IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 21, 2019, accepted January 21, 2020, date of publication February 11, 2020, date of current version March 3, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2973170*

# Knowledge Discovery and Recommendation With Linear Mixed Model

**ZHIYI CHEN[1,2], SHENGXIN ZHU[1,3], QIANG NIU[1,3], AND TIANYU ZUO[1]**

[1]Department of Mathematical Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[2]Department of Statistics, Columbia University, New York, NY 10025, USA
[3]Laboratory for Intelligent Computing and FinTech, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China

Corresponding authors: Shengxin Zhu (shengxin.zhu@xjtlu.edu.cn) and Qiang Niu (qiang.niu@xjtlu.edu.cn)

This work was supported in part by the Laboratory of Computational Physics under Grant 6142A05180501, in part by the Jiangsu Science and Technology Basic Research Program under Grant BK20171237, in part by the Key Program Special Fund through the Xi'an Jiaotong-Liverpool University (XJTLU) under Grant KSF-E-21, Grant KSF-E-32, and Grant KSF-P-02, in part by the Research Development Fund of XJTLU under Grant RDF-2017-02-23, and in part by the National Natural Science Foundation of China (NSFC) under Grant 11571002, Grant 11571047, Grant 11671049, Grant 11671051, Grant 61672003, and Grant 11871339.

**ABSTRACT** We give a concise tutorial on knowledge discovery with linear mixed model in movie recommendation. The versatility of mixed effects model is well explained. Commonly used methods for parameter estimation, confidence interval estimate and evaluation criteria for model selection are briefly reviewed. Mixed effects models produce sound inference based on a series of rigorous analysis. In particular, we analyze millions of movie rating data with LME4 R package and find solid evidences for a general social behavior: the young tend to be more censorious than senior people when evaluating the same object. Such a social behavior phenomenon can be used in recommender systems and business data analysis.

**INDEX TERMS** Knowledge discovery in database (KDD), linear mixed-effects model (LMM), recommender system (RS), R software.

## I. INTRODUCTION

With the dramatic evolvement of digital concept and data analysis, humans are constantly attaching new physical meanings to the massive data. Main sources of such data can be the figures in checkup reports, the specific grade in academic transcripts, or even inconspicuous behaviors derived from individuals. As a result, people nowadays might be shocked by their huge amount of personal data, such as their historical visits to some places, or detailed evaluations about some commodities on the internet. The volume of such data is bound to jump up exponentially in the data centric-era [1]. The main focus of our study is how to unearth useful information or knowledge from digital data that grows rapidly. Such an exploring process is called knowledge discovery [2].

Recommender systems (RS), a popular and updated method for knowledge discovery, is achieving a resounding success in e-commerce nowadays [3]. Taobao, a Chinese on-line shopping platform, has the ability to learn from customers' consumption behaviors and shopping records. With those data its system can apply models and corresponding algorithms to predict. The prediction from models provides recommendation of products for customers. From consumer's perspective, when confronting overwhelming items online, users can readily find and purchase desirable items they like with the help of recommender system in Taobao [4]. Obviously, one of the key challenges of a recommender system should be how to offer accurate recommendations with high confidence. In order to achieve this goal, the optimization of algorithms and repeated experiments will be unavoidable steps.

One application of recommender systems is in entertainments. Watching movie has become a popular entertainment and an increasingly number of audiences are getting used to posting their remarks about movies online [5]. In this circumstance, the database of movie ratings is created and its volume is expanding rapidly. Relying on the recommendation system, business corporations will gain lots of commercial benefits from their efforts on creating sophisticated algorithms and models. Netflix emphasized on the importance of a better recommender system by setting up the Netflix Price, which is a contest providing 1 million dollars for the team that can make the best improvement in recommender system [5]. To put it simply, the developers should consider what kind

The associate editor coordinating the review of this manuscript and approving it for publication was Fabrizio Messina.

of information the databases can indicate and how to obtain these information by using a simpler knowledge discovery method. Moon, S. (2010) analyzed the movie database and revealed the relationship between movie genres and movie ratings. He concluded that sequel movies can achieve lower rating than the original ones based on the decline of viewers' interests in sequels [5], [6]. The database delivers the information that original and innovative movies can be more attractive so that they inspire movie firms to produce more original works. However, such a conclusion is too general to give detailed and accurate recommendations since it does not consider the fact that individuals possess diverse traits and tastes.

Traditional recommender systems for movies base on users' watching history. For example, if *The Avengers* is in one's watching history, then some similar movies might be recommended to this user. The similarity can be explored in various perspectives: when it comes to names, it is possible to recommend *The Avengers 2* and relate to sequential movies; from producers' standpoint, they are more likely to offer movies manufactured by Marvel. We can even find something from genre angle: some science fiction movies, such as *Interstellar*, are likely to be recommended. Such a recommendation is based on *profile inference*. And currently there are various methods for profile inference, ranging from collaborative filtering [7], [8], knowledge graph embedding to heterogenous information network [9].

However, it is difficult to recommend movies to new users with no watching history. Such difficulty is referred to as the *cold start problem*. The introduction of linear mixed model (LMM) can help us with predicting the preference of these new users if we have some information other than the watching history [10]. Thinking about the favorite movies of people of different ages, we may have the intuition that people of different ages are likely to show distinct tastes or preferences. For example, the adventure or science fiction movies may be teenagers' favorites whereas the senior possibly prefer comedy films. Such group profiles/preferences once available can be used to solve the *cold start problem*. Such group preferences are usually hypothesized and then tested by rigorous statistical inference procedures. These tested and verified hypotheses can be used for various business development. After mining millions of movie rating data, we can provide statistically significant evidence that the youth on average are pickier and more censorious than the senior when rating the same movie [11].

It should be noted that the linear mixed model has long been used for knowledge/science discovery in breeding [12], and now widely used in ecology [13], genetics [14], [15], and genome-wide association study [16], [17]. However, the application of LMM in recommendation systems was examined just recently by sparse publications, for example [10], [11], [18]. The underlying mathematics foundation and theoretical details were inadequate or even omitted in all such recommendation literatures. In addition, so far the corresponding experiments are somewhat incomplete since some essential constituents, such as computational costs and interval estimation, are missing. In light of this, our main contribution in this paper is to provide relevant theoretical details based on our understanding of linear mixed models [20], [21], [23], [25], as well as to show the comprehensive experimental process. Some newly introduced techniques, such as Wald confidence interval, will be applied to show the credibility of our conclusion. In this way, we will show the power of LMM and related packages in handling *cold start problem*.

In the remaining part of this paper, we shall first introduce the basic idea of linear mixed models, and then discuss how to estimate the parameters in the model, how to do interval estimate, how to select a proper model, and finally we shall apply this model to analyze movie rating data.

## II. LINEAR MIXED-EFFECTS MODEL (LMM)

The traditional linear model is written as

$$Y = X\beta + \varepsilon, \tag{1}$$

where $Y$ is the $n \times 1$ vector of responses, $\beta$ is the $p \times 1$ vector of fixed effects, $X$ is the $n \times p$ matrix of fixed effects, $\varepsilon$ is the $n \times 1$ vector of random errors. There are three important assumptions on traditional linear models:

1) Normality, which means responses in vector $Y$ follow normal distribution from population.
2) Independence, which means responses are independent so that all their correlation coefficients are zero.
3) Homogeneity of variance, which means every response has the same variance.

However, when we analyze the actual data, it is very common to find limitations of traditional linear models because the data sets mostly violate these assumptions.

The linear mixed-effects model can be used to overcome these limitations by introducing additional random effects. Besides, it suffices to take account of the correlation of observations contained in a data set. Therefore, it is reasonable to model the relationship between users' traits and ratings of movies.

### A. THE STRUCTURE OF LMM

Linear mixed-effects model (LMM) is a more useful and realistic model to analyze real data sets. It is also called Hierarchical Linear model and as its name implies, this model divides data sets into several levels according to certain grouping factors [26]. For multilevel data, we are able to show the expression of the classical linear mixed-effects at a given factor as follow:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{2}$$

where $y_i$ is a $n_i \times 1$ vector of responses, $X_i$ is a $n_i \times p$ design matrix of fixed effects, $\beta$ is a $p \times 1$ vector of fixed effects, $b_i$ is a $q \times 1$ vector of random effects in factor $i$, $Z_i$ is another $n_i \times q$ design matrix of covariates which shows the correlation between responses $y_i$ and random effects $b_i$, $\varepsilon_i$ is the vector of residual errors for factor $i$. What should be emphasized is

that $Z_i$ contains known values of $q$ covariates corresponding to $q$ random effects chosen from its distribution [26]:

$$Z_i = (z_i^{(1)}, z_i^{(2)}, \cdots, z_i^{(q)}). \tag{3}$$

Moreover, $b_i$ is unobservable. It implies that random effects lack patterns, which causes difficulties for researchers to figure out its real value.

