Annals of the Institute of Statistical Mathematics (2023) 75:833–853
https://doi.org/10.1007/s10463-023-00865-7

Mixture of shifted binomial distributions for rating data

Shaoting Li[1] · **Jiahua Chen**[2,3]

Received: 19 April 2021 / Revised: 18 October 2022 / Accepted: 10 January 2023 /
Published online: 10 February 2023

Abstract

Rating data are a kind of ordinal categorical data routinely collected in survey sampling. The response value in such applications is confined to a finite number of ordered categories. Due to population heterogeneity, the respondents may have several different rating styles. A finite mixture model is thus most suitable to fit datasets of this nature. In this paper, we propose a two-component mixture of shifted binomial distributions for rating data. We show that this model is identifiable and propose a numerically stable penalized likelihood approach for parameter estimation. We adapt an expectation-maximization algorithm for the penalized maximum likelihood estimation. Our simulation results show that the penalized maximum likelihood estimator is consistent and effective. We fit the proposed model and other models in the literature to some real-world datasets and find the proposed model can have much better fits.

Keywords Binomial distribution · Categorical data · Identifiability · EM algorithm · Mixture model · Ordinal data · Rating data

1 Introduction

Survey samplings are widely used in applications to learn about the current status of various populations. The response variable can be binary such as employment status, quantitative such as annual income, or ordinary categorical variable such as the rating of a movie. Rating is a typical example of ordinal categorical observation. We collect rating data to evaluate the quality of the product and learn about the preferences of the general public (Agresti 2010). Consumers are asked

✉ Jiahua Chen
jhchen@stat.ubc.ca

[1] School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, Liaoning, China

[2] School of Mathematics and Statistics, Yunnan University, Kunming 650091, Yunnan, China

[3] Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

to rate movies, games, restaurants, hotels, stores, and other goods and services. The information derived from the analysis of rating data is useful to both merchants and consumers. For merchants, they would like to know about consumer satisfaction with their products. For consumers, the information helps them obtain quality products they need at competitive prices.

It is generally believed that the score data with the centesimal system reflect the cumulation of many aspects of an individual's ability. We may hence fit these data to normal distribution supported by the central limit theorem. For discretized scores, shifted binomial distribution is proposed for rating data in the literature (Zhou and Lange 2009). Suppose a rating outcome has m ordered categories. Let R be the rating given to a product by a consumer. Let $\text{Binom}(m, \theta)$ and $\text{Binom}(r;m, \theta)$ stand for a binomial distribution and its probability mass function resulted from m Bernoulli trials with probability of success θ. For a positive integer $s(s \le m)$, if $R - s$ is distributed as $\text{Binom}(m - s, 1 - \theta)$, we say that R has a shifted binomial distribution $B_s(m, \theta)$ with size parameter m, shift parameter s, and feeling parameter θ. Its probability mass function is

$$B_s(r;m, \theta) = P(R = r) = \text{ Binom }(r - s;m - s, 1 - \theta) = \binom{m-s}{r-s}\theta^{m-r}(1-\theta)^{r-s} \tag{1}$$

for $r = s, s + 1, \dots, m$. When $s = m$, $B_s(m, \theta)$ for any θ value degenerates to $P(R = m) = 1$. In this case, we let $\theta = 0$ when presenting this distribution. Parameter θ in $B_s(m, \theta)$ determines the degree of skewness of the rating distribution. When θ is large, the respondent is more likely to assign lower scores. Iannario (2010) and Simone (2021) referred to it as a feeling parameter. We note that when R has shifted binomial distribution given above, $m - R$ has a binomial distribution $\text{Binom}(m - s, \theta)$.

In many applications, the model (1) does not appropriately fit the rating data. One explanation is that besides the normal score, there are some confounding factors influencing the rating of the consumers. These confounding factors lead to the presence of subpopulations. In this case, the finite mixture model allows us to describe the proportions of subpopulations as well as their distributional properties (Lindsay 1995; McLachlan and Peel 2000). For example, Zhou and Lange (2009) modeled the rating data of movies by a two-component shifted binomial mixture distribution. Oh (2014) studied the maximum likelihood estimation for the mixture of shifted binomial distribution, and Breen and Luijkx (2010) used an ordered logit mixture model for ordinal data.

Due to the existence of indecision in the response, every discrete ordinal choice has an intrinsic degree of uncertainty. Taking this into consideration, Piccolo (2003) and D'elia and Piccolo (2005) proposed a two-component mixture model (CUB) consisting of a shifted binomial distribution and a uniform distribution, with one component being a shifted binomial distribution as in (1), with $s = 1$, and another being a discrete uniform distribution on $\{1, 2, \dots, m\}$. The corresponding probability mass function is given by, for some mixing proportion $\pi \in [0, 1]$,

$$P(R=r) = \pi\, B_1(r;m,\theta) + (1-\pi)\Big(\frac{1}{m}\Big) \tag{2}$$

for $r = 1, 2, \dots, m$. The second component in this mixture suggests that a proportion of $1-\pi$ customers give out random scores. Iannario (2010) studied the identifiability of the CUB model (2). Simone (2021) developed a CUB regression model to describe the influence of some covariates on scores and proposed an accelerated expectation-maximization (EM) algorithm (Dempster et al. 1977) for parameter estimation under this model.

We propose a new model for rating data from a different angle. In many situations, some customers are reluctant to assign the lowest score to a product. This generalizes further, for example, to the second lowest score. The customers in this subgroup or subpopulation, therefore, assign scores only from a higher score range, say $\{s, s+1, \dots, m\}$, rather than within the suggested range $\{1, 2, \dots, m\}$. Because of this, a new two-component shifted binomial mixture (SBM) model can be a better choice for many rating datasets. We propose a shifted binomial mixture (SBM) model with two subpopulations $B_1(m,\theta_1)$ and $B_s(m,\theta_2)$, for a shift parameter s and mixing proportions π and $1-\pi$.

We show that the proposed SBM is identifiable in its mixing distributions. We develop a numerically stable penalized likelihood approach to estimate the model parameters, with the penalty being a function of the mixing proportions. We devise an EM algorithm to compute the penalized maximum likelihood estimator (PMLE). We show that the PMLE has good statistical properties through a large-scale simulation study. We also fit the proposed model to some real-world rating datasets on movies and consumer goods. Our model provides the best fit compared with some existing models.

We organize this article as follows. The next section introduces the SBM model. In Sect. 3, we present the identifiability proof of the SBM model. In Sect. 4, we introduce the PMLE and the EM algorithm. In Sect. 5, we evaluate the performance of the SBM model through some simulation experiments and real data analysis. Section 6 contains conclusions and discussions.

2 Model

The proposed SBM model for the rating random variable R can be written as

$$\text{SBM}\,(m, G) = \pi B_1(m,\theta_1) + (1-\pi) B_s(m,\theta_2)$$

with G being the mixing distribution

$$G = \pi\{(1,\theta_1)\} + (1-\pi)\{(s,\theta_2)\},$$

which is a probability measure assigning probabilities π and $(1-\pi)$ to the subpopulation parameter vectors $(1,\theta_1)$ and (s,θ_2), respectively. We denote the space of G by

$$\mathbb{G} = \Big\{\pi\{(1,\theta_1)\} + (1-\pi)\{(s,\theta_2)\} \,:\, \pi,\theta_1,\theta_2 \in [0,1],\ s \in \{1,2,\dots,m-1\}\Big\} \\ \cup \Big\{\pi\{(1,\theta_1)\} + (1-\pi)\{(m,0)\} \,:\, \pi,\theta_1 \in [0,1]\Big\}. \tag{3}$$

The first portion of $\mathbb{G}$ contains mixing distributions with $s < m$. The second portion of $\mathbb{G}$ has $s = m$ and $\theta_2 = 0$.

