Consistency of the MLE under a two-parameter Gamma mixture model with a structural shape parameter

Mingxing He1 · Jiahua Chen2,3

Received: 12 December 2020 / Accepted: 12 December 2021 / Published online: 11 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Finite Gamma mixture models are often used to describe randomness in income data, insurance data, and data in applications where the response values are intrinsically positive. The popular likelihood approach for model fitting, however, does not work for this model because its likelihood function is unbounded. Because of this, the maximum likelihood estimator is not well-defined. Other approaches have been developed to achieve consistent estimation of the mixing distribution, such as placing an upper bound on the shape parameter or adding a penalty to the log-likelihood function. In this paper, we show that if the shape parameter in the finite Gamma mixture model is structural, then the direct maximum likelihood estimator of the mixing distribution is well-defined and strongly consistent. We also present simulation results demonstrating the consistency of the estimator. We illustrate the application of the model with a structural shape parameter to household income data. The fitted mixture distribution leads to several possible subpopulation structures with regard to the level of disposable income.

Keywords  EM algorithm · Finite Gamma mixture model · Maximum likelihood estimator · Strong consistency · Structural parameter

Mathematics Subject Classification 62H30 · 62H25

This work was supported by the National Natural Science Foundation of China (Grant No. 11871419) and the Natural Sciences and Engineering Research Council of Canada.

Jiahua Chen
jhchen@stat.ubc.ca

Mingxing He
hemingxing@kust.edu.cn

1 Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650091, China
2 Research Institute of Big Data, Yunnan University, Kunming 650221, China
3 Department of Statistics, University of British Columbia, Vancouver V7C 5K5, Canada
1 Introduction

Finite mixture distributions are widely used to model data collected from a heterogeneous population: the population contains several subpopulations, and each can be modeled with a distribution in a common parametric distribution family. Let \( \{ f(y; \theta) : \theta \in \Theta \} \) be the density functions of a parametric distribution family on a generic space \( \mathcal{Y} \) for \( y \) and parameter space. A finite mixture distribution of order \( m \) on this family has density function

\[
f(y; G) = \sum_{j=1}^{m} \alpha_j f(y; \theta_j),
\]

with mixing proportions \( \alpha_j \in [0, 1] \) such that \( \sum_{j=1}^{m} \alpha_j = 1 \), and subpopulation parameters \( \theta_j \in \Theta \). We name \( G \) a mixing distribution that assigns probability \( \alpha_j \) to value \( \theta_j \) in the subpopulation parameter space \( \Theta \). In this paper, we assume \( m \) is known. Because we do not exclude \( \alpha_j = 0 \) or equal \( \theta_j \) values, the true mixing distribution may have fewer support points. We name \( f(y; \theta) \) the subpopulation or component density function. When we include all distributions \( G \) on \( \Theta \), not merely those with a finite number of support points, and replace the summation in (1) with an integration, a general mixture model emerges.

There is a rich statistical literature on the theory and applications of mixture models. More than a hundred years ago, Pearson (1894) used a finite normal mixture distribution of order two to fit crab data suspected of containing two species. He used the method of moments because it is less computationally demanding. Kiefer and Wolfowitz (1956) proved the consistency of the maximum likelihood estimator (MLE) of \( G \) under some general conditions. Dempster et al. (1977) introduced the EM-algorithm, an easy-to-implement numerical method, to find the MLE for finite mixture models. The convergence of this algorithm was thoroughly discussed in Wu (1983). We have referred to McLachlan and Peel (2004) and Titterington et al. (1985) for early application examples and theoretical results. We refer to McLachlan et al. (2019) for a thorough review of recent developments.

A minimum requirement of a useful estimator is consistency: when the sample size goes to infinity, the error of the estimator goes to zero. Wald (1949) first showed that the MLE is consistent in general, given a set of \( n \) independent and identically distributed (IID) samples, as \( n \to \infty \). Kiefer and Wolfowitz (1956) observed that under some conditions, the consistency of the MLE under a mixture model is reduced to the same consistency problem discussed in Wald (1949). The consistency conclusion of the MLE given by Kiefer and Wolfowitz (1956), however, does not apply to some most important finite mixture models. Notably, the MLE of \( G \) under the finite normal mixture model is not well-defined because its likelihood function is unbounded. To estimate \( G \) consistently via the likelihood approach, one might apply a regularizing penalty function to the likelihood. The consistency of the resultant penalized MLE was rigorously established in Chen et al. (2008) and Chen and Tan (2009) for the univariate and multivariate cases respectively; see Ciuperca et al. (2003) and Tanaka (2009) for related developments. An alternative approach to achieving consistency would be
placing constraints on the range of the component parameters $\Theta$. For instance, the MLE is consistent if one places a finite bound on the ratio of any two-component variances; see Hathaway (1985), Tanaka and Takemura (2005), Tanaka and Takemura (2006), and Chen et al. (2016). A special case arises when the subpopulations of the finite normal mixture model share the same variance: the plain MLE is consistent, as seen in Chen (2017). Recently, Liu et al. (2019) showed that the MLE is consistent under finite mixtures of location-scale distributions with a structural scale parameter.

Although the finite mixture of Gaussian distributions has broad utility, there is increasing recognition of mixtures of asymmetric distributions (Young et al. 2019). In particular, there is an increased interest in finite mixtures of Gamma distributions. Gamma mixtures are broadly used in modeling the prices of commercial products, the cost of insurance, household income, and data in applications where the response values are intrinsically positive. Young et al. (2019) cited over 10 publications focusing on algorithm and applications of the finite Gamma mixture. A quick Google Scholar search led to a long list of application examples: Sergio et al. (2008) used a Gamma mixture for heavy-tailed distributions, Wywiał (2018) employed it for financial auditing, Muna et al. (2019) used it for crop insurance. Liu et al. (2003) and Wong and Li (2014) discussed the homogeneity test problem. A special type of Gamma mixture, the Erlang mixture, is applied to insurance (Willmot and Lin 2011) with its consistency discussed in Yin et al. (2019).

General discussions on inferential and algorithmic issues under finite Gamma mixture models can be found in Young et al. (2019) and Yu (2021). Interestingly, they were not aware of the likelihood function of the finite Gamma mixture model being unbounded (Chen et al. 2016). This led to the conceptual failure of the MLE under Gamma mixtures, the same as is the case under normal mixtures. Knowing the MLE is consistent under special normal mixtures, it is interesting and important to determine the consistency of the MLE under some special Gamma mixtures. In this paper, we prove the consistency of the MLE under the finite Gamma mixture model with a structural shape parameter.

This paper is organized as follows. In Sect. 2, we introduce some properties of the Gamma distribution and the finite Gamma mixture model as well as some technical results. In Sect. 3, we show that the MLE of the structural shape parameter almost surely falls into a compact interval $[\tau, \Delta]$. This crucial result paves the way for the final proof of the consistency of the MLE. In Sect. 4, we present and prove the consistency conclusion; the MLE under the two-parameter finite Gamma mixture model is consistent when the shape parameter is structural. In Sect. 5, we supplement the theoretical proof with simulation experiments to numerically demonstrate the consistency of the MLE. In Sect. 6, we fit the income data set with finite mixtures of Gamma distributions with various orders. We find that a model of order three or four provides a good fit, and the fitted models also suggest subpopulation structures. Sect. 7 provides a discussion.
2 Properties of the Gamma distribution and the finite Gamma mixture model

2.1 Preparation

We devote this subsection to the properties of the Gamma and the Gamma mixture models. Let $\Gamma(r)$ be the well-known Gamma function. The Gamma distribution has a density function given by

$$f(x; r, \theta) = \frac{x^{r-1} \exp(-x/\theta)}{\theta^r \Gamma(r)}$$

over $x > 0$ with shape parameter $r$ and scale parameter $\theta$. The parameter space of the two-parameter Gamma distribution family is given by

$$\Omega = \{(r, \theta) : 0 < r < \infty, 0 < \theta < \infty\} = \mathbb{R}^+ \times \mathbb{R}^+.$$

Suppose $X$ is a Gamma distributed random variable and $Y = \log X$. The density function of $Y$ with the same parameter space $\Omega$ becomes

$$g(y; r, \theta) = \frac{1}{\Gamma(r)} \exp\{r(y - \log \theta) - \exp(y - \log \theta)\}$$

for $y \in \mathbb{R}$. Clearly, this density function in $y$ is bounded for any $(r, \theta) \in \Omega$. Given a random sample from a finite Gamma mixture distribution, a logarithm transformation leads to a random sample from a finite log Gamma mixture distribution. The MLE of the mixing distribution based on the original data under the Gamma mixture model is the same as the MLE based on log-transformed data under the log Gamma mixture model. For this reason, we will regard the density in (2) as the component density to discuss the consistency of the MLE in this paper.