In a LMM, observations are considered not necessarily independent and have heteroscedasticity. The correlation between each pair of observations in the same level is reflected in the distribution of $b_i$ and $\varepsilon_i$. Since they are in the same level, they are able to follow bivariate normal distribution:

$$b_i \sim N(0, G), \quad \varepsilon_i \sim N(0, R_i), \tag{4}$$

where $b_i$ is independent of $\varepsilon_i$. Moreover, $G$ and $R_i$ can be specified as:

$$G = \sigma^2 G, \quad R_i = \sigma^2 R_i, \tag{5}$$

where $G$ and $R_i$ are variance functions which represent weights of each observation's variance decided by parameter $\theta_G$ and $\theta_{R_i}$ respectively. Therefore, when random effect $b_i$ is known, then the conditional distribution of $y_i$ can be formulated as:

$$E[y_i|b_i] = \mu_i = X_i\beta + Z_i b_i, \tag{6}$$
$$Var[y_i|b_i] = \sigma^2 R_i. \tag{7}$$

When $b_i$ is not given, the unconditional distribution of $y_i$ can be defined as:

$$E[y_i] = X_i\beta,$$
$$Var[y_i] = \sigma^2[Z_i G Z_i' + R_i].$$

Combing data sets from all factors, we can get the classical formula of LMM for all data:

$$Y = X\beta + Zb + \varepsilon, \tag{8}$$

where $Y = (y_1', y_2', \cdots, y_N')'$ is the $n \times 1$ vector of responses, where $n_1 + n_2 + \cdots + n_N = n$, $\beta$ is the $p \times 1$ vector of fixed effects, $X$ is the $n \times p$ design matrix for fixed effects, $Z$ is the $n \times (q_1 + q_2 + ... + q_N)$ matrix of random effects, $b$ is the $(q_1 + q_2 + ... + q_N) \times 1$ vector of random effects, where $b = (b_1', b_2', \cdots b_N')$, $\varepsilon$ is the $n \times 1$ vector of errors, $\varepsilon = (\varepsilon_1', \varepsilon_2', \cdots, \varepsilon_N')'$. Therefore, the unconditional distribution and conditional distribution can be expressed respectively as

$$Y \sim N(X\beta, \sigma^2(ZGZ' + R), \tag{9}$$

and

$$Y|_b \sim N(X\beta + Zb, \sigma^2 R). \tag{10}$$

This model can take level influence as random effects so that $Y$ can be expressed in a multivariate normal distribution form. On the other hand, according to $Y$'s variance presented in variance-covariance form, we can know that observations in $Y$ are not independent and error terms are also divided into different levels, better considering real data's features.

Linear mixed-effects models have been widely used in software, such as SAS, SPSS, Matlab as well as R. This article will focus on linear mixed-effects models using R and the lme4 package [34] to discover knowledge.

## III. PARAMETER ESTIMATION

In order to understand how linear mixed-effects model is obtained, we need to figure out the parameters' estimation. Two common ways to estimate parameters in linear mixed-effects model are maximum likelihood (ML) estimation and restricted maximum likelihood (REML) estimation [26]. The conditional distribution of $y_i$ given $b_i$ is not appropriate for constructing the likelihood function since we don't know the real value of random effects $b_i$. Therefore, marginal distribution of $y_i$ is applied to build up ML and REML function.

### 1) MAXIMUM LIKELIHOOD ESTIMATION

Summarizing the parameters contained in linear mixed-effects model above, we get three types of parameters: the fixed effects $\beta = (\beta_1, \beta_2, \cdot \cdot \cdot, \beta_p)^T$; the random effects $b = (b_1, b_2 m \cdots, b_p)^T$; and the variance parameters $\sigma^2$, $\theta = (\theta_G, \theta_R)$. Their estimators can be obtained by simultaneously maximizing the log-likelihood function with respect to these parameters. However, it is a numerically complex work which needs to find an optimum in a multidimensional parameters space. Fortunately it can be simplified by profile likelihood technique.

With parameters $\beta, \sigma^2, \theta$, we have likelihood expression given that:

$$\begin{aligned} &L_{ML}(\beta, \sigma^2, \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma\sqrt{det(V_i)}} exp\left\{-\frac{1}{2}\frac{(y_i - X_i\beta)^2}{\sigma^2 det(V_i)}\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \prod_{i=1}^{n} exp\left\{-\frac{1}{2}\frac{(y_i - X_i\beta)^2}{\sigma^2 det(V_i)}\right\} \end{aligned} \tag{11}$$

where $V_i = Z_i G(\theta_G) Z_i' + R(\theta_R)$. Ignoring the constant part and taking log operation, we get the log-likelihood function of the form:

$$\begin{aligned} l_{ML}(\beta, \sigma^2, \theta) = &-\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta). \end{aligned} \tag{12}$$

Assume that the variance parameters are known, the fixed effects and random effects can be determined by solving the following mixed model equations [22]:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} Z & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}.$$

In particular,

$$\hat{\beta}(\theta, \sigma^2) = (\sum_{i=1}^{n} X_i' V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i. \tag{13}$$

By plugging (13) into (12), we gain the log-profile likelihood function:

$$l^*_{ML}(\sigma^2,\theta) = l_{ML}(\hat{\beta}(\theta),\sigma^2,\theta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i' V_i^{-1} r_i. \quad (14)$$

where $r_i = r_i(\theta) = y_i - X_i\beta(\hat{\theta})$.

In this way, the function does not depend on $\beta$, which means the parameter space has lower dimension than previous one. Then use the same method, maximizing $l^*_{ML}(\sigma^2,\theta)$ with respect to $\sigma^2$ for every known value of $\theta$ leads to the estimation of $\sigma^2$:

$$\hat{\sigma}^2_{ML}(\theta) = \sum_{i=1}^{n} r_i' V_i^{-1} r_i/n. \quad (15)$$

By plugging (13) into (14), we get a log-profile likelihood function for $\theta$:

$$l^*_{ML}(\theta) = l^*_{ML}(\hat{\sigma^2}_{ML},\theta) = -\frac{n}{2}\log(\hat{\sigma^2}_{ML}) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{n}{2}. \quad (16)$$

Therefore, there are fewer parameters in the parameter space again. Then maximization of $l^*_{ML}(\theta)$ can yield an estimator $\hat{\theta}_{ML}$ of $\theta$. Plugging $\hat{\theta}_{ML}$ into (12) and (14) produces estimator $\hat{\beta}_{ML}$ of $\beta$ and $\hat{\sigma}^2_{ML}$ of $\sigma^2$ that:

$$\hat{\beta}_{ML} = \hat{\beta}(\hat{\theta}_{ML}) = (\sum_{i=1}^{n} X_i' \hat{V}_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' \hat{V}_i^{-1} y_i, \quad (17)$$

$$\hat{\sigma^2}_{ML} = \hat{\sigma^2}_{ML}(\hat{\theta}_{ML}) = \sum_{i=1}^{n} r_i' \hat{V}_i^{-1} r_i/n. \quad (18)$$

However, there is a significant limitation on maximum likelihood estimation. ML estimators $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ are both biased because they don't adjust for the uncertainty in estimation of $\beta$. In other words, the values of $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ may vary with the change of $\beta$ so that we cannot make accurate estimations of these two parameters. However, $\sigma^2$ and $\beta$ can be better estimated by restricted maximum likelihood estimation, which will be discussed in next section.

#### 2) RESTRICTED MAXIMUM ESTIMATION

In order to obtain unbiased estimates of $\sigma^2$ and $\theta$, we should use an estimation that is orthogonal to the estimation of $\beta$, which means using a way to make estimates of $\sigma^2$ and $\theta$ that are independent of estimation of $\beta$ [6]. To achieve this goal, we can consider the likelihood function of a set of $n-p$ independent contrasts of $y$, where $p$ is the dimension of $\beta$. After obtaining $\hat{\beta}(\theta)$, the restricted log likelihood function is given by:

$$l^*_{REML}(\sigma^2,\theta) = -\frac{n-p}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i' V_i^{-1} r_i - \frac{1}{2} \times \log[\det(\sum_{i=1}^{n} X_i' V_i^{-1} X_i)]. \quad (19)$$

From this function, maximizing of $l^*_{REML}(\sigma^2,\theta)$ with respect to $\sigma^2$ leads to an estimator of $\sigma^2$ that:

$$\hat{\sigma^2}_{REML} = \sum_{i=1}^{n} r_i' V_i^{-1} r_i/(n-p). \quad (20)$$

Plugging (18) into (19), we get a function with respect to $\theta$ only:

$$l^*_{REML}(\theta) = -\frac{n-p}{2}[\log(\sum_{i=1}^{n} r_i' V_i^{-1} r_i/(n-p)) + 1] - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{1}{2} \times \log[\det(\sum_{i=1}^{n} X_i' V_i^{-1} X_i)]. \quad (21)$$

Estimator of $\theta$ can be obtained by maximization from (21), which can be applied to get estimators of $\beta$ and $\sigma^2$, respectively.

