \in \Phi^{**}$ *are consistent, i.e. the estimators* $\hat{\boldsymbol{\phi}}$ *which maximize* $\prod_{i=1}^{n} f_1(X_i; \boldsymbol{\phi})$ *are such that*

$$P\{\lim_{n\to\infty} \hat{\boldsymbol{\phi}} = \boldsymbol{\phi}_0\} = 1. \tag{3.15}$$

Proof. Since $\hat{\boldsymbol{\phi}} \in \Phi^{**}$, are the penalized maximum likelihood estimators of $\prod_{i=1}^{n} f_1(X_i; \boldsymbol{\phi})$, we have

$$\prod_{i=1}^{n} f_1(X_i; \hat{\boldsymbol{\phi}}) \geq \prod_{i=1}^{n} f_1(X_i; \boldsymbol{\phi}_0) \text{ or } \frac{\prod_{i=1}^{n} f_1(X_i; \hat{\boldsymbol{\phi}})}{\prod_{i=1}^{n} f_1(X_i; \boldsymbol{\phi}_0)} \geq 1.$$

Obviously, $\hat{\boldsymbol{\phi}}$ satisfies (3.14) with $c = 1$. Thus, we have

$$P\{\lim_{n\to\infty} \hat{\boldsymbol{\phi}} = \boldsymbol{\phi}_0\} = 1. \quad \square$$

Theorem 3.2 leads to the following corollary.

Corollary 3.1. *The penalized maximum likelihood estimators* $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m)$ *are consistent, i.e.*

$$P\{\lim_{n\to\infty} \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0\} = 1.$$

where $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,m})$ *is the true weight parameters of the original Erlang mixture* (1.1) .

Proof. From Theorem 3.2 and (2.8), for $j = 1, \ldots, m$,

$$\begin{aligned} \|\hat{\alpha}_j - \alpha_{0,j}\| &= c \left\| \frac{\hat{\pi}_j}{\overline{G}(l; \gamma_j, \hat{\theta})} - \frac{\pi_{0,j}}{\overline{G}(l; \gamma_j, \theta_0)} \right\| \\ &= c \left\| \frac{\hat{\pi}_j - \pi_{0,j}}{\overline{G}(l; \gamma_j, \hat{\theta})} + \frac{\pi_{0,j}}{\overline{G}(l; \gamma_j, \hat{\theta})} - \frac{\pi_{0,j}}{\overline{G}(l; \gamma_j, \theta_0)} \right\| \\ &\leqslant c \left\| \frac{\hat{\pi}_j - \pi_{0,j}}{\overline{G}(l; \gamma_j, \hat{\theta})} \right\| + \pi_{0,j} \cdot c \left\| \frac{1}{\overline{G}(l; \gamma_j, \hat{\theta})} - \frac{1}{\overline{G}(l; \gamma_j, \theta_0)} \right\| \to 0. \quad \square \end{aligned}$$

Theorem 3.3. *If* $\liminf_{n\to\infty} n\lambda_n > 0$, *then the estimator of the order of mixture* $\hat{m}$ *is also consistent. i.e. the estimator is such that*

$$P\{\lim_{n\to\infty} \hat{m} = m_0\} = 1.$$

Remark. The condition $\liminf_{n\to\infty} n\lambda_n > 0$ is always satisfied with the 'optimal' choice of λ_n in Yin and Lin (2016) in which $\lambda_n = \frac{c(1+\sqrt{m})}{m^{3/2}\sqrt{n}}$.

Proof. As shown in Theorem 3.2, the penalized maximum likelihood estimators $\hat{\boldsymbol{\phi}}$ are consistent, i.e., $P\{\lim_{n\to\infty} \hat{\boldsymbol{\phi}} = \boldsymbol{\phi}_0\} = 1$. As a result, we have

$$P\{\liminf_{n\to\infty} \hat{m} \geq m_0\} = 1, \tag{3.16}$$

$\hat{m}$ is the estimate of the order of the mixture defined in (2.7).

Clearly,

$$P(\hat{m} > m_0) \leq P(L_n(\hat{\boldsymbol{\phi}}) \geq L_n(\boldsymbol{\phi}_0)).$$