For a finite mixture of log Gamma distribution of order $m$ where the subpopulation distributions share an equal though unknown shape parameter $r$, the density function is given by

$$g(y; r, G) = \int_{\mathbb{R}^+} g(y; r, \theta) G(\theta) = \sum_{j=1}^{m} \alpha_j g(y; r, \theta_j).$$

In this case, $g(y; r, \theta)$ is the subpopulation/component density function, and the mixing proportions are $(\alpha_1, \alpha_2, \ldots, \alpha_m)$. The parameter $\theta_j$ is a support point of $G(\theta)$. We write $G = \sum_{j=1}^{m} \alpha_j \{\theta_j\}$.

In the above setting, $G$ mixes only scale parameter and leaves $r$ as a common shape parameter. For this reason, we call $r$ a structural parameter; its parameter space is $\mathbb{R}^+$. We denote the space of all mixing distributions with at most $m$ supports (known) in $\mathbb{R}^+$ as
\[ \mathbb{G}_m = \left\{ G : G = \sum_{j=1}^{m} \alpha_j \{\theta_j\}; \alpha_j \in [0, 1], \theta_j \in \mathbb{R}^+; \sum_{j=1}^{m} \alpha_j = 1; j = 1, \ldots, m \right\}. \] (4)

Note that \( \mathbb{G}_m \) permits \( \alpha_j = 0 \) or equal \( \theta_j \) values. We are interested in the Gamma mixture model with the parameter space of \((r, G)\) being \( \mathbb{R}^+ \times \mathbb{G}_m \).

### 2.2 Finite expectation and extended Glivenko–Cantelli theorem

The finite Gamma mixture model with a structural shape parameter has some nice properties that are easy to verify. Some of them are given below for subsequent reference.

**Lemma 1** Assume that \( X \) is a random variable with finite Gamma mixture distribution with density function \( f(x; r^*, G^*) \) where \( G^* = \sum_{j=1}^{m^*} \alpha_j^* \{\theta_j^*\} \) for some \( m^* \) and \( Y = \log(X) \). Let the density function of \( Y \) be \( g(y; r^*, G^*) \) as in (3). Then the expectations of \( Y \), \( \exp(Y) \), and \( \log\{g(Y; r^*, G^*)\} \) exist and are finite.

**Proof** The moment generating function of \( Y \) is given by

\[
MY(t) = \mathbb{E}^*\{\exp(tY)\} = \sum_{j=1}^{m^*} \{\alpha_j^*(\theta_j^*)^t \Gamma(r^* + t)/\Gamma(r^*)\}
\]

which is well defined for \( t > -r^* \), where the expectation operator \( \mathbb{E}^*(\cdot) \) is computed under \( g(y; r^*, G^*) \). Since the domain of this function contains 0 as an interior point, all the moments of \( Y \) are finite. Further, \( \mathbb{E}^*\{\exp(Y)\} = MY(1) < \infty \) so \( \exp(Y) \) has finite expectation.

As a function of \( y \), \( g(y; r^*, G^*) \) has a finite upper bound. Hence, it is clear that \( \mathbb{E}^*\{\log g(Y; r^*, G^*)\} < \infty \). In addition,

\[
\log g(Y; r^*, G^*) \geq \log \{\alpha_1^* g(Y; r^*, \theta_1^*)\} = \log \alpha_1^* - \log \Gamma(r^*) + r^*Y - r^* \log \theta_1^* - (\theta_1^*)^{-1} \exp(Y).
\]

Clearly, every term on the right-hand side has a finite expectation. Hence, we have \( \mathbb{E}^*\{\log g(Y; r^*, G^*)\} > -\infty \). This completes the proof. \( \Box \)

**Lemma 2** In the setting of Lemma 1, let \( M = \sup_y g(y; r^*, G^*) \). Given a set of independent and identically distributed (IID) observations \( Y_1, Y_2, \ldots, Y_n \) from \( g(y; r^*, G^*) \), for any fixed positive number \( \delta \), we have

\[
\sup_u \frac{1}{n} \sum_{i=1}^{n} I(|Y_i - u| < \epsilon) < 2M\epsilon + \delta \quad (5)
\]

uniformly in \( \epsilon \) almost surely.
Proof Let $F_n(y) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq y)$ be the empirical distribution function and $F(\cdot)$ be the distribution function of $Y$. Based on the Glivenko–Cantelli theorem, we have

$$\sup_y |F_n(y) - F(y)| \to 0$$

almost surely as $n \to \infty$. It can be seen that

$$\sup_{u,\epsilon} n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon) - \mathbb{P}(|Y - u| < \epsilon)$$

$$= \sup_{u,\epsilon} |F_n(u + \epsilon) - F_n(u - \epsilon) - F(u + \epsilon) + F(u - \epsilon)|$$

$$\leq \sup_{u,\epsilon} |F_n(u + \epsilon) - F(u + \epsilon)| + \sup_{u,\epsilon} |F_n(u - \epsilon) - F(u - \epsilon)|$$

$$\leq 2 \sup_{y} |F_n(y) - F(y)| \to 0$$

almost surely as $n \to \infty$. Note that

$$\sup_{u} n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon)$$

$$\leq \sup_{u} |n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon) - \mathbb{P}(|Y - u| < \epsilon)| + \sup_{u} \mathbb{P}(|Y - u| < \epsilon)$$

$$\leq \sup_{u,\epsilon} |n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon) - \mathbb{P}(|Y - u| < \epsilon)| + \sup_{u} \mathbb{P}(|Y - u| < \epsilon).$$

Because $\sup_{u} \mathbb{P}(|Y - u| < \epsilon) \leq 2M\epsilon$, we find

$$\sup_{u} n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon) < 2M\epsilon + \delta,$$

for any fixed $\delta > 0$ almost surely. This completes the proof. \square

The above lemma slightly extends the Glivenko–Cantelli theorem. For any interval of length $\epsilon$, the proportion of a random sample it contains is nearly uniformly bounded by $O(\epsilon)$.

2.3 Inequalities

In this subsection, we present some inequalities related to Gamma function and the finite Gamma mixture distributions. They are useful in the consistency proof in subsequent sections.
The magnitude of the Gamma function is well investigated in mathematics. The result developed by Li and Chen (2007) gives some useful bounds for the Gamma function. We restate it as the following lemma for convenient reference later.

**Lemma 3** When \( r > 1 \), we have

\[
\frac{r^{r-\gamma}}{e^{r-1}} < \Gamma(r) < \frac{r^{r-1/2}}{e^{r-1}},
\]

where \( \gamma \approx 0.577215 \cdots \) is the Euler–Mascheroni constant. When \( 0 < r < 1 \), the left inequality holds, but the right inequality is reversed.

The following lemma gives two upper bounds for the log Gamma density function given by (2), which are altered versions of two results from Chen et al. (2016). We will not repeat the settings.

**Lemma 4** (a) For any \( r > 0 \) and \( \theta > 0 \), the density function of \( Y \) satisfies

\[
\log g(y; r, \theta) \leq \gamma \log r
\]

where \( \gamma < 1 \) is the Euler–Mascheroni constant.

(b) Let \( \epsilon_r = \sqrt{2} \log r / \sqrt{r} \). When \( r > 20 \) and \( |y - \log (r\theta)| \geq \epsilon_r \), we have

\[
\log g(y; r, \theta) \leq \gamma (\log r - \log^2 r).
\]

**Proof** It can easily be seen that

\[
\log g(y; r, \theta) = -\log \Gamma(r) + r(y - \log \theta) - \exp(y - \log \theta).
\]

As a function of \( y \), the above function attains its maximum at \( y = \log (r\theta) \). Hence,

\[
\log g(y; r, \theta) \leq \log g(\log (r\theta); r, \theta) = -\log \Gamma(r) + r \log r - r.
\]

Recall Lemma 3, which gives a lower bound on the log Gamma function:

\[
\log \Gamma(r) \geq (r - \gamma) \log r - r + 1
\]

for any \( r > 0 \). Applying this bound to the upper bound of \( \log g(y; r, \theta) \), we have

\[
\log g(y; r, \theta) \leq \gamma \log r.
\]

This proves (6).