## IV. CONFIDENCE INTERVAL FOR LMM

Confidence Interval (CI) gives a range for a random variable based on a certain confidence level, that is, the credibility of the estimator. Therefore, people can consider the values in this confidence interval to have the same or similar influence. Constructing confidence interval has significant meanings either in theory or empirical problem. When estimating parameters or predicting responses, we can get certain values of parameters or responses. Admittedly, it can not be denied that when considering the preciseness of science and mathematics, these values undoubtedly lack certainty or accuracy. However, if we can locate estimation and prediction in some certain ranges with confidence level attached, the conclusion tend to be more convincing.

There are three methods to compute the confidence intervals: the profile likelihood confidence interval, Wald confidence interval and bootstrap confidence interval. Each method possesses disparate ideas and assumptions. Next we will discuss the underlying concepts of these CIs and their applications in LMMs.

#### 3) PROFILE LIKELIHOOD CONFIDENCE INTERVAL

One of the main conditions of profile likelihood confidence interval (PLCI) is that the estimator does not necessarily have to follow normal distribution [27]. The concept of PLCI is

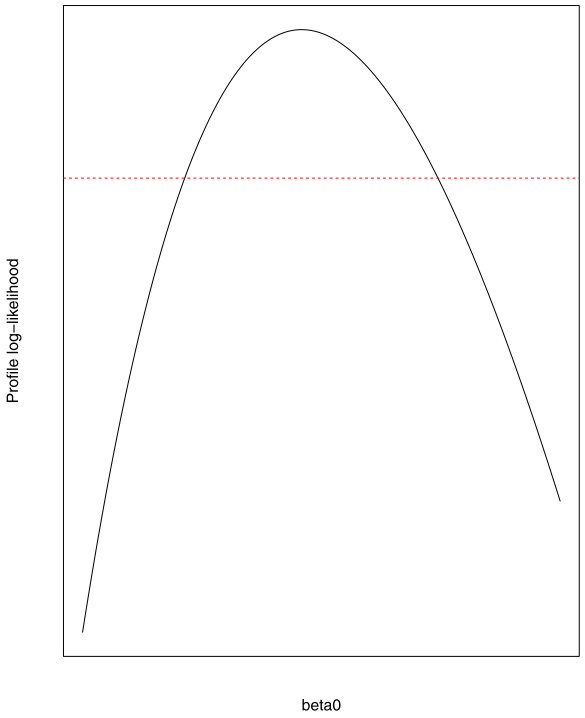



**FIGURE 1.** Profile likelihood function for $\beta_0$.

very similar to the profile likelihood technique previously mentioned. In a model, we assume $\beta$ is our interest parameter and **b** is the vector of all nuisance parameters. Thus, $L(\beta, b)$ is the maximum likelihood function based on two random variables $\beta$ and **b**, also the profile likelihood function of $\beta$ is defined as

$$L_1(\beta) = \max_b L(\beta, b), \tag{22}$$

which means the maximum likelihood function of $\beta$ with *MLE* value of $b$.

With this concept, we can consider the confidence interval next. In the hypothesis test, the null hypothesis is constructed like this: $H_0 : \beta = \beta_0$. In this circumstance, building a confidence interval is equivalent to finding all $\beta_0$, which can make the $H_0$ not be rejected under the $100(1-\alpha)\%$ confidence level. Then we use the *likelihood ratio test* [27]:

$$2[\log L(\hat{\beta}, \hat{b}) - \log L_1(\beta_0)] < \chi^2_{1-\alpha}(1), \tag{23}$$

where $L(\hat{\beta})$ is the maximum likelihood with *MLE* of all parameters and $L_1(\beta_0)$ handles one fewer parameters, that's why the left hand side of this formula follows Chi-square distribution. Therefore, all $\beta_0$ satisfying above formula can form a confidence interval for $\beta$. Since $\log L(\hat{\beta})$ and $\chi^2_{1-\alpha}(1)$ are constant, we can rearrange the expression:

$$\log L_1(\beta_0) > \log L(\hat{\beta}, \hat{b}) - \chi^2_{1-\alpha}(1)/2. \tag{24}$$

It is likely to get a graph like this below. Therefore, the part of the curve above the red line forms the confidence interval we want.

#### 4) WALD CONFIDENCE INTERVAL

Wald confidence Interval takes Wald Test into account:

$$\frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \sim N(0, 1), \tag{25}$$

with the assumption that the difference between the two will be approximately normally distributed. According to this test, when considering the confidence Interval in LMM, we only need to know the estimates and variance of the parameters included. For the fixed effects $\beta$, we get the estimate above and its variance-covariance that:

$$\hat{\beta}_{ML} = (\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1} \sum_{i=1}^{n} X'\hat{V}^{-1}y, \tag{26}$$

$$\text{varcov}(\hat{\beta}) = \hat{\sigma^2}(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1}. \tag{27}$$

Thus, we extract the diagonal of variance-covariance matrix as variance of $\hat{\beta}$, the wald confidence interval for $\hat{\beta}$ can be expressed as follow:

$$(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1} \sum_{i=1}^{n} X'\hat{V}^{-1}y \pm z_{\alpha/2} \sqrt{\text{diag}(\hat{\sigma^2}(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1})}. \tag{28}$$

#### 5) BOOTSTRAP CONFIDENCE INTERVAL

Bootstrap confidence interval comes from the idea that "pulling itself up by its own bootstrap" [28]. In other words, it means doing large number of bootstraps from the original data. Assume we have original data $\{y_1, y_2, y_3, \ldots, y_n\}$, and we build LMM for this data which contains parameter $\Theta = (\beta, \theta, \sigma^2)$. After maximum likelihood estimation or restricted maximum likelihood estimation, we are able to obtain estimates $\hat{\Theta} = (\hat{\beta}, \hat{\theta}, \hat{\sigma^2})$. In order to find the confidence interval, we should calculate the variation of $\hat{\Theta}$ around $\Theta$, that is, $\delta = \hat{\Theta} - \Theta$. Hence, confidence interval based on $\alpha\%$ confidence level can be shown as:

$$\Theta \in [\hat{\Theta} - \delta_{\alpha/2}, \hat{\Theta} + \delta_{1-\alpha/2}]. \tag{29}$$

To find $\delta$, we process bootstrap operation. Firstly, we take resamples from the original data $\{y_1, y_2, y_3, \ldots, y_n\}$ and receive $n$ new observations notated as $\left\{y_1^{(1)}, y_2^{(1)}, \ldots, y_n^{(1)}\right\}$ which has the same distribution as the original data. After the LMM estimation for the new data, we obtain new parameter estimates $\hat{\Theta}^{(1)}$ and the first variation is calculated as $\delta^{(1)} = \hat{\Theta}^{(1)} - \hat{\Theta}$. Then this kind of resampling operation should run repeatedly for $m$ times, normally more than 1000 times and a matrix of resample can be formed:

$$\begin{bmatrix} y_1^{(1)} & y_1^{(2)} & \ldots & \ldots & y_1^{(m)} \\ y_2^{(1)} & y_2^{(2)} & \ldots & \ldots & y_2^{(m)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ y_n^{(1)} & y_n^{(2)} & \ldots & \ldots & y_n^{(m)} \end{bmatrix}. \tag{30}$$

Each column represents one resampling and produces one $\delta$, thereby a sequence of $\delta$ is generated finally $\{\delta^{(1)}, \delta^{(2)}, \delta^{(3)}, \ldots, \delta^{(m)}\}$ and we sort them from the smallest to the biggest. Thus, $\delta_{\alpha/2}$ is at the $\alpha/2$ percentile and $\delta_{1-\alpha/2}$ is at the $1-\alpha/2$ percentile. In this way, the confidence interval of parameters can be calculated. Bootstrap introduces us a simple and straightforward angle to observe the variation of estimates. With the law of large number, the resample distribution can be a good approximation to the true distribution.

#### 6) COMPARISON

Since these three methods have different concepts, we should consider carefully about which method should be applied in different circumstances.

Profile confidence interval has very wide applications because of its moderate restriction, so it is still available for the estimators not normally distributed [29]. Due to this reason, the default demand *confint*() in R software adopts this profile likelihood method and it is useful when analyzing LMMs. Although Wald confidence interval is very common, it is difficult to apply such a method in LMM because it is not feasible to account for the parameters contained in random effects, that is, $\sigma^2$ and $\theta$. The Wald confidence interval cannot be calculated for these parameters in LMM. As for bootstrap confidence interval, the CI from this method is relatively valid. This kind of sampling cannot improve point estimates. It is obvious that every bootstrap is chosen from the same data pool and follows the same steps, there is no new information reflected even after all bootstraps [28]. This is also the reason why confidence interval calculated from bootstrap is more wider than profile likelihood method and Wald test, so bootstrap confidence interval is hardly used due to its less preciseness.

## V. CRITERIA FOR MODEL SELECTION

As for a data set, there might be several models applied to analyze it. Then how to measure which model is better has become a main problem. There are some important and useful criteria to evaluate those models, such as log-likelihood, Akaike information criterion (AIC), Bayesian information criterion (BIC) and p-value. Obviously, different models possess different focus and purposes which result in different values.