The probability mass function of SBM (m, G) is given by

$$\begin{aligned} \text{SBM}\,(r;m,G) =& \pi\, B_1(r;m,\theta_1) + (1-\pi)\, B_s(r;m,\theta_2) \\ =& \pi \binom{m-1}{m-r} \theta_1^{m-r}(1-\theta_1)^{r-1} + (1-\pi) \binom{m-s}{m-r} \theta_2^{m-r}(1-\theta_2)^{r-s} \end{aligned} \tag{4}$$

for $r = 1, 2, \dots, m$. By convention, $\binom{m-s}{m-r} = 0$ when $m - r > m - s$.

3 Identifiability of the SBM model

Identifiability is an essential requirement for statistical inference. Clearly, the SBM model is not identifiable in terms of π, θ_1, θ_2, and s. When the mixing proportion $\pi = 0$, the distribution in (4) is the same for any value of θ_1. When the subpopulation parameter $\theta_2 = 0$, the distribution in (4) is the same for any value of s. When $s = 1$ and $\theta_1 = \theta_2$, (4) is the same for any value of π. Nevertheless, the SBM is identifiable in terms of the mixing distribution G in the space (3).

Many researchers studied the identifiability of finite mixture models (Atienza et al. 2006; Lindsay 1995). It is known that the two-component binomial mixture model is identifiable when the number of trials is 4 or higher. In this section, we show that the SBM model (4) is identifiable when $m - 1 \geq 4$; an extra trial is needed.

Theorem 1 *Suppose $G, G' \in \mathbb{G}$ with $\mathbb{G}$ defined by (3) and $m \geq 5$. Then,* SBM $(m, G) =$ SBM (m, G') *if and only if $G = G'$.*

The rest of this section proves this result. For the convenience of presentation, we write two mixing distributions as

$$\begin{aligned} G =& \pi\{(1,\theta_1)\} + (1-\pi)\{(s,\theta_2)\}; \\ G' =& \alpha\{(1,\eta_1)\} + (1-\alpha)\{(t,\eta_2)\}. \end{aligned}$$

For a real x and positive integer k, we define its kth factorial $x^{(k)} = x(x-1)\cdots(x-k+1)$. Further, we let $x^{(0)} = 1$ when $x \neq 0$ and $0^{(0)} = 0$. When X has a binomial distribution with the parameters m and θ, it is easy to verify that $E\{X^{(k)}\} = m^{(k)}\theta^k$. When R has a shifted binomial distribution, $m - R$ has a binomial distribution. Hence, when R has SBM (m, G) or SBM (m, G') distributions, the factorial moments of $(m - R)$ are easily found to be

$$\begin{aligned} E_G\{(m-R)^{(k)}\} =& \pi(m-1)^{(k)}\theta_1^k + (1-\pi)(m-s)^{(k)}\theta_2^k, \\ E_{G'}\{(m-R)^{(k)}\} =& \alpha(m-1)^{(k)}\eta_1^k + (1-\alpha)(m-t)^{(k)}\eta_2^k. \end{aligned}$$

When SBM $(m, G) =$ SBM (m, G'), we have

$$\pi(m-1)^{(k)}\theta_1^k + (1-\pi)(m-s)^{(k)}\theta_2^k = \alpha(m-1)^{(k)}\eta_1^k + (1-\alpha)(m-t)^{(k)}\eta_2^k$$

for $k = 0, 1, \ldots, m-1$. Let us introduce column vectors

$$\mathbf{a}_1 = (a_{10}, a_{11}, \ldots, a_{1,m-1})^\top,\ \mathbf{a}_2 = (a_{20}, a_{21}, \ldots, a_{2,m-1})^\top,$$
$$\mathbf{b}_1 = (b_{10}, b_{11}, \ldots, b_{1,m-1})^\top,\ \mathbf{b}_2 = (b_{20}, b_{21}, \ldots, b_{2,m-1})^\top$$

with it entries being

$$a_{1k} = \pi(m-1)^{(k)}, \quad a_{2k} = (1-\pi)(m-s)^{(k)}; \tag{5}$$

$$b_{1k} = \alpha(m-1)^{(k)}, \quad b_{2k} = (1-\alpha)(m-t)^{(k)}. \tag{6}$$

We can then summarize the moment equations by a matrix equation:

$$\begin{bmatrix} \mathbf{a}_1 & -\mathbf{b}_1 & \mathbf{a}_2 & -\mathbf{b}_2 \end{bmatrix} \times \begin{bmatrix} 1 & \theta_1 & \theta_1^2 & \ldots & \theta_1^{m-1} \\ 1 & \eta_1 & \eta_1^2 & \ldots & \eta_1^{m-1} \\ 1 & \theta_2 & \theta_2^2 & \ldots & \theta_2^{m-1} \\ 1 & \eta_2 & \eta_2^2 & \ldots & \eta_2^{m-1} \end{bmatrix} = 0.$$

A key fact is that the second matrix in the above equation is a Vandermonde matrix. Such a matrix has full row rank 4 when $m \geq 5$ and $\theta_1, \eta_1, \theta_2, \eta_2$ have distinct values. When $m \geq 5$ and the matrix equation holds, we show $G = G'$. Although this line of proof is straightforward, we must go over each of many trivial and tedious cases. We form cases by how many distinct values $\theta_1, \eta_1, \theta_2, \eta_2$ assume.

Consider the **first case** when $\theta_1, \theta_2, \eta_1, \eta_2$ all have distinct values. In this case, the Vandermonde matrix has full row rank 4. The linear combination of four rows is not zero unless all coefficients are zero. Therefore, $\mathbf{a}_1 = \mathbf{b}_1 = \mathbf{a}_2 = \mathbf{b}_2 = 0$. This leads to equations on the mixing proportions and shift parameters:

$$\pi(m-1)^{(k)} = (1-\pi)(m-s)^{(k)} = \alpha(m-1)^{(k)} = (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \ldots, m-1$ in view of (5) and (6). Note that all the terms are nonnegative and $0^{(0)} = 0$. There exists only one solution: $\pi = \alpha = 0$ and $s = t = m$. The corresponding mixing distributions are as follows

$$G = 0\{(1, \theta_1)\} + 1\{(m, 0)\};\ G' = 0\{(1, \eta_1)\} + 1\{(m, 0)\}.$$

Clearly, $G = G'$.

Consider next the **second case** when $\theta_1, \theta_2, \eta_1, \eta_2$ assume 3 distinct values. In this case, there can be only three sub-cases due to symmetry: (a) $\theta_1 = \theta_2$, (b) $\theta_1 = \eta_1$, and (c) $\theta_2 = \eta_2$.

In sub-case (a), with $\theta_1 = \theta_2$, the matrix equation is simplified to the following

$$\begin{bmatrix} \mathbf{a}_1 + \mathbf{a}_2 & -\mathbf{b}_1 & -\mathbf{b}_2 \end{bmatrix} \times \begin{bmatrix} 1 & \theta_1 & \theta_1^2 & \dots & \theta_1^{m-1} \\ 1 & \eta_1 & \eta_1^2 & \dots & \eta_1^{m-1} \\ 1 & \eta_2 & \eta_2^2 & \dots & \eta_2^{m-1} \end{bmatrix} = 0.$$

The linear independence of the three rows implies $\mathbf{a}_1 + \mathbf{a}_2 = \mathbf{b}_1 = \mathbf{b}_2 = 0$. They imply

$$\pi(m-1)^{(k)} + (1-\pi)(m-s)^{(k)} = \alpha(m-1)^{(k)} = (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6). Note again that all the terms are non-negative. Hence, it is the same as

$$\pi(m-1)^{(k)} = (1-\pi)(m-s)^{(k)} = \alpha(m-1)^{(k)} = (1-\alpha)(m-t)^{(k)} = 0,$$

which implies $G = G'$, as shown in the first case.