Next, we prove (7) through a side result. Let \( t_0 \) be any positive constant between 0 and 1. We prove that

\[
h(t) = \exp(t) - t - 1 \geq \frac{1}{3}t_0^2
\]
for $|t| \geq t_0$. By expanding $h(t)$ to the quadratic term, we can see that the above inequality is true for $t > 0$. Note that $h(t)$ decreases monotonically when $t \in (-\infty, 0)$. Hence, the inequality in (9) is true if $h(-t_0) \geq (1/3)t_0^2$ for any $t_0 \in (0, 1)$. Next, we show $h(-t) \geq (1/3)t^2$ for any $t \in (0, 1)$. For any $t \in (0, 1)$, applying Taylor’s expansion to $h(-t)$, we have

$$h(-t) - (1/3)t^2 = (1/6)(t^2 - \tilde{t}^2) > 0$$

where $\tilde{t} \in (0, t)$. This completes the proof that $h(t) \geq (1/3)t_0^2$.

We rearrange the terms in the log density function of the log Gamma distribution to get

$$\log g(y; r, \theta) = -\log \Gamma(r) + r \log r + r(y - \log (r\theta)) - r \exp (y - \log (r\theta))$$

Applying bound (8) to the log Gamma function again, we have

$$\log g(y; r, \theta) \leq \gamma \log r - 1 - rh(t) \leq \gamma \log r - 2 \frac{1}{3} \log^2 r$$

for all $r > 20$. This completes the proof of (7) and the proof of the lemma.

The two inequalities in this lemma give us information about the log density function. First, although this function is bounded in $y$, the upper bound can be arbitrarily large when $r$ is very large. Hence, when the parameter space for $r$ has a finite upper bound, the density functions in terms of $g(y; r, \theta)$ have a uniform upper bound. This trivially implies the consistency of the MLE of $G$ when a valid upper bound is placed on $r$. Second, the density function peaks at $y = \log (r\theta)$, but its value decays quickly at a quadratic rate in $\log r$ as $y$ diverges from this value.

### 3 Almost sure range of the MLE of the structural shape parameter $r$

In this section, we prove that the MLE of $r$ is almost surely in a compact interval as $n \to \infty$. This is a key preparation result toward the final consistency proof.

Suppose we have an IID sample $Y_1, Y_2, \ldots, Y_n$ from a finite log Gamma mixture distribution specified by (3), in which $r$ is structural and $G \in \mathcal{G}_m$ for specified $m$. The log likelihood function of the structural shape parameter $r$ and the mixing distribution $G$ is given by

$$\log g(Y_1, \ldots, Y_n; r, \mathcal{G})$$
\[ \ell_n(r, G) = \sum_{i=1}^{n} \log g(Y_i; r, G). \]  

(11)

The MLE of \((r, G)\) is some \((\hat{r}, \hat{G})\) such that

\[ \ell_n(\hat{r}, \hat{G}) = \sup\{\ell_n(r, G) : (r, G) \in \mathbb{R}^+ \times G_m\}. \]

(12)

We have implicitly assumed that the MLE is the maximum point of the likelihood among the mixing distributions with at most \(m\) distinct support points.

When the shape parameter \(r\) is confined in a compact finite interval, we can easily show that the constrained MLE is consistent. Hence, if feasible, a way to prove the consistency of the MLE is to show that \(\tau \leq \hat{r} \leq \Delta\) almost surely for some positive constants \(\tau\) and \(\Delta\), where \(\hat{r}\) is the MLE of the structural shape parameter. This strategy was used by Chen and Chen (2003) in the context of the finite Gaussian mixture model; for the finite Gamma mixture model the proof is more complicated. In this section and the next, we will:

(a) show that there exists a sufficiently small positive constant \(\tau > 0\) and a sufficiently large positive constant \(\Delta > 0\) such that \(\tau \leq \hat{r} \leq \Delta\) almost surely; and

(b) verify that the sufficient conditions presented in Chen (2017) are satisfied by the finite Gamma mixture model with a reduced parameter space \([\tau, \Delta] \times G_m\) for any \(0 < \tau < \Delta < \infty\).

Lemma 5

Assume that we have a set of IID observations \(Y_1, Y_2, \ldots, Y_n\) from the finite log Gamma mixture distribution defined by (3). Denote the true structural parameter value and the mixing distribution by \(r^*\) and \(G^* = \sum_{j=1}^{m^*} \alpha_j^* \{\theta_j^*\}\) with mixing proportions \(\alpha_j^* > 0\), distinct support points \(\theta_j^*\), and order \(m^*\). Assume \(m^* \leq m\) with a known positive integer \(m\). Let \((\hat{r}, \hat{G})\) be a global maximum point of the likelihood function \(\ell_n(r, G)\) over \(\mathbb{R}^+ \times G_m\).

There exist a sufficiently small constant \(\tau > 0\) and a sufficiently large constant \(\Delta\) such that as \(n \to \infty\), \(\{\tau \leq \hat{r} \leq \Delta\}\) almost surely.

Proof

We first show that \(\{\hat{r} \geq \tau\}\) almost surely for a sufficiently small positive constant \(\tau\). Recall that \(g(y; r, G)\) is the density function of \(Y = \log X\), which is a finite mixture of log Gamma distributions. By inequality (6) in Lemma 4, we have \(\sup_y \frac{g(y; r, G)}{\gamma \log r} \leq \gamma \log r\). Hence,

\[ \sup_{0 < r < \tau} \ell_n(r, G) < n\gamma \log \tau \]

(13)

for any positive and small constant \(\tau\).

In Lemma 1, we showed that \(E^*[\log g(Y; r^*, G^*)]\) is finite where \(E^*\) is the expectation when the distribution of \(Y\) is \(g(y; r^*, G^*)\). By the strong law of large numbers, we have

\[ \ell_n(r^*, G^*) = nE^*[\log g(Y; r^*, G^*)] + o(n) \]

(14)
almost surely. Note the remainder is \( o(n) \) rather than \( o_P(n) \). Hence, combining (13) with (14), we have

\[
\sup_{0 < r < \tau} \ell_n(r, G) - \ell_n(r^*, G^*) < n \left\{ \gamma \log \tau - \mathbb{E}^* \{ \log g(Y; r^*, G^*) \} + o(1) \right\}
\]

almost surely. Note that \( \gamma \log \tau \to -\infty \) as \( \tau \to 0^+ \). By Lemma 1, \( \mathbb{E}^* \{ \log g(Y; r^*, G^*) \} \) is a finite constant depending only on parameter value \( (r^*, G^*) \) of the true density function \( g(y; r^*, G^*) \). Therefore, when \( \tau \) is sufficiently small we must have \( \gamma \log \tau - \mathbb{E}^* \{ \log g(Y; r^*, G^*) \} \) far less than 0. Further, when \( \tau \) is sufficiently small, we have

\[
n \left\{ \gamma \log \tau - \mathbb{E}^* \{ \log g(Y; r^*, G^*) \} \right\} + o(n) < 0
\]

almost surely. Hence,

\[
\sup_{0 < r < \tau} \ell_n(r, G) - \ell_n(r^*, G^*) < 0
\]

almost surely. In other words,

\[
\ell_n(r, G) < \ell_n(r^*, G^*)
\]

uniformly for \( r < \tau \) almost surely. Based on this, we conclude that almost surely, the MLE of \( r \) will not be in the region \((0, \tau)\). This proves the first inequality of the lemma.