#### 7) LOG-LIKELIHOOD

It is the simplest criterion which has the expression shown below.

$$l(\Theta) = \log L(\Theta) = \log\left(f(Y|\Theta)\right), \tag{31}$$

where $Y = (y_1, y_2, \cdots, y_n)'$ is the vector of observations; $\Theta = (\beta, \sigma^2, \theta)$ represents the vector of all parameters contained in linear mixed-effects model where $\beta = (\beta_1, \beta_2, \cdots, \beta_k)$ is the vector of fixed effects and $\theta = (\theta_G, \theta_R)$ is the vector of random effects; $f(Y|\Theta)$ is the likelihood function of observations.

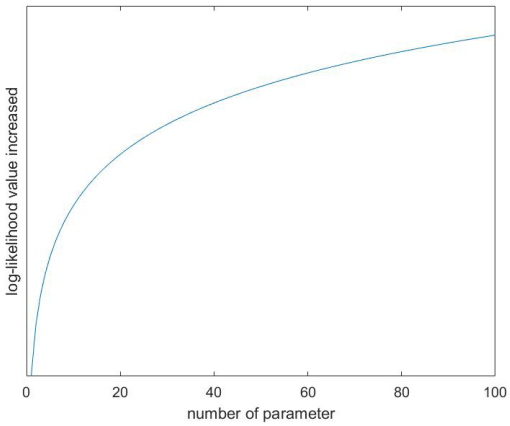



**FIGURE 2.** Relationship between log-likelihood value and parameter.

According to this expression, we can see clearly the design of this criterion: $\Theta$ is the key component of our model so that it can represent our model to some extent, $Y$ is the data set observed. Therefore, $f$ can quantify how much the data fits our candidate models and adding log is to avoid zero value in the likelihood [30]. The higher value of the log-likelihood achieves, the better the data fits our models.

However, log-likelihood criterion has a huge problem that it doesn't consider the number of parameters. Ideally we prefer a model with high log-likelihood and small number of parameters. Such models can not only guarantee high level of fitting but also require less calculations and operations. Normally, larger number of parameters would increase the value of log-likelihood and the effects are significant when the number of parameters is few. However, when the number of parameters is large enough, each parameter added in model would has little influence to log-likelihood value, which is shown FIGURE 2.

That is, it makes little sense to have too many parameters. On the contrary, we prefer a model with fewer parameters when its log-likelihood value is just a little bit less than the value of the model with more parameters. In this circumstance, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) both consider the number of parameters and overcome log-likelihood's limitation.

#### 8) AKAIKE INFORMATION CRITERION (AIC)

Akaike information criterion is a standard to measure the goodness of statistical model fitting. It was found and developed by Japanese statistician Akaike. This criterion suffices to quantify the complexity of the estimated model and the goodness of the fitting data of the model. When we use maximum likelihood estimation (MLE) in linear-mixed effect model, log-likelihood $l(\hat{\Theta})$ can be achieved where $\hat{\Theta} = (\hat{\beta}, \hat{\sigma^2}, \hat{\theta})$ contributes to maximum log-likelihood.

AIC can be expressed as $l(\hat{\Theta}) - p$. However, normally, in R software package, the AIC formula is defined as

$$AIC = 2p - 2l(\hat{\Theta}), \tag{32}$$

where $p$ is the number of parameters, $l(\hat{\Theta})$ is the maximum log-likelihood. Both expressions have the same structure with slight difference including sign and multiplies. To be clear, in this essay, we use formula (32). The lower AIC value a model has, the better the model is.

#### 9) BAYESIAN INFORMATION CRITERION (BIC)

In statistics, there are two ways to optimize models. On the one hand, adding more parameters to models can increase their complexity. On the other hand, collecting more observations or data suffices to improve models' ability to describe data sets. AIC considers the parameter problems whereas the number of observations is not included. However, BIC considers both of them and takes them as measurement for models.

BIC provides an algorithm to approximate the log marginal likelihood of candidate models and chooses the one having smaller value as the better model. The formula of BIC is:

$$BIC = p \cdot \log n - 2l(\hat{\Theta}), \tag{33}$$

where $p$ is the number of parameters, $n$ is the number of observations.

#### 10) F-TEST

Sometimes AIC and BIC will give different choices between two models. So under this circumstance, we need to refer to other criteria. F-test is able to calculate p-value which can be used to judge whether these two models are significantly different and make decisions about model selection. The logic behind F-test is the comparison of models' deviance which is defined that:

$$D = 2[l(\hat{\Theta}_{max}; y) - l(\hat{\Theta}; y)], \tag{34}$$

where $\hat{\Theta}_{max}$ is the MLE for the parameter vector in the saturated model which has $N$ parameters, $\hat{\Theta}$ is the MLE for the parameter vector in the candidate model. Assume there are two models $m0$ and $m1$ with degree of freedoms, $p$ and $q$ respectively where $p < q$ and the set of parameters of $m1$ contains $m0$'s parameters. We can calculate their deviance denoted as $D_0$ and $D_1$ which are applied for $F$ test:

$$F = \frac{(D_0 - D_1)/(p - q)}{D_1/(N - q)}. \tag{35}$$

After we obtain the $F$ value, p-value also can be achieved by referring to F table. Next the process of making decisions depends on our own confidence degree to this test. Normally 95 percent confidence level and 99 percent confidence level are preferred choices, so based on 95 percent confidence level, if the p-value is less than 5 percent, there is significant difference between $m0$ and $m1$. Because of the 'useful' parameters in $m1$, we tend to choose $m1$ with more parameters. On the contrary, when the p-value is more than 5 percent, there is no significant difference between these two models so that the model with fewer parameters $m0$ is our choice.

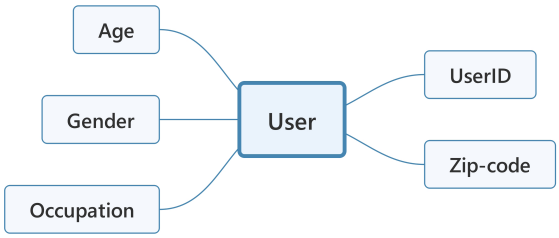



**FIGURE 3.** Data structure of user and movie.

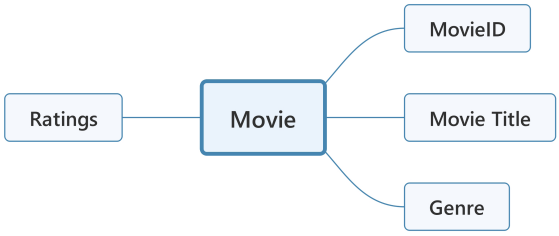



**FIGURE 4.** Data structure of user and movie.

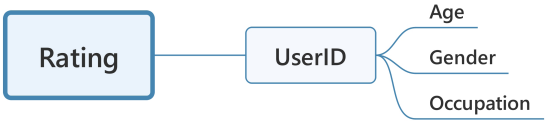



**FIGURE 5.** Data structure.

## VI. APPLICATION:KNOWLEDGE DISCOVERY FROM HIGH-THROUGHPUT MOVIE RATING DATA

Understanding the knowledge of model selection criteria and parameter estimation, firstly we need a database about detailed information of users, including age, gender, occupation, nation, hobby, etc. Moreover, some movie information, such as movie names, movie genres and movie ratings, is also the necessity. For this purpose, The ***ml-1m*** data set is used in this experiment. It provides us with 6040 users, 3952 movies and movie rating made on a 5-star scale. Among these ratings, each user has at least 20 ratings. Furthermore, users' features including age, gender, occupation and zip-code are all in category types. Movie information including Movie ID, title and genres is listed and movie genre can be found in **Appendix**. Therefore, this is a real big data and the pattern excavated from it will be highly convincing. Although some false information exists in the data set, the huge volume allows us to ignore its influence on our data analysis. The data structure is shown in FIGURE 3 and FIGURE 4.

Our task is to discover the age influence on movie rating, thereby we need to rearrange the data provided above. Since UserID from User data set is linked with the ratings in movie data set, we can ignore zip-code which makes little sense in this experiment and construct a data chain aiming for one movie or types of movies (see FIGURE 5).

According to this data structure, we can construct linear mixed-effects models. Since age is our interest variable, it is treated as fixed effect. In addition, gender and occupation are our candidate random effects. Three different models in Table 1 can be designed.

**TABLE 1.** Three models.

| Model | Fixed effect | Random effects |
|---|---|---|
| Model1 | Age | Gender |
| Model2 | Age | Occupation |
| Model3 | Age | Gender+Occupation |

| Comparison | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Random Effects | Gender+Occupation | Occupation | Gender |
| Degree of freedom | 10 | 9 | 9 |
| AIC | 3067.9 | 3067.2 | 3065.9 |
| BIC | 3118.2 | 3112.7 | 3111.3 |
| Loglik | −1523.9 | −1524.6 | −1524 |
| p-value | Model1 with Model2:0.224 | Model1 with Model3:0.68 | |

**FIGURE 6.** Results for model selection criterions.

As for how to choose the optimal candidate from these models, we should focus on the real data analysis as well as our model selection criteria.