In sub-case (b) with $\theta_1 = \eta_1$, the matrix equation can be simplified in a similar way. This time, we get $\mathbf{a}_1 - \mathbf{b}_1 = \mathbf{a}_2 = \mathbf{b}_2 = 0$. They imply

$$\pi(m-1)^{(k)} - \alpha(m-1)^{(k)} = (1-\pi)(m-s)^{(k)} = (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6). The first term implies $\pi = \alpha$. If $\pi = \alpha = 1$, it leads to $G = G'$ because $\theta_1 = \eta_1$. If $\pi = \alpha < 1$, the second and third terms are zero only if $s = t = m$. This leads to the conventional value $\theta_2 = \eta_2 = 0$ as required by (3). The corresponding mixing distributions are as follows:

$$G = G' = \pi\{(1, \theta_1)\} + (1-\pi)\{(m, 0)\}.$$

That is, we also have $G = G'$.

In sub-case (c) with $\theta_2 = \eta_2$, the matrix equation leads to $\mathbf{a}_1 = \mathbf{b}_1 = \mathbf{a}_2 - \mathbf{b}_2 = 0$. They imply

$$\pi(m-1)^{(k)} = \alpha(m-1)^{(k)} = (1-\pi)(m-s)^{(k)} - (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6). The first two entries imply $\pi = \alpha = 0$. Subsequently, the third entry has solution $s = t$. With $\theta_2 = \eta_2$, the resulting G and G' are given by

$$G = 0\{(1, \theta_1)\} + 1\{(s, \theta_2)\};\ \ G' = 0\{(1, \eta_1)\} + 1\{(s, \theta_2)\}.$$

It is seen that $G = G'$ regardless of $\theta_1 \neq \eta_1$.

The **third case** in consideration is when $\theta_1, \theta_2, \eta_1, \eta_2$ assume two distinct values. In spite of symmetry, there are still five different sub-cases: (a) $\theta_1 = \theta_2 = \eta_1 \neq \eta_2$, (b) $\theta_1 = \theta_2 = \eta_2 \neq \eta_1$, (c) $\theta_1 = \theta_2 \neq \eta_1 = \eta_2$, (d) $\theta_1 = \eta_1 \neq \theta_2 = \eta_2$, and (e) $\theta_1 = \eta_2 \neq \theta_2 = \eta_1$. They are formed by having one of $\theta_1, \theta_2, \eta_1, \eta_2$ singled out to have a different value or by having their values paired up. The proofs of these sub-cases are trivial repetitions of previous proofs. We choose sub-cases (a) and (d) for illustration and omit the other proofs. If interested, one can easily make up the proof following these examples.

For sub-case (a) $\theta_1 = \theta_2 = \eta_1 \neq \eta_2$, the simplified matrix equation implies $\mathbf{a}_1 - \mathbf{b}_1 + \mathbf{a}_2 = \mathbf{b}_2 = 0$. They imply

$$\pi(m-1)^{(k)} - \alpha(m-1)^{(k)} + (1-\pi)(m-s)^{(k)} = (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6). Suppose $s < m$. Letting $k = 0$ in the first entry, we obtain

$$\pi(m-1)^{(0)} - \alpha(m-1)^{(0)} + (1-\pi)(m-s)^{(0)} = 1 - \alpha = 0,$$

so we have $\alpha = 1$. With $\alpha = 1$ and letting $k = 1$, we get

$$\pi(m-1)^{(1)} - (m-1)^{(1)} + (1-\pi)(m-s)^{(1)} = (1-\pi)(1-s) = 0.$$

The equality holds only when $s = 1$ or $\pi = 1$. Both result in G and G' given by

$$G = \pi\{(1, \theta_1)\} + (1-\pi)\{(1, \theta_1)\};\ G' = 1\{(1, \theta_1)\} + 0\{(t, \eta_2)\}.$$

It is seen that $G = G'$ in spite of their different appearances. Suppose $s = m$ in the first place. Then $\pi(m-1)^{(k)} - \alpha(m-1)^{(k)} + (1-\pi)(m-s)^{(k)} = 0$, which implies $\pi = \alpha$. If $\pi = \alpha = 1$, then

$$G = 1\{(1, \theta_1)\} + 0\{(s, \theta_1)\};\ G' = 1\{(1, \theta_1)\} + 0\{(t, \eta_2)\}$$

and $G = G'$. If $\pi = \alpha < 1$, then $(1-\alpha)(m-t)^{(k)} = 0$ implies $t = m$. Hence, $G = G'$.

In the sub-case (d) $\theta_1 = \eta_1 \neq \theta_2 = \eta_2$, the solution to the simplified matrix equation must satisfy $\mathbf{a}_1 - \mathbf{b}_1 = \mathbf{a}_2 - \mathbf{b}_2 = 0$. They imply

$$\pi(m-1)^{(k)} - \alpha(m-1)^{(k)} = (1-\pi)(m-s)^{(k)} - (1-\alpha)(m-t)^{(k)} = 0$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6). The first entry implies $\pi = \alpha$. After which, the second entry requires $s = t$. Hence $G = G'$.

The proofs for other sub-cases are similar and omitted here.

The **fourth case** in consideration is when $\theta_1 = \theta_2 = \eta_1 = \eta_2$. The solution to the simplified matrix equation must satisfy $\mathbf{a}_1 + \mathbf{a}_2 = \mathbf{b}_1 + \mathbf{b}_2$. It implies

$$\pi(m-1)^{(k)} + (1-\pi)(m-s)^{(k)} = \alpha(m-1)^{(k)} + (1-\alpha)(m-t)^{(k)}$$

for $k = 0, 1, \dots, m-1$ in view of (5) and (6).

(i) If $s, t > 1$, then $(m-s)^{(k)} = (m-t)^{(k)} = 0$ but $(m-1)^{(k)} = (m-1)!$ when $k = m-1$. The above equation becomes $\pi(m-1)! = \alpha(m-1)!$ so $\pi = \alpha$. After which, the equation becomes $(m-s)^{(k)} = (m-t)^{(k)}$, implying $s = t$. Hence, $G = G'$.

(ii) If $s = 1$ instead, the left side of the above equation is equal to $(m-1)^{(k)}$, and the equation becomes $(m-1)^{(k)} = \alpha(m-1)^{(k)} + (1-\alpha)(m-t)^{(k)}$. The solutions to the equation are $t = 1$ or $\alpha = 1$. They all imply $G = G'$.

We have now exhausted all possibilities and thus proved the identifiability of the SBM model.

4 The PMLE under the SBM Model

Let $r_1, r_2, \dots, r_n$ be a random sample from model (4) representing n scores of a product assigned by n consumers. Based on this dataset, we aim to estimate parameters of model (4): mixing proportion π, feeling parameters θ_1, θ_2, and shift parameter s. For convenience, let $\boldsymbol{\theta}$ be the vector of two feeling parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The log-likelihood function is given by

$$\ell_n(\pi, \boldsymbol{\theta}, s) = \sum_{i=1}^{n} \log \left\{ \pi\, B_1(r_i; m, \theta_1) + (1-\pi)\, B_s(r_i; m, \theta_2) \right\}.$$

The maximum likelihood estimator (MLE) is root-n consistent and asymptotically normal with optimal asymptotic variance under regular models. The finite mixture models are not regular, because the mixing proportion can take a value 0, which is the boundary of the parameter space. One of the consequences is that the best possible rate of convergence is merely $n^{-1/4}$ (Chen 1995).

However, there is plenty of empirical evidence supporting its superior performance. The MLE can be numerically unstable when the true mixing proportions are close to 0 or 1 under finite mixture models. Under the current model, where $\pi = 0$, the likelihood function does not depend on θ_1, which makes estimating θ_1 a meaningless task. Such a phenomenon is often referred to as a partial loss of identifiability. To counter this, one may adopt a modified log-likelihood function, as suggested by Chen (1998) and Chen et al. (2001). We define the penalized maximum likelihood estimator (PMLE) to be

$$(\hat{\pi}, \hat{\boldsymbol{\theta}}, \hat{s}) = \arg \max_{(\pi, \boldsymbol{\theta}, s)} p\ell_n(\pi, \boldsymbol{\theta}, s)$$

with

$$p\ell_n(\pi, \boldsymbol{\theta}, s) = \ell_n(\pi, \boldsymbol{\theta}, s) + p(\pi),$$

where $p(\pi)$ is a continuous function that approaches negative infinity as π approaches 0 or 1. In this study, we use the following

$$p(\pi) = \lambda \log(1 - |1 - 2\pi|)$$

with some $\lambda > 0$, following Li et al. (2009). Obviously, this penalty function decreases to negative infinity when π approaches 0 or 1 and reaches the maximum at $\pi = 0.5$. With the help of this penalty, the PMLE of π avoids 0 and 1 and is pushed toward the middle value of 0.5. This ensures an appropriate number of observations for estimating parameters in each subpopulation. Consequently, the PMLE is numerically more stable.