We now show that \( \bar{r} \leq \Delta \) almost surely for a sufficiently large constant \( \Delta \). We first observe that

\[
g(y; r, G) = \sum_{j=1}^{m} \alpha_j g(y; r, \theta_j) \leq \max_{1 \leq j \leq m} g(y; r, \theta_j).
\]

Recall (see (7)) that for any \( G \) and \( r \), when \( r > 20 \) and \( |y - \log (r \theta_j)| \geq \sqrt{2} \log r / \sqrt{r} \), we have

\[
\log g(y; r, \theta) \leq \gamma (\log r - \log^2 r).
\]

This upper bound is very small when \( r \) is large. For \( y \) such that \( |y - \log (r \theta_j)| \geq \sqrt{2} \log r / \sqrt{r} \) for all \( j \) in \( 1, 2, \ldots, m \), we have

\[
\log g(y; r, G) \leq \log \left\{ \max_{1 \leq j \leq m} g(y; r, \theta_j) \right\} \leq \gamma (\log r - \log^2 r).
\]

For convenience, let \( \mu_j = \log (r \theta_j) \) and \( \epsilon_r = \sqrt{2} \log r / \sqrt{r} \). For \( r > 20 \), applying (6) and (15), we have

\[
\log g(y; r, G) \leq \gamma \log r - \{ \gamma \log^2 r \} \left( \min_{1 \leq j \leq m} |y - \mu_j| \geq \epsilon_r \right)
\]
\[\ell_n(r, G) = \sum_{i=1}^{n} \log\{g(Y_i; r, G)\} \]

\[\leq n\gamma \log r - n\gamma \log^2 r + \gamma \log^2 r \sum_{j=1}^{m} \left\{ \sup_{\mu_j} \sum_{i=1}^{n} \mathbb{I}(|Y_i - \mu_j| < \epsilon_r) \right\} \]

\[\leq n\gamma \{\log r - \log^2 r + m\gamma \log^2 r \sup_{u} \left\{ n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon_r) \right\} \}. \quad (16)\]

By the upper bound in Lemma 2, we have

\[\sup_{u} \left\{ n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon_r) \right\} < 2M^*\epsilon_r + \delta \]

almost surely for any fixed \(\delta > 0\), where \(M^*\) is the upper bound on the true density function \(g(y; r^*, G^*)\). Obviously, \(M^*\) is finite. Furthermore, by choosing a sufficiently large \(\Delta\), we can ensure that \(2M^*\epsilon_r < \delta\) uniformly over \(r > \Delta\). Hence,

\[\sup_{u} \left\{ n^{-1} \sum_{i=1}^{n} \mathbb{I}(|Y_i - u| < \epsilon_r) \right\} < 2\delta \quad (17)\]

almost surely for any \(\delta > 0\). By choosing a sufficiently small \(\delta \in (0, 1/(2M^*))\) and the corresponding \(\Delta\) and then applying (17) for (16), we have that uniformly over \(r > \Delta\),

\[\ell_n(r, G) < n\gamma \{\log r - (1 - 2m\delta) \log^2 r\} < n[\mathbb{E}^*[\log g(Y; r^*, G^*)] - 1]. \quad (18)\]

almost surely. Recalling (14), we have

\[\ell_n(r^*, G^*) = n[\mathbb{E}^*[\log g(Y; r^*, G^*)] + o(1)]\]

almost surely. Therefore, applying above formula for (18), we obtain

\[\sup_{r > \Delta} \ell_n(r, G) < \ell_n(r^*, G^*)\]
almost surely. This is the second inequality of the lemma. □

4 Consistency of the MLE after the domain of the structural shape parameter is reduced

In the last section, we have accomplished Task (a): the MLE of the structural shape parameter is almost surely in a finite interval \([\tau, \Delta] \subset \mathbb{R}^+\). The consistency problem for the finite Gamma mixture model with structural shape parameter \(r\) has therefore been reduced to the problem where the parameter space of \((r, G)\) is \([\tau, \Delta] \times \mathbb{G}_m\). In this section, we show the MLE of \((r, G)\), constrained on the parameter space \([\tau, \Delta] \times \mathbb{G}_m\), is consistent by verifying the consistency conditions presented in Chen (2017). When this is done, it further implies the MLE of \((r, G)\) defined on the whole parameter space \(\mathbb{R}^+ \times \mathbb{G}_m\) is consistent.

To conveniently discuss the consistency of the MLE \((\hat{r}, \hat{G})\), we introduce a distance \(D_{KW}(\cdot, \cdot)\), used in Kiefer and Wolfowitz (1956), on the space of \((r, G)\). For any shape parameter values \(r_1, r_2 \in \mathbb{R}^+\) and any mixing distributions \(G_1, G_2 \in \mathbb{G}_m\), define

\[
D_{KW}((r_1, G_1), (r_2, G_2)) = |\arctan(r_1) - \arctan(r_2)| + \int_{\mathbb{R}^+} |G_1(\theta) - G_2(\theta)| \exp(-\theta) \, d\theta. \tag{19}
\]

An important property is that \(D_{KW}((r_k, G_k), (r^*, G^*)) \to 0\) if and only if \(r_k \to r^*\) and \(G_k \to G^*\) in distribution/measure. Hence, the consistency of the MLE \((\hat{r}, \hat{G})\) can be conveniently interpreted as \(D_{KW}((\hat{r}, \hat{G}), (r^*, G^*)) \to 0\) almost surely as the sample size \(n \to \infty\). We now state our main result.

Theorem 1 Assume that we have a set of IID observations \(Y_1, Y_2, \ldots, Y_n\) from (3) with mixing distribution \(G^* = \sum_{j=1}^{m^*} \alpha_j^* \{\theta_j^*\}\), structural shape parameter value \(r^*\), and order \(m^*\). We assume \(m^* \leq m\) for some known \(m\).

Then the MLE of \((r, G)\) defined in (12), \((\hat{r}, \hat{G})\), is a consistent estimator. That is, almost surely as \(n \to \infty\)

\[
D_{KW}((\hat{r}, \hat{G}), (r^*, G^*)) \to 0.
\]

The proof of Theorem 1 will be presented in the next two subsections.

4.1 Four conditions

This subsection presents the four conditions specified in Chen (2017) under which the consistency of MLE holds. The verification will be given in the next subsection.

There is a rich literature on the consistency of the MLE under mixture models. According to Chen (2017), there are three popular ways to establish the consistency of the MLE: the approaches of Kiefer and Wolfowitz (1956), Redner (1981), and Pfanzagl (1988). They give similar but not completely equivalent results. Our proof follows the approach of Kiefer and Wolfowitz (1956).
To avoid confusion in notation between the specific finite Gamma mixture and the general mixture model, we used $f(x; \psi)$ for the density function of the component distribution for the general mixture. A general mixture model has the density function

$$f(x; G) = \int_{\psi} f(x; \psi) dG(\psi)$$

(20)

for $G \in \mathcal{G}$. Note that the finite mixture model is a special case where $\mathcal{G}$ is reduced to $\mathcal{G}_m$. The parameter space of $\psi$ is $\Psi \subset \mathbb{R}^d$ for some positive integer $d$.

Kiefer and Wolfowitz (1956) observed that under some conditions, the consistency of the MLE under a mixture model based on IID observations is reduced to the general consistency problem discussed in Wald (1949). The conditions for consistency given by Kiefer and Wolfowitz (1956) are detailed but hard to comprehend. Chen (2017) streamlined their conditions and replaced them by the following four high-level conditions applicable to (20):

A1 Identifiability: Let $F(x; G)$ be the cumulative distribution function of $f(x; G)$.

The mixture model is identifiable, i.e.,

$$F(x; G_1) = F(x; G_2)$$

for all $x$ implies $G_1 = G_2$.

A2 Finite Kullback–Leibler Information: Let the true mixing distribution be $G^*$ and for any subset $B$ of the space of mixing distributions, define

$$f(x; B) = \sup_{G \in B} f(x; G).$$

Let $B_\epsilon(G) = \{G' : D_{KW}(G, G') < \epsilon\}$ be an open ball of radius $\epsilon$ centered at $G$. For any $G \neq G^*$, there exists an $\epsilon > 0$ such that

$$\mathbb{E}^*[\log \{f(X; B_\epsilon(G))/f(X; G^*)\}]^+ < \infty.$$

The expectation operator $\mathbb{E}^* (\cdot)$ is taken under $f(x; G^*)$.

A3 Continuity: The component parameter space $\Psi$ is a closed set. For all $x$ and any given $G_0$, we have

$$\lim_{G \to G_0} f(x; G) = f(x; G_0).$$

A4 Compactness: The definition of the mixture density $f(x; G)$ in $\mathcal{G}$ can be continuously extended to a compact space $\bar{\mathcal{G}}$ while retaining the validity of Condition A2.