Given the data as well as candidate models above, our study on age discovery starts from one single movie to different types of movies. During this process, we are willing to discover the implicit general pattern.

### A. ONE SINGLE MOVIE: LIFE IS BEAUTIFUL

Firstly, we are willing to see the relationship between users' ages and ratings for one specific movie. Here we choose the film *Life Is Beautiful* labeled as comedy and romantic movie which won the best foreign language film at the 71st Oscar Awards. Since it is a known work, it received lots of rating records from users so that there are 1152 ratings available.

In our model, user's age is regarded as our interest, i.e. the fixed effect and we need to consider how to choose random effects based on the models provided above. When we try to build candidate models for comparison, the computational time for constructing Model 1, Model 2 and Model 3 are 0.036s, 0.025s and 0.026s respectively, which are fairly short. After using **Anova()** test, the values of model selection criteria are calculated in FIGURE 6.

According to these results, Model 3 has the lowest AIC and BIC value so Model 3 is the best model based on AIC and BIC criteria, but Model 1 obtains the largest log-likelihood value. Literally, Model 1 is the best one referring to log-likelihood and it has more parameters. However, after the comparison on p-value, we can make our decision. Under 95% confidence interval, the p-value of Model 1 and Model 2 is larger than 0.05 which means there is little difference between Model 1 and Model 2. Since Model 2 has the fewer parameters, it is better than Model 1. Due to the same reason, Model 3 is better than Model 1. In the meantime, observing from AIC, BIC and log-likelihood criteria, we can see the superiority of Model 3 compared with Model 2. To conclude, Model 3 is the best model for the data *Life is beautiful*.

***lmer*** command is capable of estimating the parameters contained in models and we can have the result in FIGURE 7.

Results in FIGURE 7 show the standard deviation of random effect, gender and residual as well as what we desire: fixed effects of all age levels. To be clear, the radar chart

```
Random effects:
 Groups    Name        Std.Dev.
 GenderLIB (Intercept) 0.08759
 Residual              0.90991
Number of obs: 1152, groups:  GenderLIB, 2
Fixed Effects:
 AgeLIB1  AgeLIB18  AgeLIB25  AgeLIB35  AgeLIB45  AgeLIB50  AgeLIB56
   4.333     4.568     4.399     4.223     3.961     4.194     4.256
```

**FIGURE 7.** Results for parameter estimation.

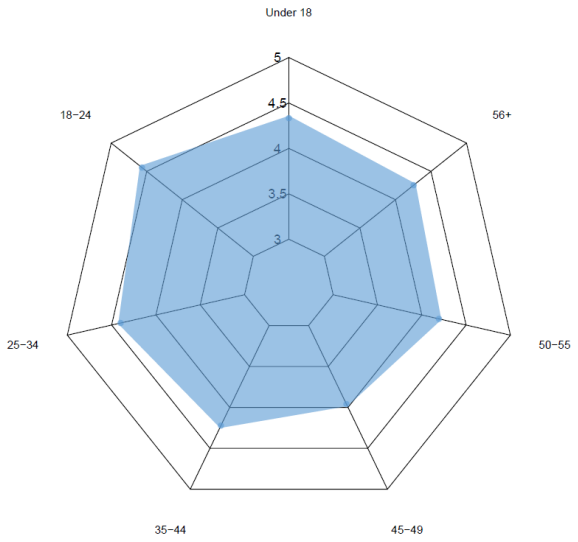



**FIGURE 8.** Age effect distribution for *Life Is Beautiful*.

| Age group | Under 18 | 18-24 | 25-34 | 35-44 | 45-49 | 50-55 | 56+ |
|---|---|---|---|---|---|---|---|
| Age effect | β 1 | β 2 | β 3 | β 4 | β 5 | β 6 | β 7 |

**FIGURE 9.** Age effects expression.

```
             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
ModelLIBtest  8 3069.5 3109.9 -1526.8   3053.5
ModelLIB      9 3065.9 3111.3 -1524.0   3047.9 5.6018      1    0.01794
```

**FIGURE 10.** Anova analysis of $\beta_2$ and $\beta_3$.

in FIGURE 8 is able to show the fixed effects and their differences intuitively.

In this graph, visually we can see the rating from users in '18-24' group is much higher than those from other age groups. However, this deduction is unreasonable since we are not sure about whether the rating of users from 18-24 is significantly higher than others. To judge the significance of differences, we need to operate hypothesis test. The effects of every age group, that is, the parameters of age levels can be represented in FIGURE 9.

Therefore, our task is to prove whether $\beta_2$ is significantly the largest effect. From the result in FIGURE 6, $\beta_3$ is the closest one to $\beta_2$ which leads to our hypothesis test that:

- $H_0 : \beta_2 = \beta_3$
- $H_1 : \beta_2 \neq \beta_3$

$H_1$ represents the original model and we can replace all data belonging to 25-34 age level to 18-24 age level (i.e. combining $\beta_2$ and $\beta_3$ together), which brings new model. We compare these two models and obtain result in FIGURE 10.

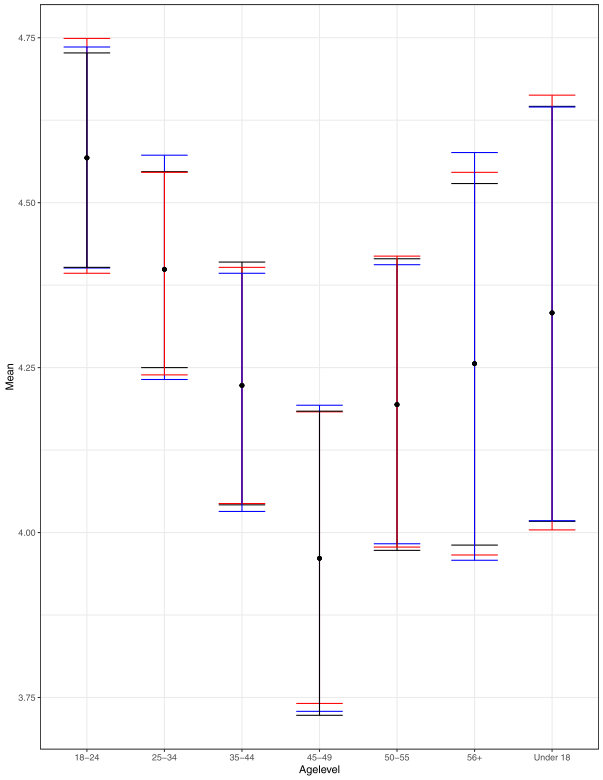



**FIGURE 11.** Three types of CI for LIB.

The low $p$-value indicates that it is significant to reject $H_0$ under 95% confidence interval and there is a big difference between $\beta_2$ and $\beta_3$, thereby the conclusion that 18-24 users give higher ratings for *Life is Beautiful* than any other age groups can be drawn.

Besides observing the estimates of age effects, we are able to use package *lm4* to calculate the confidence interval of age effects. Based on three types of confidence interval (profile CI, Wald CI, bootstrap CI), their ranges can be expressed respectively in one graph as shown in FIGURE 11 and FIGURE 12.

As shown in FIGURE 12, the differences among these three confidence intervals are negligible so that the ranges shown in the picture almost overlap with each other. As a result, it is reasonable for us to treat all confidence intervals as the same. Therefore, taking the Wald confidence interval, we are capable of building a radar chart for CI of age effects where the area between red line and green line is the CI place showed above.

Back to FIGURE 12, it is not difficult to find that the CI ranges of 18-24, 25-34, 35-44 are narrower than those for 45-49, 50-55, 56+ and under 18. According to formula of the Wald confidence interval, it can be deduced that the standard deviations of age effects of 18-24, 25-34, 35-44 are smaller than those of 45-49, 50-55, 56+, under 18. The narrower a confidence interval is, the more new information is reflected. Therefore, the age effects of 18-44 for *Life Is Beautiful* are more meaningful and it can better display the range where the true age effects lie.

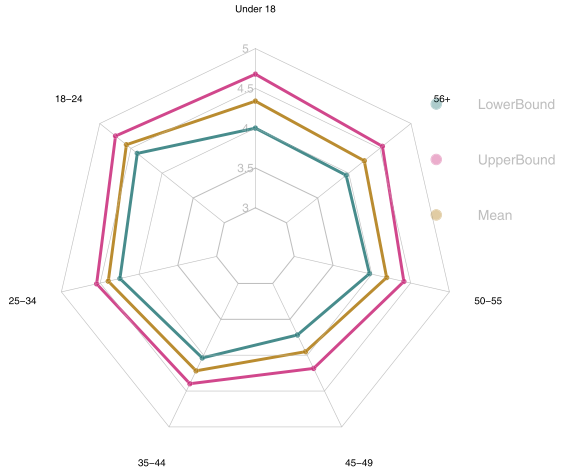



**FIGURE 12.** Confidence Intervals of *Life Is Beautiful*: profile CI (black), Wald CI (red) and the bootstrap CI (blue).

| Comparison | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Random Effects | Occupation+Gender | Occupation | Gender |
| Degree of freedom | 10 | 9 | 9 |
| AIC | 327679 | 327847 | 327997 |
| BIC | 327775 | 327934 | 328083 |
| Loglik | −163829 | −163915 | −163989 |
| p-value | Model1 with Model2: (2.2*10^-16) | Model1 with Model3: (2.2*10^-16) | |

**FIGURE 13.** Model selection.

Nonetheless, the analysis above only comes from one movie, which may not have a high reliability to generate a tenable conclusion for future application and reference. What we want to explore is the general pattern of different audiences' rating behaviors. Therefore, although we now have some interesting discoveries of *Life Is Beautiful*, they still have poor influence on our main target. On the other hand, if the general pattern of one specific movie genre can be excavated, then this knowledge would be more plausibly used for recommending movies for people.