Because the mixing distribution space $\mathbb{G}$ is compact, and the SBM model is identifiable; from Kiefer and Wolfowitz (1956), it is easy to show that the PMLE,

$$\hat{G} = \hat{\pi}\{(1, \hat{\theta}_1)\} + (1 - \hat{\pi})\{(\hat{s}, \hat{\theta}_2)\},$$

is a consistent estimator of the mixing distribution G.

We can easily adapt the commonly used EM algorithm to obtain the PMLE of G under the SBM model. For each fixed shift parameter $s \in \{1, 2, \dots, m\}$, the EM algorithm obtains the restricted PMLE that maximizes $p\ell_n(\pi, \theta, s)$ with respect to π and $\boldsymbol{\theta}$ but fixed s. Let

$$(\hat{\pi}(s), \hat{\boldsymbol{\theta}}(s)) = \arg\max_{(\pi,\boldsymbol{\theta})} p\ell_n(\pi, \boldsymbol{\theta}, s).$$

We then get

$$\hat{s} = \arg\max_{s} p\ell_n(\hat{\pi}(s), \hat{\boldsymbol{\theta}}(s), s)$$

and the PMLE of G through $\hat{\pi}(\hat{s}), \hat{\boldsymbol{\theta}}(\hat{s})$ and $\hat{s}$.

To compute the restricted PMLE given s, the adopted EM algorithm proceeds as follows. Recall that each observation r_i is a random outcome of either subpopulations $B_1(m, \theta_1)$ or $B_s(m, \theta_2)$. Let $z_i = 1$ if it is the former and $z_i = 0$ otherwise. In applications, we do not observe z_i, or it is missing. If the complete data (r_i, z_i), $i = 1, 2, \dots, n$ are available, the complete data penalized log likelihood is given by

$$\begin{aligned} p\ell_n^c(\pi, \boldsymbol{\theta}, s) = &\sum_i z_i \log B_1(r_i; m, \theta_1) + \sum_i (1 - z_i) \log B_s(r_i; m, \theta_2) \\ &+ \sum_i z_i \log \pi + \sum_i (1 - z_i) \log(1 - \pi) + \lambda \log(1 - |1 - 2\pi|). \end{aligned}$$

Suppose $G^{(0)}$ is a proposed initial estimate of G. The EM algorithm replaces the unobserved z_i values by its conditional expectation $E(z_i | r_{(n)}, G)$, computed based on the current mixing distribution $G = G^{(0)}$ (or $G = G^{(t)}$ after t iterations) given data $r_{(n)} = \{r_i : i = 1, \dots, n\}$. This leads to the E step of the algorithm.

E-step. The conditional expectations are given by

$$\omega_i^{(t)} = E(z_i | r_{(n)}, G^{(t)}) = \frac{\pi^{(t)} B_1(r_i; m, \theta_1^{(t)})}{\pi^{(t)} B_1(r_i; m, \theta_1^{(t)}) + (1 - \pi^{(t)}) B_s(r_i; m, \theta_2^{(t)})}.$$

Note that $E(z_i | r_{(n)}, G^{(t)}) = 1$ when $r_i < s$.

Ignoring additive constants in the log likelihood and replacing z_i by their conditional expectations, we obtain the Q function

$$\begin{aligned} Q(\pi, \boldsymbol{\theta}) = &\sum_i \omega_i^{(t)} \{(m - r_i) \log \theta_1 + (r_i - 1) \log(1 - \theta_1)\} \\ &+ \sum_i (1 - \omega_i^{(t)}) \{(m - r_i) \log \theta_2 + (r_i - s) \log(1 - \theta_2)\} \\ &+ \sum_i \omega_i^{(t)} \log \pi + \sum_i (1 - \omega_i^{(t)}) \log(1 - \pi) + \lambda \log(1 - |1 - 2\pi|). \end{aligned}$$

Note that this function is a sum of three functions. One is a function of θ_1 only, another is a function of θ_2 only, and the third is a function of π only. This property simplifies the numerical problem of maximizing Q function which is the next step.

M-step. Let $\omega^{(t)}(s) = \sum_i \omega_i^{(t)}$. Owing to the decomposition of Q into additive functions, it is simple to find the maximum point in π is given by

$$\pi^{(t+1)} = \begin{cases} \min\{0.5, (n+\lambda)^{-1}(\omega^{(t)}(s)+\lambda)\}, & \text{if } \omega^{(t)}(s) < n/2; \\ \max\{0.5, (n+\lambda)^{-1}\omega^{(t)}(s)\}, & \text{if } \omega^{(t)}(s) \ge n/2. \end{cases},$$

The maximum point of Q with respect to $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is given by

$$\theta_1^{(t+1)} = \frac{\sum_i \omega_i^{(t)}(m - r_i)}{(m-1)\omega^{(t)}(s)};$$

$$\theta_2^{(t+1)} = \begin{cases} \dfrac{\sum_i (1-\omega_i^{(t)})(m-r_i)}{(m-s)(n-\omega^{(t)}(s))}, & s < m; \\ 0, & s = m. \end{cases}$$

For the case of $s = m$, the likelihood value does not depend on θ_2. Hence, any legitimate θ_2 value is a maximum point of Q function.

Given $G^{(0)}$, we iterate between the E step and M step. We terminate the iteration when the increment in $p\ell_n(G^{(t)})$ is below a tolerance level of 10^{-6}. The EM iteration is known to produce converging $p\ell_n(G^{(t)})$ value. The $G^{(t)}$ sequence converges to one of the potentially many local maxima. We use multiple instances of initial $G^{(0)}$ to ensure a greater chance of locating the global maximum.

5 Simulation study and real data examples

To demonstrate the utility of the SBM model and its usefulness in applications, we report some results based on several simulation experiments and apply this model to some rating data.

5.1 Simulation experiments

We first study the performance of the PMLEs under the proposed SBM model. In simulations, we choose the penalty function $p(\pi) = \lambda \log(1 - |1 - 2\pi|)$ with the size $\lambda = 1$. The size of λ within a reasonable range, such as [0.5, 5], has been repeatedly found to have little effect on the performance of the estimator despite being useful for numerical stability (Chen 1998; Chen et al. 2001; Chen and Li 2009). To increase the chance of obtaining the global maximum, we started the iterations with 10 random initial values in the EM algorithm. Without loss of generality, we let $\theta_1 \ge \theta_2$ when $s = 1$ and $\theta_2 = 0$ when $s = m$.