The above sufficient conditions are applicable to the nonparametric MLE $\hat{G}_n$ for general mixture models. They are equally applicable to finite mixture models when $\mathcal{G}$ is reduced to $\mathcal{G}_m$ and $G^* \in \mathcal{G}_m$. If we regard the shape parameter $r$ as part of the mixing distribution, then the mixing distribution degenerates in this margin. However,
4.2 Conditions A1–A4 on the reduced parameter space

We have already shown that $\hat{r}$ is almost surely part of $[\tau, \Delta] \subset \mathbb{R}^+$. In this subsection, we show that Conditions A1–A4 are satisfied under the finite Gamma mixture model with structural shape parameter $r \in [\tau, \Delta]$.

To begin with, we verify Condition A1: the finite Gamma mixture model with structural shape parameter $r$ is identifiable. This model is a special case of the general Gamma mixture model: its bivariate mixing distribution degenerates in its structural element $r$. Teicher (1963) gave a set of sufficient conditions for the identifiability of finite mixture models. The general finite mixture of the two-parameter Gamma model satisfies these conditions. Therefore, the general finite mixture of the two-parameter Gamma model is identifiable. This further implies the identifiability of the finite Gamma (or log Gamma) mixture model with a structural shape parameter and hence Condition A1 is verified.

Next, we verify Conditions A2–A4. Note that Conditions A2 and A3 are part of Condition A4. When Condition A4 is verified, Conditions A2 and A3 are implied. For this reason, we start with Condition A4. We will:

(c) show that the definition of the mixture density $g(y; r, G)$ in $[\tau, \Delta] \times \mathbb{G}_m$ can be continuously extended to a compact space $[\tau, \Delta] \times \bar{\mathbb{G}}_m$;

(d) verify that Condition A2 is also satisfied when $(r, G) \in [\tau, \Delta] \times \bar{\mathbb{G}}_m$.

We now verify (c). The first task is to extend the component parameter space of $\theta$ from $\mathbb{R}^+$ to its closure $[0, \infty]$. For any given $r \in [\tau, \Delta]$, it can easily be seen that

$$\lim_{\theta \to 0^+} g(y; r, \theta) = \lim_{\theta \to \infty} g(y; r, \theta) = 0$$

for all $y$. Therefore, for any given $r \in [\tau, \Delta]$, we have a continuity preserving extension of the subpopulation parameter space of $\theta$ to $[0, \infty]$:

$$g(y; r, 0) = g(y; r, \infty) = 0$$

for all $y \in \mathbb{R}$. The density function $g(y; r, \theta)$ remains continuous in $r$ and $\theta$ on $[\tau, \Delta] \times [0, \infty]$.

Next, we define

$$\bar{\mathbb{G}}_m = \{ \rho G : G \in \mathbb{G}_m, \rho \in [0, 1] \}.$$ 

For $\bar{G} \in \bar{\mathbb{G}}_m$, we have $\rho = 1$. For any $G_1, G_2 \in \mathbb{G}_m$, define a distance $D_{K\text{W}}'(\cdot, \cdot)$ on $\mathbb{G}_m$:

$$D_{K\text{W}}'(G_1, G_2) = \int_{\mathbb{R}^+} |G_1(\theta) - G_2(\theta)| \exp(-\theta) d\theta,$$
which is the second term of $D_{KW}(\cdot,\cdot)$ defined by (19). According to Helly’s selection theorem as explained in van der Vaart (2000), each sequence of probability measures \( \{G_k \in G_m : k = 1, 2, \ldots \} \) has a converging subsequence \( \{G_k'\} \) such that 
\[
D'_{KW}(G_k', \bar{G}) \to 0 \quad \text{with some} \quad \bar{G} \in \bar{G}_m.
\]
Equipped with the distance $D'_{KW}(\cdot,\cdot)$, $\bar{G}_m$ is compact. Note that

\[
D_{KW}((r_1, G_1), (r_2, G_2)) \leq \pi + \int_{\mathbb{R}^+} \exp(-|\theta|) \, d\theta = \pi + 1
\]

for any \((r_1, G_1)\) and \((r_2, G_2)\) in $[\tau, \Delta] \times \bar{G}_m$. Then $D_{KW}(\cdot,\cdot)$ is a bounded distance on $[\tau, \Delta] \times \bar{G}_m$. Hence, $[\tau, \Delta] \times \bar{G}_m$ is compact with respect to $D_{KW}(\cdot,\cdot)$. We remark here that the completeness is implied by Helly’s theorem.

For any \(r \in [\tau, \Delta]\) and \(\bar{G} \in \bar{G}_m\), we further define

\[
g(y; r, \bar{G}) = \int_0^\infty g(y; r, \theta) \, d\bar{G} = \rho \int_0^\infty g(y; r, \theta) \, d\bar{G} = \rho g(y; r, G).
\]

When $\bar{G} \in G_m$, we have $g(y; r, \bar{G}) = g(y; r, G)$. This extends the domain of density function $g(y; r, G)$ from $[\tau, \Delta] \times G_m$ to the compact space $[\tau, \Delta] \times \bar{G}_m$.

Next we show that the density function $g(y; r, \bar{G})$ is continuous on the compact space $[\tau, \Delta] \times \bar{G}_m$ under $D_{KW}(\cdot,\cdot)$. That is, for any sequence \((r_k, G_k) : (r_k, G_k) \in [\tau, \Delta] \times \bar{G}_m, k = 1, 2, \ldots\) and \((\bar{r}, \bar{G}) \in [\tau, \Delta] \times \bar{G}_m\) such that $D_{KW}((r_k, G_k), (\bar{r}, \bar{G})) \to 0$, we have $g(y; r_k, G_k) \to g(y; \bar{r}, \bar{G})$ for each $y \in \mathbb{R}$. Since $D_{KW}((r_k, G_k), (\bar{r}, \bar{G})) \to 0$ if and only if $r_k \to \bar{r}$ and $G_k \to \bar{G}$, our task is to show that $r_k \to \bar{r}$ and $G_k \to \bar{G}$ imply $g(y; r_k, G_k) \to g(y; \bar{r}, \bar{G})$ for each $y \in \mathbb{R}$.

The function $g(y; r, \bar{G})$ is continuous with respect to $r \in [\tau, \Delta]$ for any given $y \in \mathbb{R}, \bar{G} \in G_m$. Hence, the task is reduced to show that $g(y; r, G_k) \to g(y; r, \bar{G})$ when $G_k \to \bar{G}$. Given such $y$ and $r$, $g(y; r, \theta)$ has a finite upper bound by (2).

One equivalent definition of convergence in measure (van der Vaart (2000)) is that $G_k \to \bar{G}$ in distribution if and only if $\int h(\theta) \, dG_k(\theta) \to \int h(\theta) \, d\bar{G}(\theta)$ for all bounded and continuous functions $h(\cdot)$. Using this definition, we find

\[
g(y; r, G_k) = \int_0^\infty g(y; r, \theta) \, dG_k \to \int_0^\infty g(y; r, \theta) \, d\bar{G} = g(y; r, \bar{G})
\]

as $G_k \to \bar{G}$. Therefore, the density function $g(y; r, \bar{G})$ is continuous in compact space $[\tau, \Delta] \times \bar{G}_m$ under $D_{KW}(\cdot,\cdot)$. This verifies Condition A3 after $[\tau, \Delta] \times \bar{G}_m$ is compactified and completes Task (c).