### B. ONE GENRE: COMEDY

Comedy is a very big genre in **ml-1m** because it has the largest number of rating data (107009 ratings) among all movie genres. Therefore, the analysis of this big data set will be more meaningful than one certain movie. The target is still to find the relationship between age and rating whereas the research range expands to comedy, one certain genre.

The first thing we need to do is constructing candidate models. Since this time the volume of data set is much larger than the previous case, the computational time become longer: 25s, 9.6s and 8.2s for building Model 1, Model2 and Model 3 respectively. Then we still need to choose the optimal model according to the results shown in FIGURE 13:

The comparison is very clear that Model 1 has the smallest AIC and BIC values and the largest log-likelihood value. Moreover, the p-values of Model 1 with Model 2 and Model 1 with Model 3 are very small which gives the information that although Model 1 contains one more parameter than Model 2 and Model 3, the deviance of it is significantly lower than the other two. Therefore, we should choose Model 1 as our optimal model.

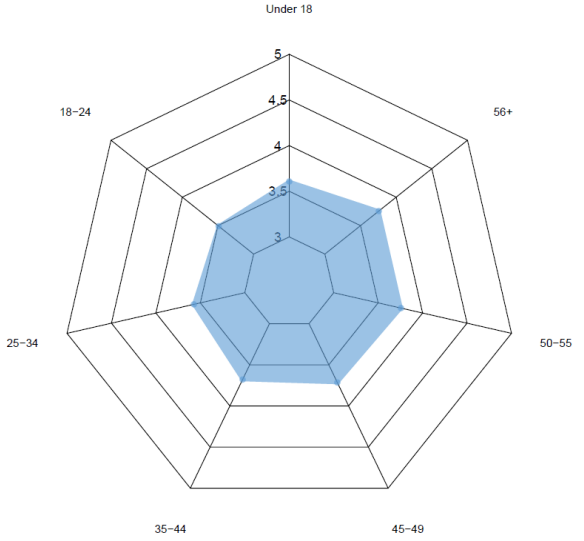



**FIGURE 14.** Age effects for comedy.

According to the result from ***lmer***, we can obtain the radar chart of age effects for comedy movies showed FIGURE 14.

Since age effect mentioned here is just the mean value of corresponding parameter, we cannot be fully confident to make an assertion that the age effect to one's rating for comedy is the number displayed in the graph above. One convincing way is to provide a confidence interval for each age effect, here we choose 99% confidence interval. Note that the comedy data is much larger than the data of *Life Is Beautiful*, that is why in this experiment we want a higher confidence level. Surprisingly, these confidence intervals here perform great differences in FIGURE 15 and we should make a choice among them. Firstly, there is a clear observation that profile confidence interval has much wider ranges for every age effects compared with the other two so that it contains less new information for the location of age effects' estimates, so we will not use this type of confidence interval here. Moreover, although the Wald confidence interval and bootstrap one look very similar, one problem on bootstrap CI is remarkable on the age level 18-24. At this age level, bootstrap CI is fairly narrow whereas it does not contain the estimate which is a serious issue indeed. Considering the previous definition of bootstrap CI, every bootstrap is random and its merit is just based on the large number of repetitions to simulate whereas no improvement is produced. Hence, it is reasonable for the existence of the estimate exclusion. Then let's have a look at the wald CI, for every age level, the range performs high proximity and it demonstrates the significance of every age effect CI is similar. Thereby, we can trust the CIs of different age levels evenly. In this circumstance, the Wald confidence interval should be taken to account for the range of age effects and its radar chart is illustrated in FIGURE 16.

From age effect of comedy graph, on the first cursory glimpse, age above 45 seems to have higher effects and people in 18-24 age group are likely to have relatively lower effect, whereas the whole effects look like a round pie.

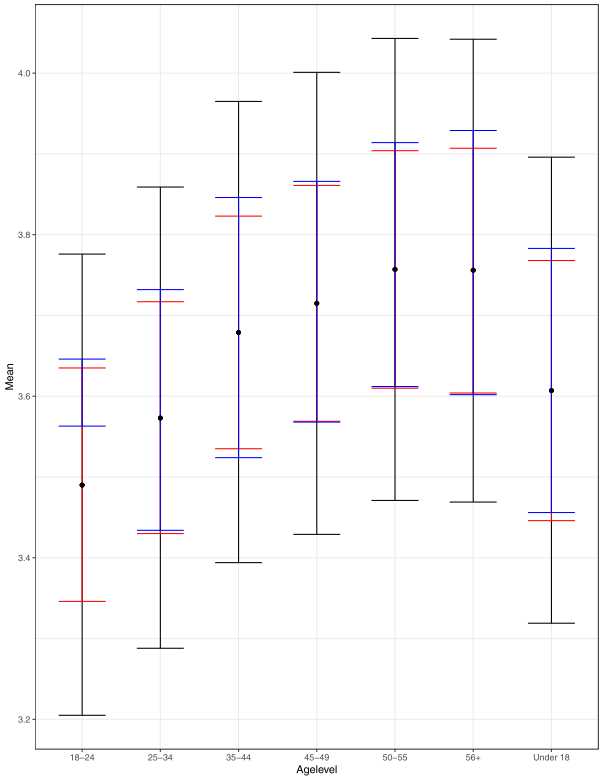



**FIGURE 15.** Confidence intervals of comedy: profile CI (black), Wald CI (red) and the bootstrap CI (blue).

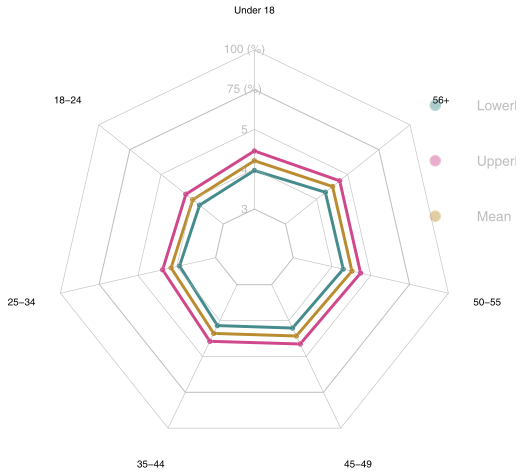



**FIGURE 16.** Confidence intervals of comedy.

To investigate the difference between levels of effects, we are trying to process hypothesis test for all differences. Since it would be a tedious task, we use a simple method to operate it.

In the first step, we choose two effects having the largest difference i.e. in our two candidate models, they will share the same parameters but one model contains one fewer parameter. Then, we use **Anova()** to compare previous model and the changed model with 99% confidence level. If the result shows they are not significant, then all differences between age effects are of no significance. On the contrary, when the result reflects significance, we need to contrast the value of

```
Models:
comedy_model2: comedy$Rate ~ -1 + agetest + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
               Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model2  9 327931 328017 -163957   327913
comedy_model1 10 327679 327775 -163829   327659 254.29      1  < 2.2e-16 ***
---
```

**FIGURE 17.** Comedy: Model 1 vs Model 2.

```
Models:
comedy_model3: comedy$Rate ~ -1 + agetest2 + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
               Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model3  9 327817 327904 -163900   327799
comedy_model1 10 327679 327775 -163829   327659 140.33      1  < 2.2e-16 ***
---
```

**FIGURE 18.** Comedy: Model 1 vs Model 3.

```
Models:
comedy_model4: comedy$Rate ~ -1 + agetest3 + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
               Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model4  9 327683 327769 -163833   327665
comedy_model1 10 327679 327775 -163829   327659 6.0828      1    0.01365 *
---
```

**FIGURE 19.** Comedy: Model 1 vs Model 4.

the second high difference and do hypothesis test again until the result shows the difference of models are not significant. In this way, we are able to sort the age effects according to significance.