In our simulations, we evaluate the performance of the PMLE from two perspectives. First, we generate the random samples so that R has a distribution in the following four groups:

$$\begin{aligned}
&\text{Model I:} && 0.4\,B_1(5,0.5)+0.6\,B_3(5,\theta_2),\ \theta_2=0.1,0.3,0.5,0.7,0.9;\\
&\text{Model II:} && \pi\,B_1(5,0.4)+(1-\pi)\,B_3(5,0.6),\ \pi=0.1,0.3,0.5,0.7,0.9;\\
&\text{Model III:} && 0.4\,B_1(10,0.5)+0.6\,B_6(10,\theta_2),\ \theta_2=0.1,0.3,0.5,0.7,0.9;\\
&\text{Model IV:} && \pi\,B_1(10,0.4)+(1-\pi)\,B_6(10,0.6),\ \pi=0.1,0.3,0.5,0.7,0.9.
\end{aligned}$$

Figure 1 shows the probability mass functions of these distributions for R. Models I and II contain SBM distributions with five categories ($m=5$), and models III and IV contain SBM distributions with 10 categories ($m=10$). The mixture distributions in models I and III have a fixed first subpopulation distribution. Their second subpopulation distribution has a shift parameter values of 3 and 6 and a feeling parameter θ_2 values of 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. A large θ_2 implies a

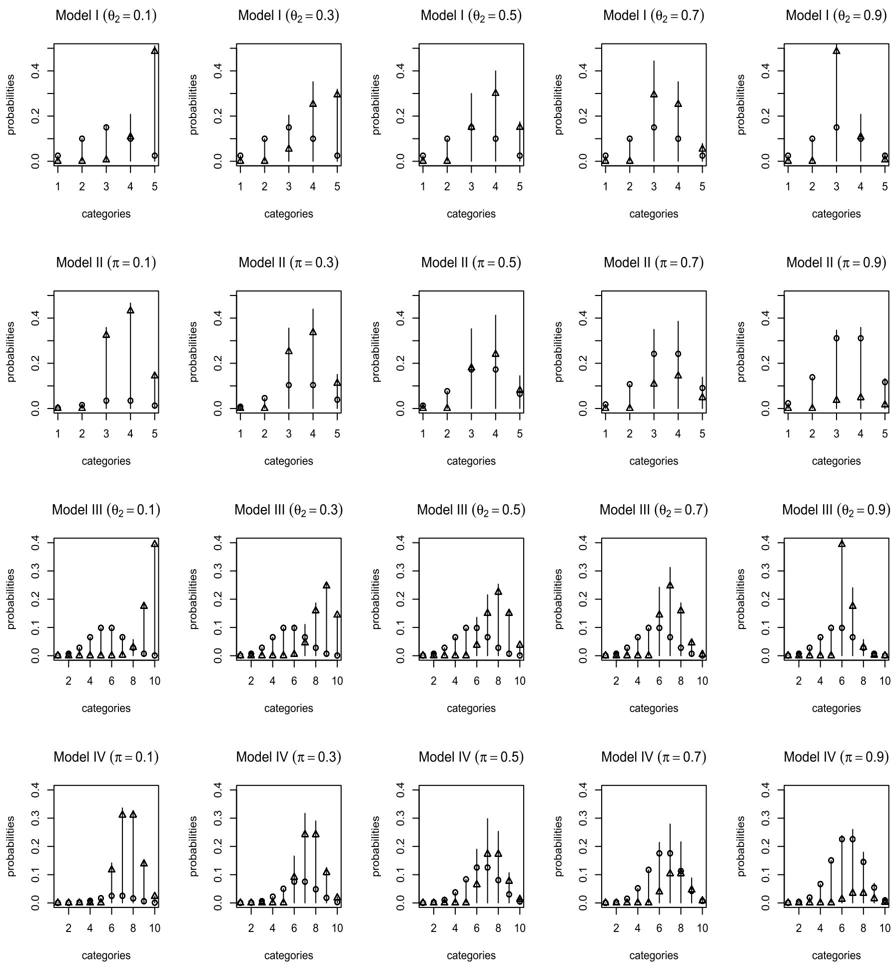

Fig. 1 The lengths of the black lines equal SBM(r; m, G), the probability mass function of the SBM; the heights of "o" and "△" equal $\pi B_1(r;m,\theta_1)$ and $(1-\pi)B_s(r;m,\theta_2)$

higher probability of assigning a score at the lower end. The distributions in models II and IV also allow us to examine the performance of PMLE with different mixing proportions. A large π implies fewer samples are from the second subpopulation making it hard to estimate s and θ_2 accurately. In the simulation, we generate a sample from each of the mixtures of sizes $n = 500, 1,000, 2,000$, and $4, 000$. The performance of the PMLE is based on $N = 1,000$ repetitions.

For each distribution and sample size, we recorded the number of times out of 1,000 repetitions when $\hat{s} = s^*$ or when $|\hat{s} - s^*| \leq 1$, where s^* is the true shift parameter value. We compute the mean and root mean squared error (RMSE) of the PMLEs $\hat{\pi}$, $\hat{\theta}_1$, $\hat{\theta}_2$ based on 1,000 repetitions. The results are presented in Tables 1 and 2.

The simulation results show that the PMLEs of π, θ_1, and θ_2 have good accuracy. When n increases, the RMSEs decrease and the accuracy of $\hat{s}$ improves. This supports the consistency of the PMLE of the mixing distribution G. Also, the accuracy of $\hat{s}$ is low when θ_2 is lower or when π is larger as we anticipated. The probability of $\hat{s} = s^*$ is larger when θ_2 is larger, and smaller when π increases or θ_2 decreases. Comparing the simulation results between these two tables, it reveals the accuracy of $\hat{s}$ is higher when $m = 5$ than when $m = 10$.

In the next simulation experiment, we study the performance of the PMLEs under the SBM with a degenerate second subpopulation. We generate R from the following four special SBM distributions:

$$\begin{aligned} \text{Model V:} &\quad 0.6\,B_1(5, 0.5) + 0.4\,B_5(5, 0); \\ \text{Model VI:} &\quad 0.4\,B_1(5, 0.7) + 0.6\,B_5(5, 0); \\ \text{Model VII:} &\quad 0.6\,B_1(10, 0.5) + 0.4\,B_{10}(10, 0); \\ \text{Model VIII:} &\quad 0.4\,B_1(10, 0.7) + 0.6\,B_{10}(10, 0). \end{aligned}$$

Fig. 2 depicts the probability mass functions of these distributions for R. The true value in these models is $s^* = m$. Algebraically, both $\theta_2^* = 0$ for any s^* value, and $s^* = m$ for any θ_2^* value correspond to the true SBM distribution. To handle this situation, for each fitted $\hat{G}$, we compute the follows:

$$\begin{aligned} D(\hat{s}, \hat{\theta}_2) &= \sum_{r=\hat{s}}^{m} (m - r)^2 B_{\hat{s}}(r; m, \hat{\theta}_2) \\ &= \sum_{r=\hat{s}}^{m} (m - r)^2 \binom{m - \hat{s}}{m - r} \hat{\theta}_2^{m-r} (1 - \hat{\theta}_2)^{r-\hat{s}}. \end{aligned}$$

When $\hat{s}$ is close to m or $\hat{\theta}_2$ is close to 0, the measure $D(\hat{s}, \hat{\theta}_2)$ will decrease to 0 and the fitted second component approaches the true model. Therefore, $D(\hat{s}, \hat{\theta}_2)$ is a good performance metric in this case.

We then compute root mean square distance (RMSD) over the repetitions (marked with a superscript (i)) as follows:

Table 1 Simulation results of PMLEs under models I and II

			$\hat{s}$		$\hat{\pi}$		$\hat{\theta}_1$		$\hat{\theta}_2$	
Model	Parameter	n	$\hat{s}=s^*$	$\lvert\hat{s}-s^*\rvert \le 1$	Mean	RMSE	Mean	RMSE	Mean	RMSE
I	$\theta_2=0.1$	500	69	739	0.415	0.051	0.494	0.036	0.130	0.070
		1,000	80	717	0.410	0.036	0.495	0.025	0.128	0.065
		2,000	80	654	0.407	0.028	0.497	0.019	0.122	0.062
		4,000	86	635	0.405	0.020	0.498	0.013	0.120	0.060
	$\theta_2=0.3$	500	459	907	0.457	0.107	0.481	0.057	0.330	0.106
		1,000	540	953	0.443	0.100	0.485	0.047	0.329	0.083
		2,000	581	941	0.427	0.091	0.491	0.041	0.318	0.073
		4,000	642	920	0.403	0.078	0.500	0.034	0.301	0.061
	$\theta_2=0.5$	500	818	969	0.449	0.128	0.483	0.064	0.495	0.084
		1,000	925	996	0.438	0.113	0.489	0.049	0.501	0.059
		2,000	982	997	0.425	0.089	0.493	0.042	0.499	0.033
		4,000	995	1,000	0.413	0.060	0.497	0.031	0.498	0.017
	$\theta_2=0.7$	500	1,000	1,000	0.438	0.113	0.492	0.057	0.717	0.042
		1,000	1,000	1,000	0.426	0.091	0.494	0.044	0.711	0.028
		2,000	1,000	1,000	0.417	0.070	0.496	0.034	0.707	0.020
		4,000	1,000	1,000	0.410	0.052	0.498	0.025	0.703	0.013
	$\theta_2=0.9$	500	1,000	1,000	0.402	0.066	0.504	0.036	0.903	0.028
		1,000	1,000	1,000	0.400	0.048	0.502	0.025	0.901	0.020
		2,000	1,000	1,000	0.399	0.034	0.501	0.018	0.901	0.014
		4,000	1,000	1,000	0.400	0.025	0.501	0.012	0.900	0.011
II	$\pi=0.1$	500	1,000	1,000	0.227	0.157	0.319	0.126	0.632	0.045
		1,000	1,000	1,000	0.185	0.115	0.340	0.101	0.620	0.031
		2,000	1,000	1,000	0.155	0.084	0.361	0.083	0.612	0.020
		4,000	1,000	1,000	0.136	0.063	0.373	0.067	0.608	0.014
	$\pi=0.3$	500	1,000	1,000	0.393	0.156	0.374	0.071	0.631	0.053
		1,000	1,000	1,000	0.361	0.129	0.384	0.058	0.620	0.036
		2,000	1,000	1,000	0.344	0.101	0.388	0.046	0.613	0.026
		4,000	1,000	1,000	0.324	0.078	0.395	0.037	0.607	0.018
	$\pi=0.5$	500	956	999	0.487	0.119	0.405	0.050	0.611	0.070
		1,000	992	1,000	0.492	0.106	0.406	0.040	0.609	0.043
		2,000	1,000	1,000	0.500	0.087	0.403	0.030	0.607	0.030
		4,000	1,000	1,000	0.501	0.075	0.402	0.024	0.604	0.023
	$\pi=0.7$	500	719	991	0.531	0.196	0.426	0.050	0.568	0.105
		1,000	823	998	0.572	0.168	0.424	0.043	0.578	0.093
		2,000	938	1,000	0.624	0.127	0.417	0.031	0.591	0.062
		4,000	987	1,000	0.659	0.099	0.410	0.027	0.598	0.040
	$\pi=0.9$	500	336	892	0.530	0.378	0.440	0.067	0.508	0.137
		1,000	330	916	0.548	0.364	0.442	0.057	0.497	0.141
		2,000	343	952	0.576	0.343	0.443	0.050	0.493	0.140
		4,000	429	978	0.612	0.318	0.440	0.045	0.497	0.136

Table 2 Simulation results of PMLEs under models III and IV

			$\hat{s}$		$\hat{\pi}$		$\hat{\theta}_1$		$\hat{\theta}_2$	
Model	Parameter	n	$\hat{s} = s^*$	$\|\hat{s} - s^*\| \leq 1$	Mean	RMSE	Mean	RMSE	Mean	RMSE
III	$\theta_2 = 0.1$	500	38	177	0.454	0.106	0.391	0.221	0.237	0.245
		1,000	84	288	0.431	0.081	0.435	0.171	0.176	0.187
		2,000	151	479	0.416	0.057	0.467	0.121	0.139	0.135
		4,000	212	605	0.403	0.027	0.494	0.054	0.108	0.069
	$\theta_2 = 0.3$	500	142	522	0.467	0.134	0.415	0.172	0.387	0.176
		1,000	240	656	0.445	0.112	0.442	0.143	0.351	0.142
		2,000	387	791	0.421	0.078	0.472	0.099	0.321	0.104
		4,000	542	924	0.404	0.042	0.493	0.051	0.302	0.063
	$\theta_2 = 0.5$	500	471	877	0.447	0.134	0.479	0.070	0.529	0.094
		1,000	620	923	0.421	0.106	0.490	0.054	0.511	0.070
		2,000	734	965	0.405	0.073	0.497	0.037	0.498	0.051
		4,000	888	996	0.397	0.038	0.501	0.018	0.495	0.028
	$\theta_2 = 0.7$	500	940	1,000	0.408	0.077	0.501	0.033	0.696	0.032
		1,000	974	999	0.405	0.057	0.500	0.025	0.698	0.021
		2,000	994	1,000	0.403	0.035	0.499	0.015	0.699	0.011
		4,000	1,000	1,000	0.403	0.024	0.499	0.010	0.700	0.006
	$\theta_2 = 0.9$	500	1,000	1,000	0.402	0.048	0.501	0.021	0.900	0.013
		1,000	1,000	1,000	0.404	0.035	0.499	0.014	0.901	0.010
		2,000	1,000	1,000	0.400	0.245	0.501	0.011	0.900	0.007
		4,000	1,000	1,000	0.400	0.018	0.500	0.008	0.900	0.005
IV	$\pi = 0.1$	500	987	1,000	0.205	0.152	0.350	0.088	0.610	0.028
		1,000	999	1,000	0.154	0.096	0.373	0.066	0.604	0.013
		2,000	1,000	1,000	0.129	0.060	0.382	0.048	0.601	0.007
		4,000	1,000	1,000	0.114	0.036	0.391	0.034	0.600	0.005
	$\pi = 0.3$	500	838	998	0.383	0.167	0.385	0.053	0.619	0.065
		1,000	938	1,000	0.338	0.122	0.395	0.042	0.606	0.040
		2,000	983	1,000	0.322	0.080	0.395	0.028	0.601	0.020
		4,000	999	1,000	0.309	0.044	0.398	0.017	0.600	0.006
	$\pi = 0.5$	500	649	954	0.492	0.128	0.403	0.036	0.608	0.086
		1,000	756	989	0.489	0.123	0.405	0.031	0.604	0.063
		2,000	877	998	0.493	0.099	0.403	0.023	0.599	0.043
		4,000	946	1,000	0.491	0.070	0.403	0.017	0.599	0.025
	$\pi = 0.7$	500	304	692	0.549	0.197	0.415	0.041	0.572	0.127
		1,000	366	777	0.572	0.185	0.417	0.034	0.574	0.111
		2,000	460	895	0.604	0.158	0.414	0.025	0.579	0.088
		4,000	589	971	0.630	0.132	0.411	0.021	0.584	0.064
	$\pi = 0.9$	500	71	268	0.553	0.370	0.414	0.050	0.539	0.153
		1,000	66	223	0.550	0.371	0.420	0.048	0.510	0.154
		2,000	71	197	0.562	0.362	0.422	0.038	0.494	0.159
		4,000	91	247	0.589	0.344	0.425	0.034	0.495	0.154

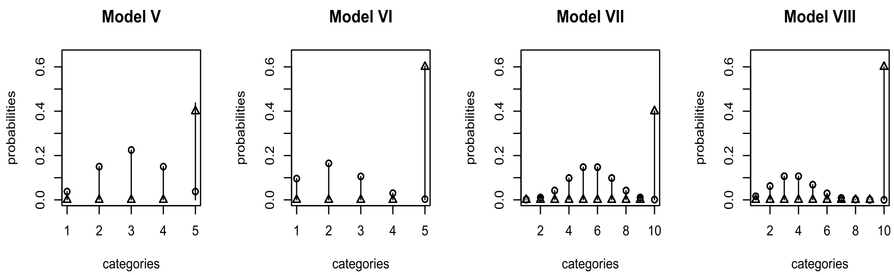

Fig. 2 The lengths of the black lines equal SBM(r; m, G), the probability mass function of the SBM. Heights of "∘" and "△" equal $\pi B_1(r;m,\theta_1)$ and $(1-\pi)B_s(r;m,\theta_2)$

$$\text{RMSD} = \big[N^{-1}\sum_{i=1}^{N}\sum_{r=\hat{s}^{(i)}}^{m}(m-r)^2 B_{\hat{s}^{(i)}}(r;m,\hat{\theta}_2^{(i)})\big]^{1/2}.$$