We now work on Task (d): verifying that Condition A2 is also satisfied on the space of $(r, G) \in [\tau, \Delta] \times \bar{G}_m$. By (6), we have $\log g(y; r, G) \leq y \log \Delta$ when $r \in [\tau, \Delta]$. Hence, for any $(r, G) \in [\tau, \Delta] \times \bar{G}_m$ and any constant $\epsilon > 0$, let

\[
B_\epsilon(r, G) = \{(r', G') : D_{KW}((r, G), (r', G')) < \epsilon\}
\]
be the open ball of radius \( \epsilon \) centered at \((r, G)\). Clearly,

\[
   g(y; B_\epsilon(r, G)) = \sup_{(r', G') \in B_\epsilon(r, G)} g(y; r', G') \leq \Delta^y.
\]

Therefore, we have

\[
   \mathbb{E}^*[\log g(Y; B_\epsilon(r, G))] \leq \gamma \log \Delta,
\]

where the expectation operator \( \mathbb{E}^*(\cdot) \) is taken under \( g(y; r^*, G^*) \). Hence, we have

\[
   \mathbb{E}^* \log\{g(Y; B_\epsilon(r, G))/g(Y; r^*, G^*)\} \leq \gamma \log \Delta - \mathbb{E}^* \log\{g(Y; r^*, G^*)\}. \tag{21}
\]

Note \( \Delta \) in (21) is a sufficiently large constant chosen before and \( \mathbb{E}^*\{\log g(Y; r^*, G^*)\} \) is also a finite constant only depending on the true parameter value \((r^*, G^*)\) of true density function \( g(y; r^*, G^*) \) based on Lemma 1. Therefore, the upper bound in (21) is finite and we have

\[
   \mathbb{E}^* \log\{g(Y; B_\epsilon(r, G))/g(Y; r^*, G^*)\} < \infty.
\]

Therefore, we have

\[
   \mathbb{E}^*[\log\{g(Y; B_\epsilon(r, G))/g(Y; r^*, G^*)\}]^+ < \infty,
\]

for all \((r, G) \in [\tau, \Delta] \times \bar{G}_m\). Thus, Condition A2 is satisfied on the compactified space \([\tau, \Delta] \times \bar{G}_m\). This completes Task (d).

We have shown Conditions A1–A4 are satisfied. Hence, the MLE over the space of \( r \in [\tau, \Delta] \) and \( G \in G_m \) is strongly consistent. Combined with conclusions of Lemma 5, we have proved Theorem 1.

One potential concern is that \((\hat{r}, \hat{G})\) is the maximum point of \( \ell_n(r, G) \) over the compactified space \( \mathbb{R}^+ \times \bar{G}_m \). It appears that \( \hat{G} \) could be a subdistribution with a total probability below 1, i.e., \( \hat{G} = \rho \check{G} \) for some \( \rho \in (0, 1) \) and \( \check{G} \in G_m \). This is impossible because

\[
   \ell_n(\hat{r}, \check{G}) = \ell_n(\hat{r}, \rho \check{G}) < \ell_n(\hat{r}, \hat{G}).
\]

Therefore, such a \((\hat{r}, \hat{G})\) is not an MLE. We needed subdistributions only for the compact property in the proof.

5 Numerical computations and simulation experiments

5.1 EM algorithm

We provide specific details of the EM-algorithm for fitting finite Gamma mixtures based on IID observations. The EM algorithm is the most popular numerical approach
for finding the MLE of the mixing distribution. The algorithm is iterative: it updates
the initial mixing distribution $G^{(0)}$ proposed by the user to obtain $G^{(k)}: k = 1, 2, \ldots$
It is well known that $\ell_n(G^{(k)})$ is monotonic, which leads to the convergence property.
The properties of the EM algorithm for finite mixtures have been thoroughly discussed
in Wu (1983). We follow Chen et al. (2016) and use a slightly adapted EM algorithm
in our simulation experiments.

In a finite mixture model, each observation $x_i$ may be regarded as part of the com-
plete observations $(x_i, z_i)$ on the $i$th sample unit. In this setup, $z_i$ is an unknown latent
variable; given $z_i = j$, $x_i$ is a sample from the $j$th subpopulation, $j = 1, 2, \ldots, m$.
Hence, the complete-data log-likelihood for the finite Gamma mixture model with a
structural $r$ is given by

$$
\sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{I}(z_i = j) \{ \log \alpha_j + \log f(x_i; r, \theta_j) \}.
$$

To improve the finite-sample performance, one may modify the log-likelihood function
by adding an $O_p(1)$ term without altering the consistency conclusion. We use

$$
\ell_c(r, G) = \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{I}(z_i = j) \{ \log \alpha_j + \log f(x_i; r, \theta_j) \} + \epsilon \sum_{j=1}^{m} \log \alpha_j
$$

for some $\epsilon > 0$. In our simulation and real-data experiments, we chose $\epsilon = 0.001$.
This strategy was first employed in the modified likelihood approach of Chen (1998)
and has been widely adopted, e.g., Chen et al. (2016). Since the $z_i$’s are missing, one
cannot estimate $(r, G)$ by the maximum point of $\ell_c(r, G)$. To overcome this obstacle,
the EM algorithm replaces $\mathbb{I}(z_i = j)$ with its expected value. We now discuss the two
steps of the algorithm.

**E-step.** Given a shape parameter $r^{(0)}$ and a mixing distribution $G^{(0)} \in \mathbb{G}_m$, we find
the conditional expectation of $\mathbb{I}(z_i = j)$ given the data:

$$
w_{ij}^{(0)} = \mathbb{E}_n^{(0)}(\mathbb{I}(z_i = j)|x_1, \ldots, x_n) = \frac{\alpha_j^{(0)} f(x_i; r^{(0)}, \theta_j^{(0)})}{\sum_{k=1}^{m} \alpha_k^{(0)} f(x_i; r^{(0)}, \theta_k^{(0)})}.
$$

Replacing $\mathbb{I}(z_i = j)$ with its conditional expectation given above, we obtain

$$
Q(r, G; r^{(0)}, G^{(0)}) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_{ij}^{(0)} \{ \log \alpha_j + \log f(x_i; r, \theta_j) \} + \epsilon \sum_{j=1}^{m} \log \alpha_j.
$$

**M-step.** In this step, we maximize $Q(r, G; r^{(0)}, G^{(0)})$ with respect to $r \in \mathbb{R}^+$ and
$G \in \mathbb{G}_m$. For $j = 1, \ldots, m$, denote

$$
\hat{w}_j^{(0)} = n^{-1} \sum_{i=1}^{n} w_{ij}^{(0)},
$$

\[ Springer \]
\[ x_j^{(0)} = (n \bar{w}_j^{(0)})^{-1} \sum_{i=1}^{n} w_{ij}^{(0)} x_i, \quad \text{and} \]
\[ y_j^{(0)} = (n \bar{w}_j^{(0)})^{-1} \sum_{i=1}^{n} w_{ij}^{(0)} \log x_i. \]

We then get the expression
\[ Q(r, G; r^{(0)}, G^{(0)}) = n \sum_{j=1}^{m} \left\{ \bar{w}_j^{(0)} \left[ (r-1) \bar{y}_j^{(0)} - (\bar{x}_j^{(0)}/\theta_j) - \log \Gamma(r) - r \log \theta_j \right] \right\} + \sum_{j=1}^{m} \left( (n \bar{w}_j^{(0)} + \epsilon) \log \alpha_j \right). \] (22)

Note that the component parameters in \( Q(r, G; r^{(0)}, G^{(0)}) \) are well separated. Maximizing \( Q(r, G; r^{(0)}, G^{(0)}) \) with respect to \( \alpha_j \) gives
\[ \alpha_j^{(1)} = (n \bar{w}_j^{(0)} + \epsilon)/(n + m \epsilon). \]

The extra positive constant \( \epsilon \) makes the above iteration step numerically stable.

For each fixed \( r \), \( Q(r, G; r^{(0)}, G^{(0)}) \) is maximized with respect to \( \theta_j \) when
\[ \theta_j^{(1)} = \bar{x}_j^{(0)}/r. \]

Replacing \( \theta_j \) with \( \theta_j^{(1)} \) in \( Q(r, G; r^{(0)}, G^{(0)}) \), we find that the maximization solution in \( r \) is given by
\[ r^{(1)} = \arg \max_r \left\{ \sum_{j=1}^{m} \left[ \bar{w}_j^{(0)} \left( \bar{y}_j^{(0)} - \log \bar{x}_j^{(0)} \right) \right] r + \left[ r \log r - \log \Gamma(r) - r \right] \right\}. \]

It is a simple task to optimize a single-variable function. Once \( r^{(1)} \) is obtained, the updated mixing distribution is given by
\[ G^{(1)} = \sum_{j=1}^{m} \alpha_j^{(1)} \{ \theta_j^{(1)} \}. \]

Starting from the initial value \( (r^{(0)}, G^{(0)}) \), the E- and M-steps give us \( (r^{(1)}, G^{(1)}) \). Repeating these two steps, we get a sequence \( (r^{(k)}, G^{(k)}) \), \( k = 1, 2, \ldots \). The slightly modified log likelihood function has its value increased after each iteration. We terminate the algorithm when the modified log likelihood value stabilizes; in the simulations, we set the tolerance to \( 10^{-6} \).
5.2 Simulation experiments

We specify the settings of simulation experiments and present the findings in this subsection.