In this experiment, we know the effects $\beta_{1-7}$ are 3.607, 3.490, 3.573, 3.679, 3.715, 3.757, 3.756 respectively. Therefore, we choose $\beta_6$, $\beta_2$ and do the first hypothesis test:

- $H_0 : \beta_6 = \beta_2$
- $H_1 : \beta_6 \neq \beta_2$

where $H_0$ shows comedyModel2, $H_1$ is the comedyModel1 defined previously. we can get the **Anova()** result in FIGURE 17

The p-value is quite small and less than 0.01, so we should do the second hypothesis test with $\beta_7$ and $\beta_2$:

- $H_0 : \beta_7 = \beta_2$
- $H_1 : \beta_7 \neq \beta_2$

where $H_1$ indicates comedyModel3 and we obtain the **Anova()** result in FIGURE 18. The p-value still shows the non-significance. Then, according to this rule and after several runs of tests, we come to the difference between $\beta_4$ and $\beta_5$:

- $H_0 : \beta_4 = \beta_5$
- $H_1 : \beta_4 \neq \beta_5$

where $H_0$ indicates comedyModel4 so the result is in FIGURE 19.

The p-value is larger than 0.01, so the difference is significant and loop hypothesis tests are finished. According to the tests, there are some surprising discoveries which can be mentioned:

- The differences of effects of rating from people with age more than 35 are insignificant so that they can be regard as a unit with high rating on comedy which we can notate as ''The high rating group''. In the meantime, the differences of effects of rating from people with age less than 35 are also insignificant which we notate as ''The low rating group''.
- The difference between every member from ''The high rating group'' and every member from ''The low rating group'' is significant.

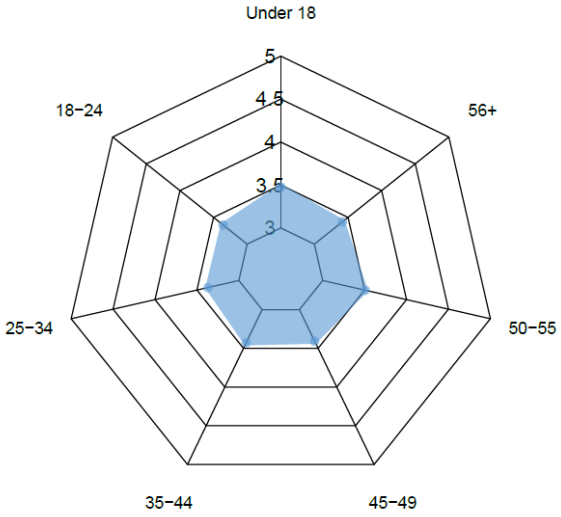



**FIGURE 20.** Age effects for sci-fi movie.

In this circumstance, we can define young people as the ones with age 1-34 and senior people as the ones with age more than 35. Therefore, the hidden pattern from comedy analysis is that **the young people are more picky and particular than the senior for comedy**. Thus, it provides a strategy for some video websites so that they can recommend more comedy movies to the senior and show this type of programs less frequently to young people. In this way video websites may receive more positive feedback. However, how about movies of other genres? We have not done researched on the movies of other types and the whole movies recorded. If they give the similar patterns in the radar chart, then it can deliver the information that people from the same age group have the same attitudes to all movies of different genre. Thus, now we continue to analyze other types of movies and observe the related outcomes.

### C. OTHERS MOVIE GENRES

Furthermore, we use linear mixed-effects models to analyze data of other genres where age is regarded as fixed effect, occupation and gender are considered as random effects. Here we choose science fiction movie, children movie and adventure movie to show in radar chart. The age effects for these models are shown from FIGURE 20 to FIGURE 23.

After hypothesis tests respectively, these three graphs all give patterns that young people who are defined as ones in age groups ''Under 18'', ''18-24'', ''25-34'' mark the movies lower than the old people which are defined as those in age groups ''45-49'', ''50-55'', ''56+''. It contributes to our idea that as for ranking movie, young people are pickier than old ones. This conclusion is also supported by comedy analysis which is displayed above. Moreover, in order to provide more evidences, rather than analyzing other types of movies, we directly investigate age effects distribution in all movies recorded, i.e 3,000,000 pieces of users' rating for movies, from which we get the chart below.

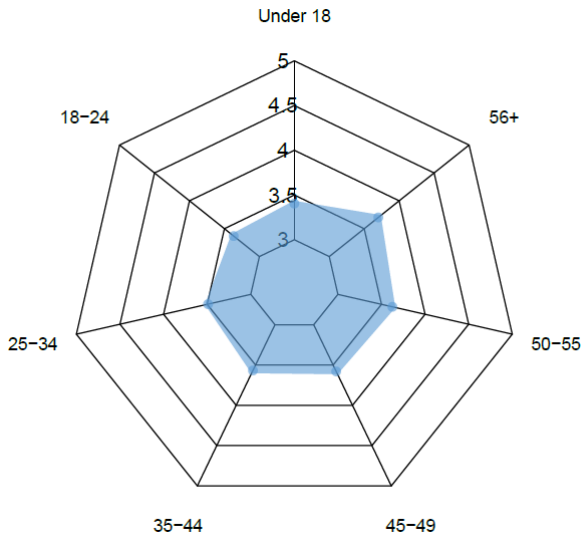



**FIGURE 21.** Age effects for sci-fi Children movie.

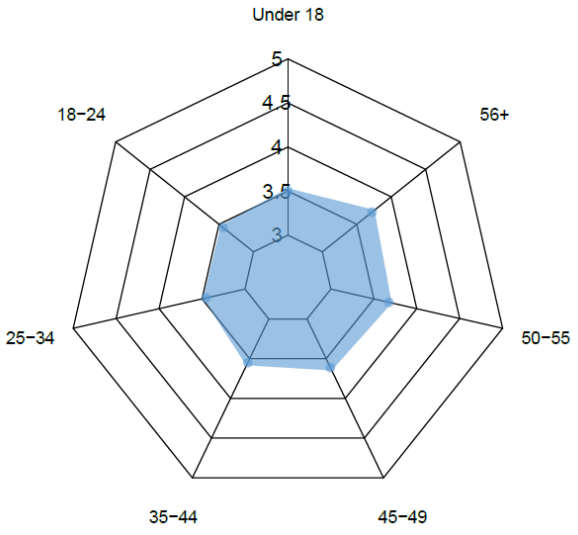



**FIGURE 22.** Age effects for adventure movie.

Clearly, by visual observation and hypothesis test, age effect differences between young people (Under 34) and old people (Beyond 45) are considerable. Therefore, to summarize, by analyzing the data from ***ml-1m***, we are able to come to a conclusion that young people are pickier than old people to mark movies.

## VII. DISCUSSION

In this paper, we introduced how to use linear mixed model for movie recommendation, in order to address *cold start problem* where there are few users' historical ratings available for data analysis. Compared with traditional algorithms, this model can explore the general pattern of users' rating behaviors, based on their features in different group levels. After some careful analysis we can dig out some implicit information or interesting social behaviors from millions of rating data. Linear mixed models are conceptually simple,

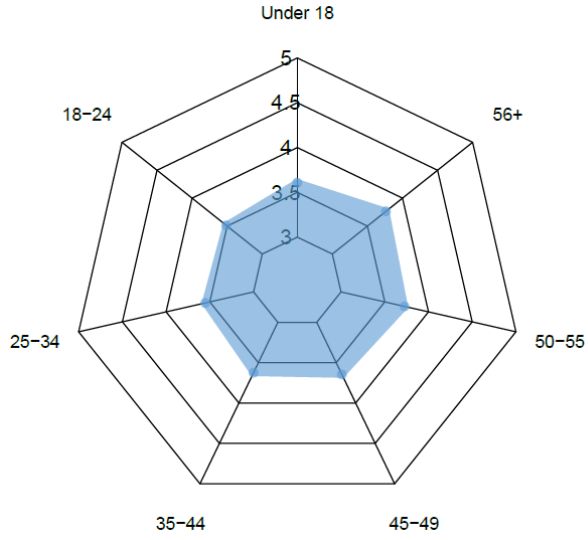



**FIGURE 23.** Age effects for all movie.

but the underlying computation is mathematically involved. In this essay, we showed the mathematical background knowledge of LMM. For readers who are interested in the details of related techniques, they can refer to [22] and [24]. Thanks to the lme4 R package, we can apply these mixed models to analyze huge recommendation data. In our future research, other properties of rating behaviors might be discovered, if we are given more types of users' traits and features. We shall also combine such a model with the traditional SVD approach [40], and consider how to do recommendation with multi-source data set [39].