We also compute the mean total variation distance (MTVD) and mean Hellinger distance (MHD) as additional performance metrics:

$$\text{MTVD} = N^{-1}\sum_{i=1}^{N}\frac{1}{2}\sum_{r=\hat{s}^{(i)}}^{m}|B_{\hat{s}^{(i)}}(r;m,\hat{\theta}_2^{(i)}) - B_m(r;m,0)|,$$

$$\text{MHD} = N^{-1}\sum_{i=1}^{N}\big[\sum_{r=\hat{s}^{(i)}}^{m}\frac{1}{2}(\sqrt{B_{\hat{s}^{(i)}}(r;m,\hat{\theta}_2^{(i)})} - \sqrt{B_m(r;m,0)})^2\big]^{1/2}.$$

The other simulation arrangements are the same as before. We present the simulation results in Table 3. The simulation results show that the PMLEs are very accurate for π and θ_1. As n increases, the RMSEs of π and θ_1 decrease, as well as the RMSDs, MTVDs and MHDs under models V-VIII decrease. When $n = 4000$ under model VIII, the RMSD = 0.006, the MTVD = 3.85e−5, the MHDs = 0.003. Since Model VII has a larger m and a smaller mixing proportion of the second component, it is harder to accurately estimate s under model VII. Nevertheless, the consistency trend is clear.

5.2 Real data analysis

In this section, we apply the proposed SBM model to some real-world rating data.

Amazon Inc. is a multinational electronic commerce company in the USA. It is the largest online retailer in the world. We obtained several rate datasets from Amazon.com of two consumer products: Hisense 55-Inch Class H8 (TV-1) and H9 (TV-2) Quantum Series TVs. The rating data were collected before 2020–11–03. There are 862 and 205 ratings with 5 grades. Their histograms are shown in Fig. 3.

Table 3 Simulation results of PMLEs under models V, VI, VII and VIII

		$\hat{\pi}$		$\hat{\theta}_1$		$(\hat{s}, \hat{\theta}_2)$		
Model	n	Mean	RMSE	Mean	RMSE	RMSD	MTVD	MHD
V	500	0.581	0.038	0.510	0.023	0.187	0.033	0.090
	1,000	0.588	0.026	0.506	0.016	0.157	0.023	0.075
	2,000	0.592	0.019	0.504	0.011	0.128	0.019	0.059
	4,000	0.594	0.014	0.503	0.008	0.108	0.011	0.051
VI	500	0.397	0.023	0.705	0.019	0.078	0.006	0.034
	1,000	0.398	0.016	0.703	0.013	0.065	0.004	0.029
	2,000	0.398	0.011	0.702	0.009	0.054	0.003	0.024
	4,000	0.398	0.008	0.701	0.007	0.045	0.002	0.020
VII	500	0.550	0.097	0.381	0.245	2.322	0.244	0.259
	1,000	0.577	0.068	0.444	0.169	1.597	0.117	0.135
	2,000	0.595	0.030	0.490	0.073	0.687	0.023	0.042
	4,000	0.599	0.010	0.500	0.016	0.154	0.003	0.019
VIII	500	0.402	0.023	0.700	0.025	0.202	0.001	0.004
	1,000	0.400	0.015	0.700	0.008	0.015	2.36e−4	0.004
	2,000	0.400	0.011	0.700	0.005	0.013	1.75e−4	0.005
	4,000	0.400	0.002	0.700	0.001	0.006	3.85e−5	0.003

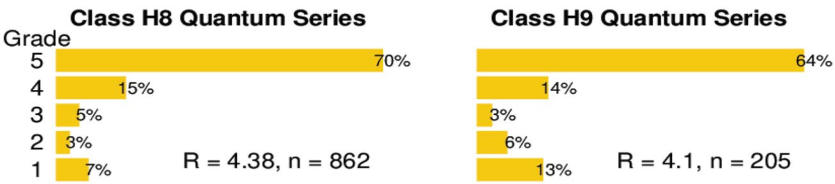

Fig. 3 Ratings of Hisense 55-Inch Class H8 and H9 TVs from amazon.com

Douban.com is a popular Chinese site for rating movies. We obtained rating datasets of three new movies shown in September in 2020: "Enola Holmes" (Movie-1), "The Social Dilemma" (Movie-2) and "The Devil All the Time" (Movie-3). The rating data were collected before 2020–11–03, There are 12,858, 15,863, and 10,470 ratings with the highest grade 5. Their histograms are shown in Fig. 4.

To determine the usefulness of the proposed SBM model, we fit these dataset with the four models with $m = 5$:

$$\begin{array}{lr} \text{Model A:} & B_1(m, \theta); \\ \text{Model B:} & \pi B_1(m, \theta) + (1 - \pi) U[1 : m]; \\ \text{Model C:} & \pi B_1(m, \theta_1) + (1 - \pi) B_1(m, \theta_2); \\ \text{Model D:} & \pi B_1(m, \theta_1) + (1 - \pi) B_s(m, \theta_2), \end{array}$$

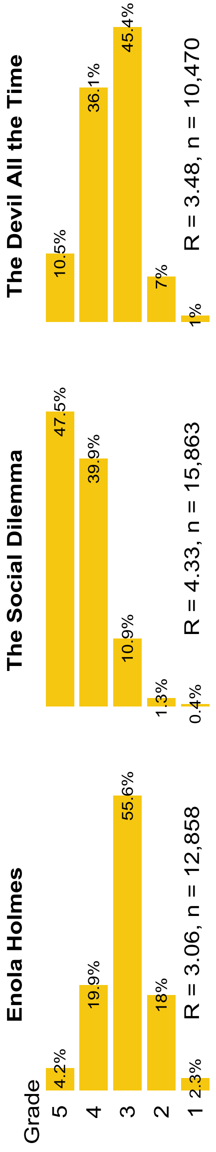

Fig. 4 Rating data for three movies from douban.com

Table 4 Fitted models for TV rating datasets

Dataset	Model	Fitted result	$\ell_n(p\ell_n)$	AIC	BIC
TV-1	A	$B_1(4, 0.155)$	−1192.71	2387.41	2396.17
	B	$0.759\,B_1(4, 0.037) + 0.241\,U[1:4]$	−846.85	1697.69	1707.21
	C	$0.127\,B_1(4, 0.830) + 0.873\,B_1(4, 0.057)$	−849.86(−851.29)	1705.72	1720.00
	D	$0.127\,B_1(4, 0.830) + 0.873\,B_1(4, 0.057)$	−849.86(−851.29)	1707.72	1726.76
TV-2	A	$B_1(4, 0.226)$	−370.22	742.45	745.77
	B	$0.630\,B_1(4, 0.026) + 0.370\,U[1:4]$	−233.397	470.79	477.44
	C	$0.209\,B_1(4, 0.891) + 0.791\,B_1(4, 0.054)$	−225.54(−226.41)	457.09	459.09
	D	$0.209\,B_1(4, 0.891) + 0.791\,B_1(4, 0.054)$	−225.54(−226.41)	467.06	472.37

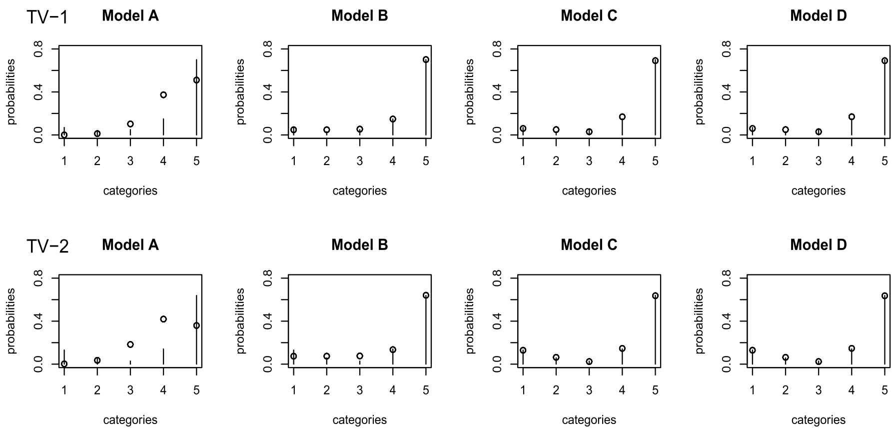

Fig. 5 Fitted models for the TV rating datasets. The black lines represent the histogram of the real data, the heights of "∘" are the fitted values

where $U[1:m]$ denotes the discrete uniform distribution on $\{1, 2, \dots, m\}$. We note that model A is a homogeneous shifted binomial model, model B is the CUB model introduced in (2), model C is the SBM with fixed $s = 1$, and model D is the proposed SBM model.