We conducted simulation experiments to illustrate the consistency of the MLE under a finite Gamma mixture model with structural $r$. We generated data from the six Gamma mixture distributions specified in the following table. Models I and II each contain three mixtures of orders $m = 2$ and $m = 3$. We selected distinct subpopulation scale parameter values for both models.

| Model | Density function | $r$         |
|-------|------------------|-------------|
| I     | $0.4 f(x; r, 0.5) + 0.6 f(x; r, 5)$ | 0.5, 5, 50 |
| II    | $0.35 f(x; r, 0.5) + 0.55 f(x; r, 2) + 0.1 f(x; r, 6)$ | 0.5, 10, 30 |

We generated samples of sizes 60, 240, 960, and 3,840 from distributions in Models I and II. For each random sample, we used 30 sets of initial values to drive the EM algorithm. The first 11 sets contained the true values of the model that generated the sample and 10 randomly and mildly perturbed true values. The remaining 19 sets were randomly generated, and they could be quite different from the true value. With these 30 initial values, up to 30 local maxima were found for each random sample; we took the MLE to be the one with the highest $\ell_n$ value. We performed $K = 1,000$ repetitions for each mixture distribution and sample-size combination.

We found the root mean square error (RMSE) for each parameter in these six mixtures. Let $\hat{\psi}$ be a generic parameter estimator and $\psi^*$ be the true value of the corresponding parameter in the selected distribution. The root mean square error (RMSE) of $\psi$ is

\[
\text{RMSE}(\psi) = \left\{ K^{-1} \sum_{k=1}^{K} (\hat{\psi}^{(k)} - \psi^*)^2 \right\}^{1/2}.
\]

In this expression, we used the superscript $(k)$ for the estimate based on the $k$th sample, and $\psi$ is a generic notation for either the mixing proportions $\alpha_j$, the subpopulation scale $\theta_j$, or the structural shape parameter $r$. We used R-CRAN (Core Team (2020)) for the simulation and the results are presented in Tables 1 and 2.

We kept a record of the initial values (perturbed or random) that led to the highest modified likelihood values. Let $K_0$ be the number of times in the 1,000 repetitions that the highest values were obtained from the perturbed values. We computed $\eta = K_0/K$ and reported it in the last column of Tables 1 and 2. The simulation results led to the following two observations:

1. As shown in the first column of the tables, for each $r$, we increased the sample size in multiples of four from 60 to 3,840. The increase often halved the RMSE of a parameter. Overall, the RMSEs decreased markedly as the sample size increased. These observations support the theoretical consistency conclusion.
Table 1  RMSEs of the MLE based on data generated from distributions in Model I

| n  | $\alpha_1$ | $\alpha_2$ | $r$  | $\theta_1$ | $\theta_2$ | $\eta$ |
|----|------------|------------|------|------------|------------|-------|
| $r = 0.5$ | 60 | 0.210 | 0.210 | 0.153 | 0.511 | 3.210 | 0.361 |
| 60 | 0.113 | 0.113 | 0.063 | 0.321 | 1.296 | 0.413 |
| 240 | 0.050 | 0.050 | 0.028 | 0.120 | 0.455 | 0.431 |
| 960 | 0.025 | 0.025 | 0.014 | 0.059 | 0.216 | 0.453 |
| 3,840 | 0.067 | 0.067 | 1.083 | 0.104 | 0.982 | 0.822 |
| 240 | 0.016 | 0.016 | 0.228 | 0.026 | 0.240 | 0.792 |
| 960 | 0.008 | 0.008 | 0.116 | 0.013 | 0.123 | 0.865 |
| 3,840 | 0.0016 | 0.0016 | 1.164 | 0.012 | 0.116 | 1.000 |

Table 2  RMSEs of the MLE based on data generated from distributions in Model II

| n  | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $r$  | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\eta$ |
|----|------------|------------|------------|------|------------|------------|------------|-------|
| $r = 0.5$ | 60 | 0.251 | 0.285 | 0.384 | 0.387 | 0.444 | 2.742 | 5.188 | 0.288 |
| 60 | 0.246 | 0.210 | 0.272 | 0.121 | 0.412 | 1.405 | 5.195 | 0.405 |
| 240 | 0.209 | 0.134 | 0.186 | 0.042 | 0.316 | 0.875 | 3.785 | 0.237 |
| 960 | 0.123 | 0.080 | 0.088 | 0.016 | 0.169 | 0.502 | 2.211 | 0.578 |
| 3,840 | 0.065 | 0.070 | 0.046 | 2.670 | 0.118 | 0.465 | 1.694 | 0.848 |
| 240 | 0.031 | 0.035 | 0.022 | 1.106 | 0.056 | 0.225 | 0.778 | 0.874 |
| 960 | 0.017 | 0.017 | 0.011 | 0.519 | 0.027 | 0.108 | 0.377 | 0.910 |
| 3,840 | 0.008 | 0.009 | 0.005 | 0.261 | 0.014 | 0.055 | 0.193 | 0.890 |
| $r = 10$ | 60 | 0.060 | 0.062 | 0.040 | 6.708 | 0.093 | 0.369 | 1.211 | 0.910 |
| 240 | 0.030 | 0.032 | 0.019 | 2.885 | 0.047 | 0.186 | 0.599 | 0.899 |
| 960 | 0.016 | 0.016 | 0.010 | 1.363 | 0.023 | 0.091 | 0.288 | 0.891 |
| 3,840 | 0.007 | 0.008 | 0.005 | 0.693 | 0.012 | 0.046 | 0.147 | 0.890 |

There is one exception: $r = 0.5$ from Model II. Table 2 shows that when the sample size increases from 60 to 240, the RMSE for $\theta_3$ increases slightly. This unexpected outcome may be attributed to the slow action of the asymptotic when $r < 1$ and to the random nature of the simulation experiment. Note that in this case, the density function goes to infinity when $x$ approaches 0.

2. When $r = 0.5$, $\eta$ is usually below 50%, but it increases with the sample size. This also indicates that in this case the asymptotic requires a large sample. If $r$ is large, the MLE is near the true parameter value, so it is usually located when the EM algorithm starts from the perturbed values.
Table 3  Summary statistics for household income (in thousand yuan)

| Minimum | 25th Percentile | Median | 75th Percentile | Maximum |
|---------|----------------|--------|-----------------|---------|
| 0.0407  | 24.4149        | 42.4569| 70.7991         | 1958.9434|

| Sample size | Mean   | SD     | Skewness | Kurtosis |
|-------------|--------|--------|----------|----------|
| 17,159      | 55.7077| 53.5846| 7.7611   | 187.3505 |

6 Data example: disposable income

Finite Gamma mixture distributions are often used to model, for example, insurance payments, household incomes, and the cost of medicine: see Liu et al. (2003), Wong and Li (2014), Willmot and Lin (2011), and Yin et al. (2019). In this section, we illustrate the use of the finite Gamma mixture distribution with a structural parameter for data on disposable income and expenditure. We obtained the data from the China Institute for Income Distribution (CHIP13 2016). They were collected in the fifth-wave survey in July and August 2014. The CHIP13 data contain many attributes; we analyzed only the household income and expenditures.

The data set contains 17,244 records for disposable income, but 85 of them are either missing or nonpositive. We excluded these records and fitted a finite mixture model to the remaining 17,159 observations. Table 3 gives summary statistics for household income, and we make three remarks below:

1. The maximum income is about 35 times the mean income, and the mean is much larger than the median. Both features reflect the uneven wealth distribution.
2. Slightly over 10% of the households have an income that is 50% above the 75th percentile. The data are seriously skewed to the right.
3. The majority of households (over 98%) have an annual income below 196 thousand yuans. For the top 2%, the range is from 195 to 959 thousand yuans.

These characteristics suggest that the population can be segregated into several homogeneous subpopulations. A finite Gamma mixture distribution with a structural shape parameter is a reasonable choice.