## APPENDIX
## MOVIE GENRE AND OCCUPATION

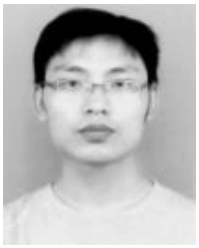
| Number | Movie Genre | Occupation |
|---|---|---|
| 0 | | other or not specified |
| 1 | Action | academic/educator |
| 2 | Adventure | artist |
| 3 | Animation | clerical/admin |
| 4 | Children | college/grad student |
| 5 | Comedy | customer service |
| 6 | Crime | doctor/health care |
| 7 | Documentary | executive/managerial |
| 8 | Drama | farmer |
| 9 | Fantasy | homemaker |
| 10 | Film-Noir | K-12 student |
| 11 | Horror | lawyer |
| 12 | Musical | programmer |
| 13 | Mystery | retired |
| 14 | Romance | sales/marketing |
| 15 | Sci-fi | scientist |
| 16 | Thriller | self-employed |
| 17 | War | technician/engineer |
| 18 | Western | tradesman/craftsman |
| 19 | | unemployed |
| 20 | | writer |

## REFERENCES

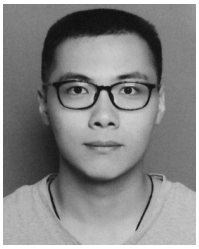
[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discoveries in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.
[2] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro, "Systems for knowledge discovery in databases," *IEEE Trans. Knowl. Data Eng.*, vol. 5, no. 6, pp. 903–913, Dec. 1993.
[3] B. M. Sarwar, J. A. Konstan, A. Borchers, and J. T. Riedl, "Applying knowledge from KDD to recommender systems," Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. 99-013, 1999.
[4] J. Schafer, "The application of data-mining to recommender systems," in *Encyclopedia Data Warehousing Mining*, 2nd ed. Suzhou, China: Information Science, 2009.
[5] S. Moon, P. K. Bergey, and D. Iacobucci, "Dynamic effects among movie ratings, movie revenues, and viewer satisfaction," *J. Marketing*, vol. 74, no. 1, pp. 108–121, Jan. 2010.
[6] D. Yang and X. Zhong, "The perception of film attractiveness and its effect on the audience satisfaction, intention and investment," *J. Service Sci. Manage.*, vol. 9, no. 1, pp. 21–27, 2016.
[7] J. Liu, M. Tang, Z. Zheng, X. Liu, and S. Lyu, "Location-aware and personalized collaborative filtering for Web service recommendation," *IEEE Trans. Services Comput.*, vol. 9, no. 5, pp. 686–699, Sep. 2016.
[8] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Syst. Appl.*, vol. 92, pp. 507–520, Feb. 2018.
[9] X. Lu, S. Zhu, Q. Niu, and Z. Chen, "Profile inference from heterogeneous data," in *Business Information Systems* (Lecture Notes in Business Information Processing), vol. 353, W. Abamowicz and R. Corchuelo, Eds. Cham, Switzerland: Springer, 2019, pp. 122–136.
[10] B. Gao, G. Zhan, H. Wang, Y. Wang, and S. Zhu, "Learning with linear mixed model for group recommendation systems," in *Proc. 11th Int. Conf. Mach. Learn. Comput. (ICMLC)*, 2019, pp. 81–85.
[11] Z. Chen, S. Zhu, Q. Niu, and X. Lu, "Censorious young: Knowledge discovery from high-throughput movie rating data with LME4," in *Proc. IEEE 4th Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2019, pp. 32–36.
[12] C. R. Henderson, O. Kempthorne, S. R. Searle, and C. M. Von Krosigk, "The estimation of environmental and genetic trends from records subject to culling," *Biometrics*, vol. 15, no. 2, pp. 192–218, 1959.
[13] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S.-S. White, "Generalized linear mixed models: A practical guide for ecology and evolution," *Trends Ecol. Evol.*, vol. 24, no. 3, pp. 127–135, Mar. 2009.
[14] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "FaST linear mixed models for genome-wide association studies," *Nature Methods*, vol. 8, no. 10, pp. 833–835, Oct. 2011.
[15] J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman, "Improved linear mixed models for genome-wide association studies," *Nature Methods*, vol. 9, no. 6, pp. 525–526, Jun. 2012.
[16] Z. Zhang, E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler, "Mixed linear model approach adapted for genome-wide association studies," *Nature Genet.*, vol. 42, no. 4, pp. 355–360, Apr. 2010.
[17] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nature Genet.*, vol. 44, no. 7, pp. 821–824, Jul. 2012.
[18] X. Zhang, Y. Zhou, Y. Ma, B.-C. Chen, L. Zhang, and D. Agarwal, "GLMix: General linear mixed models for large-scale response prediction," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 363–372.
[19] O. A. Montesinos-López, "Prediction of multiple-trait and multiple-environment genomic data using rcommender systems," *G3, Genes, Genomes, Genet.*, vol. 8, no. 1, pp. 131–147, 2018.
[20] S. Welham, S. Zhu, and A. J. Wathen, "Big data, fast models: Faster calculation of models from high-throughput biological data sets," Knowl. Transf. Project, Smith Inst., Univ. Oxford, Oxford, U.K., Tech. Rep. IP12-0009, 2013.
[21] S. Zhu, T. Gu, and X. Liu, "AIMS: Average information matrix splitting," 2016, *arXiv:1605.07646*. [Online]. Available: http://arxiv.org/abs/1605.07646
[22] S. Zhu, T. Gu, X. Xu, and Z. Mo, "Information spliting for big data analytics," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discivery (CyberC)*, 2016, pp. 294–302, doi: 10.1109/CyberC.2016.64.
[23] S. Zhu, "Fast calculation of restricted maximum likelihood methods for unstructured high-throughput data," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 40–43, doi: 10.1109/icbda.2017.8078871.
[24] S. Zhu and A. J Wathen, "Essential formulae for restricted maximum likelihood and its derivatives associated with the linear mixed models," 2018, *arXiv:1805.05188*. [Online]. Available: http://arxiv.org/abs/1805.05188
[25] S. Zhu and A. J Wathen, "Sparse inversion for derivative of log determinant," 2019, *arXiv:1911.00685*. [Online]. Available: http://arxiv.org/abs/1911.00685
[26] A. Galecki and T. Burzykowski, *Linear Mixed-Effects Models Using R* (Springer Texts in Statistics). Cham, Switzerland: Springer, 2013.
[27] C. J. H. Stryhn. *Confidence Intervals By the Profile Likelihood Method, With Applications in Veteinary Epidemiology*. Accessed: Nov. 2003. [Online]. Available: http://people.upei.ca/hstryhn/stryhn208.pdf
[28] J. B. J. Orloff. (2014). *Bootstrap Confidence Interval*. [Online]. Available: https://ocw.mit.edu/terms
[29] H. T. K. Akdur, D. Özonur, and H. Bayrak, "A comparison of confidence interval methods of fixed effect in nested error regression model," *J. Natural Appl. Sci.*, vol. 20, no. 2, pp. 167–175, 2016.
[30] L. Wasserman, "Bayesian model selection and model averaging," *J. Math. Psychol.*, vol. 44, no. 1, pp. 92–107, Mar. 2000.
[31] A. Kuznetsova, P. Brockhoff, and R. Christensen, "LmerTest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–27, 2017.
[32] R. H. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2008
[33] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *J. Memory Lang.*, vol. 59, no. 4, pp. 390–412, Nov. 2008, doi: 10.1016/j.jml.2007.12.005.
[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," 2014, *arXiv:1406.5823*. [Online]. Available: http://arxiv.org/abs/1406.5823
[35] B. Winter. (2013). *A Very Basic Tutorial for Performing Linear Mixed Effects Analyses*. [Online]. Available: http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf
[36] J. Deleeuw, "Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 599–609.
[37] X. Zhang, Y. Zhou, Y. Ma, B.-C. Chen, L. Zhang, and D. Agarwal, "Glmix: Generalized linear mixed models for large-scale response prediction," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining(KDD)*, New York, NY, USA, 2016, pp. 363–372.
[38] H. Zhao, Q. Yao, Y. Song, J. Kwok, and D. Lun Lee, "Learning with heterogeneous side information fusion for recommender systems," 2018, *arXiv:1801.02411*. [Online]. Available: http://arxiv.org/abs/1801.02411
[39] Y. Wang, T. Wu, F. Ma, and S. Zhu, "Personalized recommender systems with multi-scource data," in *Proc. Comput. Conf.*, London, U.K., 2020, p. 15.
[40] T. Zuo, S. Zhu, and J. Lu, "A hybrid recommender system combing singular value decomposition and linear mixed model," in *Proc. Comput. Conf.*, London, U.K., 2020, p. 19.

**ZHIYI (OLIVER) CHEN** received the B.S. degree in financial mathematics from Xi'an Jiaotong-Liverpool University, in 2019. He is currently a Graduate Student majoring in statistics with Columbia University. He is working on the application of data science and statistical machine learning for internet industry.

**SHENGXIN ZHU** received the Ph.D. degree in mathematics from the University of Oxford, in 2015. He is currently a Lecturer with the Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include numerical analysis, high performance computing, computational statistics, and data science. He is a member of SIAM, CSIAM, and ACM. He is a reviewer for Math Review. He was a Principal Investigator for Natural Science Foundation of China, Natural Science Foundation of Jiangsu province, and co-sponsored by more than six NSFC projects.



**QIANG NIU** received the Ph.D. degree from Xiamen University, in 2008. He is currently an Associate Professor with the Department of Applied Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests are scientific computing and data mining. He has authored or coauthored more than 20 research articles in many international journals.



**TIANYU ZUO** is currently pursuing the bachelor's degree majoring in applied mathematics with the Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University. He is currently very interested in the application of data science and machine learning.