For models C and D, we use the penalty function $p(\pi) = \lambda \log(1 - |1 - 2\pi|)$ with the size $\lambda = 1$ to compute the PMLEs of the parameters. In the EM iterations, we use 10 random initial values to increase the chance of obtaining the global maximum. For the dataset of "Movie-2," we obtain two local maximum and retained the outcome with the larger log-likelihood value for in 10 replications. For other datasets, we obtain the unique local maximum values.

Figure 5 contains the probability mass functions of the fitted models together with the data histograms of the TV rating datasets. Table 4 contains the estimated parameter values. Model A fits these two datasets poorly from the plots in Fig. 5.

Table 5 Fitted models for movie rating datasets

Dataset	Model	Fitted model	$\ell_n(p\ell_n)$	AIC	BIC
Movie-1	A	$B_1(4, 0.486)$	– 16,064.35	32,130.70	32,138.16
	B	$1\,B_1(4, 0.486) + 0\,U[1 : 4]$	– 16,064.35	32,132.70	32,147.62
	C	$0.623\,B_1(4, 0.486) + 0.377\,B_1(4, 0.486)$	– 16,064.35(– 16065.80)	32,134.70	32,157.08
	D	$0.677\,B_1(4, 0.487) + 0.323\,B_3(4, 0.968)$	– 15,187.43(– 15187.87)	30,382.86	30,412.71
Movie-2	A	$B_1(4, 0.168)$	– 16,569.08	33,140.17	33,147.84
	B	$0.987\,B_1(4, 0.164) + 0.013\,U[1 : 4]$	– 16,524.41	33,052.82	33,068.17
	C	$0.003\,B_1(4, 0.999) + 0.997\,B_1(4, 0.165)$	– 16,514.88 (– 16,520.58)	33,035.76	33,058.78
	D	$0.008\,B_1(4, 0.846) + 0.992\,B_2(4, 0.217)$	– 16,504.94 (– 16,509.08)	33,017.88	33,048.56
Movie-3	A	$B_1(4, 0.380)$	– 13,033.82	26,069.64	26,076.90
	B	$1\,B_1(4, 0.380) + 0\,U[1 : 4]$	– 13,033.82	26,071.64	26,086.15
	C	$0.494\,B_1(4, 0.380) + 0.506\,B_1(4, 0.380)$	– 13,033.82 (– 13,035.21)	26,073.64	26,095.41
	D	$0.525\,B_1(4, 0.376) + 0.475\,B_3(4, 0.769)$	– 12,514.42 (– 12,514.47)	25,036.83	25,065.86

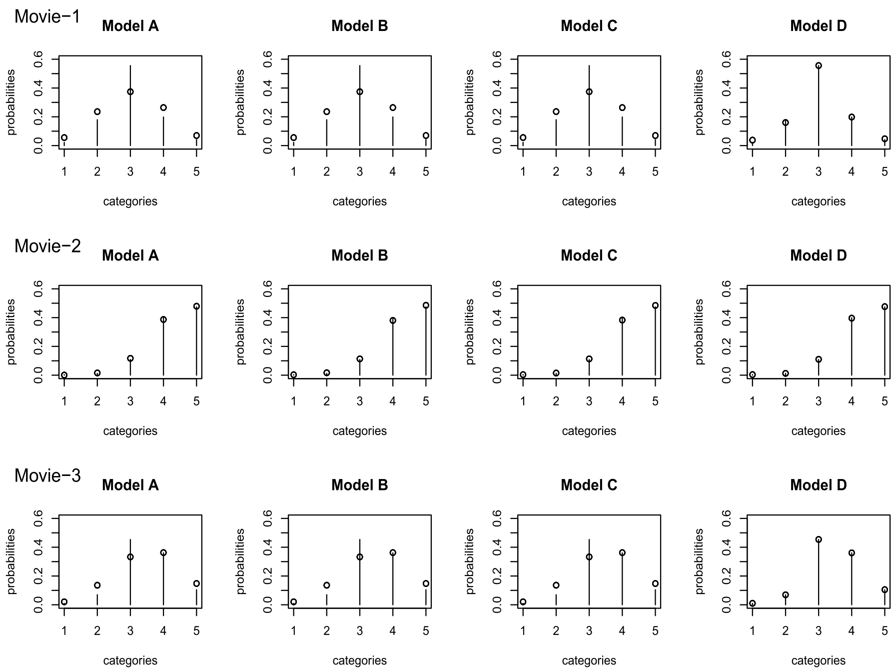

Fig. 6 Fitted models for movie rating datasets. The black lines represent the histogram of the real data, and the heights of "∘" are the fitted values

The other three models fit two datasets satisfactorily according to Fig. 5. When we look into the maximum log-likelihood values, Model B is slightly better than Models C and D for TV-1, and it is worse than models C and D for TV-2. Models C and D lead to the same fitted models (or distributions).

Figure 6 contains the probability mass functions of the fitted models together with the data histograms of the movie rating datasets. Table 5 contains the estimated parameter values. For Movie-1 and Movie-3, the fitted models A, B, and C are identical. The proposed SBM model (model D) outperforms other models significantly. For Movie-2, all four models fit data well as indicated by the plots in Fig. 6. The proposed SBM model (model D) has the best performance.

6 Conclusions

In this study, we propose a new shifted binomial mixture model for rating data. The new model has two components: The first component describes the normal rating, and the second component is for a subpopulation not willing to assign low scores. We suggest that this model should tally with the rating behavior of the general population. This is evident by its better fittings with several real datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00865-7.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 11701071, 11871419), the Natural Science and Engineering Research Council (Grant No. 2019–04204) and the Scientific Research Projects of Dongbei University of Finance and Economics (Grand No. 20210261). We thank the AE and referee for helpful comments and suggestions.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken: John Wiley and Sons.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J. M. (2006). A new condition for identifiability of finite mixture distributions. *Metrika, 63,* 215–221.

Breen, R., Luijkx, R. (2010). Mixture models for ordinal data. *Sociological Methods and Research, 39,* 3–24.

Chen, H., Chen, J., Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *63,* 19–29.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics, 23,* 221–233.

Chen, J. (1998). Penalized likelihood ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics, 26,* 583–599.

Chen, J., Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics, 37,* 2523–2542.

D'elia, A., Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics Data Analysis, 49,* 917–934.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. *Series B* (*Methodological*), *39,* 1–38.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data. *Metron, 68,* 87–94.

Kiefer, J., Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics, 27,* 887–906.

Li, P., Chen, J., Marriott, P. (2009). Non-finite Fisher information and homogeneity: The EM approach. *Biometrika, 96,* 411–426.

Lindsay, B. G. (1995). *Mixture models*: *theory. Geometry and applications*. Hayward: Institute for Mathematical Statistics.

McLachlan, G. J., Peel, D. (2000). *Finite mixture models*. New York: John Wiley and Sons.

Oh, C. (2014). A maximum likelihood estimation method for a mixture of shifted binomial distributions. *Journal of the Korean Data and Information Science Society, 25,* 255–261.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica, 5,* 85–104.

Simone, R. (2021). An accelerated EM algorithm for mixture models with uncertainty for rating data. *Computational Statistics, 36,* 691–714.

Zhou, H., Lange, K. (2009). Rating movies and rating the raters who rate them. *The American Statistician, 63,* 297–307.