Rewrite the right hand side of the above as

$$P(L_n(\hat{\boldsymbol{\phi}}) \geq L_n(\boldsymbol{\phi}_0)) = P\left(\frac{L_n^0(\hat{\boldsymbol{\phi}}) - L_n^0(\boldsymbol{\phi}_0)}{n\sum_{k=1}^m P_{\lambda_n}(\pi_{0,k})} - \frac{\sum_{j=1}^m P_{\lambda_n}(\hat{\pi}_j)}{\sum_{k=1}^m P_{\lambda_n}(\pi_{0,k})} + 1 \geq 0\right),$$

where $L_n^0(\boldsymbol{\phi})$ is the likelihood without penalty. Since (3.15) implies that $L_n^0(\hat{\boldsymbol{\phi}}) \to L_n^0(\boldsymbol{\phi}_0)$, as $n \to \infty$, and under the condition of this theorem

$$\liminf_{n\to\infty} n\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j}) = \liminf_{n\to\infty} \sum_{\pi_{0,j}>0, j=1,\ldots,m} n\lambda_n\{\log\frac{a\lambda_n+\varepsilon_n}{\varepsilon_n} + \frac{a^2\lambda_n^2}{2} - \frac{a\lambda_n}{a\lambda_n+\varepsilon_n}\} = \infty,$$

we have

$$\frac{L_n^0(\hat{\boldsymbol{\phi}}) - L_n^0(\boldsymbol{\phi}_0)}{n\sum_{k=1}^m P_{\lambda_n}(\pi_{0,k})} \to 0, \text{ as } n \to \infty. \tag{3.17}$$

Based on (3.8), if $\hat{m} > m_0$,

$$\frac{\sum_{j=1}^m P_{\lambda_n}(\hat{\pi}_j)}{\sum_{k=1}^m P_{\lambda_n}(\pi_{0,k})} = \frac{\hat{m}}{m_0} > 1, \text{ for sufficient large } n. \tag{3.18}$$

It follows from (3.17) and (3.18) that

$$\frac{L_n^0(\hat{\boldsymbol{\phi}}) - L_n^0(\boldsymbol{\phi}_0)}{n\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j})} - \frac{\sum_{j=1}^m P_{\lambda_n}(\hat{\pi}_j)}{\sum_{j=1}^m P_{\lambda_n}(\pi_{0,j})} + 1 < 0, \text{ for sufficient large } n.$$

Thus,

$$P(L_n(\hat{\boldsymbol{\phi}}) \geq L_n(\boldsymbol{\phi}_0)) = 0, \text{ for sufficient large } n.$$

Equivalently,

$$P\{\hat{m} > m_0, \text{ for sufficient large } n\} = 0,$$

Combining (3.16), we have

$$P\{\lim_{n\to\infty} \hat{m} = m_0\} = 1. \quad \square$$

Acknowledgments

This research was partly supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-06684) and the Natural Science Foundation of China (No. 11471272).

References

Ahn, S.M., 2009. Note on the consistency of a penalized maximum likelihood estimate. Commun. Stat. Appl. Methods 16 (4), 573–578.
Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2006. Statistics of Extremes: Theory and Applications. John Wiley & Sons.
Chen, J., Khalili, A., 2008. Order selection in finite mixture models with a nonsmooth penalty. J. Amer. Statist. Assoc. 103 (484), 1674–1683.
Chen, J., Li, P., Fu, Y., 2012. Inference on the order of a normal mixture. J. Amer. Statist. Assoc. 107 (499), 1096–1105.
Ciuperca, G., Ridolfi, A., Idier, J., 2003. Penalized maximum likelihood estimator for normal mixtures. Scand. J. Stat. 30 (1), 45–59.
Cossette, H., Cote, M.P., Marceau, E., Moutanabbir, K., 2013. Multivariate distribution defined with Farlie–Gumbel–Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation. Insurance Math. Econom. 52 (3), 560–572.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Hashorva, E., Ratovomirija, G., 2015. On Sarmanov mixed Erlang risks in insurance applications. ASTIN Bulletin J. IAA 45 (01), 175–205.
James, L.F., Priebe, C.E., Marchette, D.J., 2001. Consistent estimation of mixture complexity. Ann. Statist. 29 (5), 1281–1296.
Keribin, C., 2000. Consistent estimation of the order of mixture models. Sankhya Indian J. Stat., Ser. A 49–66.
Lee, S.C., Lin, X.S., 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. N. Am. Actuar. J. 14 (1), 107–130.
Leroux, B.G., 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20 (3), 1350–1360.
Newey, W.K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. In: Handbook of Econometrics, vol. 4, pp. 2111–2245.
Porth, L., Zhu, W., Tan, K.S., 2014. A credibility-based Erlang mixture model for pricing crop reinsurance. Agric. Finance Review 74 (2), 162–187.
Redner, R., 1981. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann. Statist. 225–228.
Teicher, H., 1963. Identifiability of finite mixtures. Ann. Math. Stat. 1265–1269.
Verbelen, R., Antonio, A., Gong, L., Lin, X.S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the em algorithm. ASTIN Bulletin J. IAA 45 (3), 729–758.
Verbelen, R., Antonio, K., Claeskens, G., 2016. Multivariate mixtures of Erlangs for density estimation under censoring. Lifetime Data Anal. 22 (3), 429–455.
Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20 (4), 595–601.
Yin, C., Lin, X.S., 2016. Efficient estimation of Erlang mixtures using iSCAD penalty with insurance application. ASTIN Bulletin J. IAA 46 (3), 779–799.