We fitted a set of finite Gamma mixture models with a structural shape parameter \( r \) and order \( m = 1, 2, \ldots, 7 \). When \( m \) was greater than 2, we first used 50 random initial values to drive the EM algorithm and obtained up to 50 local maxima of the likelihood function. We took the local maximum with the highest likelihood value as the tentative MLE. Next, we created 10 initial values by perturbing the tentative MLE and generated a further 19 randomly. When the EM iteration stopped, we selected the estimate with the highest \( \ell_n \) value as the MLE. For numerical stability, we adopted the modified likelihood with \( \epsilon = 0.001 \). Table 4 gives the MLEs and corresponding likelihood values. By the nature of the maximum likelihood, \( \ell_n(\hat{G}) \) increases as the order \( m \) of the mixture models increases. The size of the increment stalls at \( m = 7 \).

Figure 2 gives a QQ-plot for the fitted Gamma mixture models (log transformed), together with the 45-degree line. The data points nearly perfectly align with the straight line when \( m = 3 \) and 4. There is practically no room for further improvement by increasing \( m \).
Fig. 1 Histogram of log disposable income and density function of fitted log Gamma mixtures with order $m = 3$ and $m = 4$

Table 4 MLEs of Gamma mixtures for disposable income

| $m$ | 1    | 2    | 3    | 4    | 5    | 6    | 7    |
|-----|------|------|------|------|------|------|------|
| $\hat{r}$ | 1.6945 | 2.0261 | 2.1841 | 2.2433 | 2.3293 | 2.3526 | 2.4391 |
| $\hat{\alpha}_1$ | 0.9129 | 0.7436 | 0.6009 | 0.0008 | 0.0002 | 0.0002 |
| $\hat{\alpha}_2$ | 0.0871 | 0.2531 | 0.3740 | 0.5486 | 0.0008 | 0.0009 |
| $\hat{\alpha}_3$ | 0.0033 | 0.0247 | 0.4214 | 0.5363 | 0.0094 |
| $\hat{\alpha}_4$ | 0.0004 | 0.0288 | 0.4323 | 0.5311 |
| $\hat{\alpha}_5$ | 0.0004 | 0.0300 | 0.4280 |
| $\hat{\alpha}_6$ | 0.0004 | 0.0300 | 0.4280 |
| $\hat{\alpha}_7$ | 0.0004 |
| $\hat{\theta}_1$ | 32.876 | 23.744 | 19.290 | 17.189 | 0.2408 | 0.0526 | 0.0509 |
| $\hat{\theta}_2$ | 66.805 | 41.620 | 33.102 | 15.714 | 0.4046 | 0.4037 |
| $\hat{\theta}_3$ | 190.82 | 78.326 | 30.811 | 15.353 | 5.0650 |
| $\hat{\theta}_4$ | 484.69 | 73.740 | 30.271 | 14.804 |
| $\hat{\theta}_5$ | 466.97 | 72.579 | 29.496 |
| $\hat{\theta}_6$ | 462.15 | 70.756 |
| $\hat{\theta}_7$ | 449.30 |
| $\ell_n$ | −84,911 | −84,542 | −84,487 | −84,478 | −84,468 | −84,467 | −84,467 |
With a structural shape parameter \( r \), the number of parameters in the finite Gamma mixture model does not increase as quickly with \( m \). Although the mixture model is not regular, the Bayes information criterion still provides some guidance. For the current sample size, a higher order of mixture model would be recommended when the log-likelihood increases by 40. With this general guidance, a finite Gamma mixture of order \( m = 3 \) is recommended; \( m = 4 \) is also acceptable. Figure 1 shows a histogram of the log disposable income data and the density function of the fitted Gamma mixture model of order \( m = 3 \) and \( m = 4 \).

The fitted finite Gamma mixture distribution of order \( m = 3 \) suggests that about 75% of the households had a low mean annual disposable income of 19,000 yuan. A small percentage of household (0.3%) had 10 times this value. Setting \( m = 4 \) changes the picture of the low- and medium-income households only slightly. However, it separates a much smaller percentage (0.04%) of super-rich households. They had nearly 30 times the mean income of the low-income households.
7 Discussion and observations

The finite Gamma mixture distributions are useful for modeling positive data that is suspected to come from a heterogeneous population. However, the MLE of the general Gamma mixture model is inconsistent. We have shown that the MLE of the finite Gamma mixture model with a structural shape parameter is strongly consistent. The simulation results indicate that the MLE has respectable finite-sample properties and an observable consistency trend. The real-data example demonstrated that this model can reveal potential subpopulation structures.

Acknowledgements We thank the China Institute for Income Distribution for the income data. We thank the referee for the helpful comments.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Chen J (1998) Penalized likelihood-ratio test for finite mixture models with multinomial observations. Can J Stat 26(4):583–599
Chen J (2017) Consistency of the MLE under mixture models. Stat Sci 32(1):47–63
Chen H, Chen J (2003) Tests for homogeneity in normal mixtures in the presence of a structural parameter. Stat Sin 13:351–365
Chen J, Tan X (2009) Inference for multivariate normal mixtures. J Multivar Anal 100(7):1367–1383. https://doi.org/10.1016/j.jmva.2008.12.005
Chen J, Tan X, Zhang R (2008) Inference for normal mixtures in mean and variance. Stat Sin 18(2):443–465
Chen J, Li S, Tan X (2016) Consistency of the penalized MLE for two-parameter gamma mixture models. Sci China Math 59(12):2301–2318
CHIP13 (2016) Chinese household income and expenditure project. http://www.ciidbnu.org/chip/chips.asp?year=2013
Ciuperca G, Ridolfi A, Idier J (2003) Penalized maximum likelihood estimator for normal mixtures. Scand J Stat 30(1):45–59
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38
Hathaway RJ (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13(2):795–800
Kiefer J, Wolowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. Ann Math Stat 27(4):887–906
Li X, Chen C (2007) Inequalities for the Gamma function. J Inequal Pure Appl Math 8(1):article 28
Liu X, Pasarica C, Shao Y (2003) Testing homogeneity in Gamma mixture models. Scand J Stat 30(1):227–239
Liu G, Li P, Liu Y, Pu X (2019) On consistency of the MLE under finite mixtures of location-scale distributions with a structural parameter. J Stat Plann Inference 199:29–44
McLachlan G, Peel D (2004) Finite mixture models. Wiley, Hoboken
McLachlan GJ, Lee SX, Rathnayake SJ (2019) Finite mixture models. Annu Rev Stat Appl 6:355–378
Muna S, Purnaba I, Setiawaty B (2019) Premium rate determination of crop insurance product based on rainfall index consideration. In: IOP conference series: earth and environmental science. IOP Publishing, vol 299, p 012047
Pearson K (1894) Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A 185:71–110
Pfanzagl J (1988) Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. J Stat Plan Inference 19(2):137–158
R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Redner R (1981) Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann Stat 9(1):225–228
Sergio V, Francesca D, Giovanni P (2008) Gamma shape mixtures for heavy-tailed distributions. Ann Appl Stat
Tanaka K (2009) Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. Scand J Stat 36(1):171–184
Tanaka K, Takemura A (2005) Strong consistency of MLE for finite uniform mixtures when the scale parameters are exponentially small. Ann Inst Stat Math 57(1):1–19
Tanaka K, Takemura A (2006) Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small. Bernoulli 12(6):1003–1017
Teicher H (1963) Identifiability of finite mixtures. Ann Math Stat 34(4):1265–1269
Titterington DM, Smith AF, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York
van der Vaart AW (2000) Asymptot Stat. Cambridge University Press, New York
Wald A (1949) Note on the consistency of the maximum likelihood estimate. Ann Math Stat 20(4):595–601
Willmot GE, Liu XS (2011) Risk modelling with the mixed Erlang distribution. Appl Stoch Model Bus Ind 27(1):2–16
Wong S, Li W (2014) Test for homogeneity in Gamma mixture models using likelihood. Comput Stat Data Anal 70:127–137
Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11(1):95–103
Wywiał JL (2018) Application of two gamma distributions mixture to financial auditing. Sankhya B 80(1):1–18
Yin C, Lin XS, Huang R, Yuan H (2019) On the consistency of penalized MLEs for Erlang mixtures. Stat Probab Lett 145:12–20
Young DS, Chen X, Hewage DC, Nilo-Poyanco R (2019) Finite mixture-of-gamma distributions: estimation, inference, and model-based clustering. Adv Data Anal Classif 13(4):1053–1082
Yu Y (2021) An introduction to mixR

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.