Censorious Young: Knowledge Discovery from High-throughput Movie Rating Data with LME4

Zhiyi Chen, Shengxin Zhu†, Qiang Niu†, Xin Lu
Department of Mathematics, Xi'an Jiaotong-Liverpool University
Email† : {Shengxin.Zhu; Qiang.Niu}@xjtlu.edu.cn {Zhiyi.Chen15;Xin.Lu15}@student.xjtlu.edu.cn

***Abstract*—Quantitative analysis of high throughput movie rating data provides supports for one general social behavior: the young are usually more censorious than senior people when rating/evaluating the same thing. Millions of movie rating data with users' categorical age information are analyzed by the linear mixed model with the lme4 R package. When the age factor is viewed as fixed effects, the rating scores for movies are positively related to age. In general the young people are tends to give lower score than senior people. Such a social behavior phenomenon should be carefully examined in a recommendation system and in data collection.**

***Index Terms*—Knowledge discovery in databases(KDD), Linear-mixed effects model(LMM), recommender system (RS), lme4 software.**

I. Introduction

Knowledge discovery is the process to unearth useful information or high-level knowledge from low or raw digital data that grow rapidly [1][2]. It is known that behavior study may lead to more sale amounts to customers [3]. For example, firms can forecast customers' behaviors and obtain their preference by analyzing customer databases. Recommender system, which is a popular and updated method for knowledge discovery is receiving broad success in E-Commerce nowadays [3][4]. Movie becomes one of the important entertainments in people's daily life and increasingly individuals are posting evaluations about movies online [5]. In this circumstance, the database of movie rating is created and its volume is expanding rapidly. Relying on the recommendation system, it is no doubt that efforts on creating model and sophisticated algorithm is worthwhile. Netflix emphasized on the importance of better recommender system indicated by Netflix Price, which is a contest willing to provide 1 million dollars for the team giving the best method on recommender system improvement [5]. More simply, we should think about what the movie rating databases indicate and can we obtain useful information from simple knowledge discovery. Is there any social behavior hidden in those voluminous data?

Moon, S. (2010) did analysis about movie database exploring the relationship between movie genre and movie rating and it gives a conclusion that sequel movies can achieve lower rating than the original ones, which is resulted from the reason that viewers' satisfaction is declined [6][5]. It delivers the information that original and innovative ones can be more attractive so that it inspires movie firms to produce more original movies. Consider the fact that individual possesses diverse traits and viewers can be clustered by numbers of different factors like gender, age, occupation, etc. and movies can be categorized into different genres, Gao et al. examined group recommendation and concluded that certain movies are more popular to people from certain occupation [7]. Here we continued the study of recommendation with linear mixed model (LMM), and find an interesting social psychological phenomenon: the young people tends to more censorious than senior people.

II. Linear Mixed Models

Linear mixed models are developed based on linear models. Compared with linear models, LMMs take additional random effects into consideration, and thus they are more flexible than linear models whose assumptions on the normality, independence and homogeneity of response variables. LMMs are based on more practical assumptions than linear models do. They have widespread applications in breeding, genome wide association studies , and recently are used in recommendation systems [8].

The advantage of linear mixed models in recommendation systems lies in that it not only can be easily implemented but also can handle the cold start problem. Within a LMM, users' features (age, gender, occupation, etc) can be viewed as categorical variables, the association between these categorical variable and movie features(ID, genre, etc) can be assumed, modeled, analyzed and tested through a rigorous hypothesis test reasoning process. When one new user or new movie has little or even no rating history, we can still give some prediction through pre-learned parameters for users' categorical information or movies categorical information. If historical records are abundant, the user's preference can be reflected more clearly. Therefore the cold start problem can be well solved in the linear mixed model approach.

Linear mixed-effects model has implemented in several 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 [9] giving lmer() function to construct linear mixed-effects models for discovering knowledge.

LMMs divide data sets into levels according to certain grouping factors and thus are also referred as hierarchical linear models [10]. For multilevel data, a LMM at a given level of a grouping factor as following:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i \tag{1}$$

where $\mathbf{y}_i$ is a $n_i \times 1$ vector of responses, $\mathbf{X}_i$ is a $n_i \times p$ design matrix of fixed effects, β is a $p \times 1$ vector of fixed effects, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects in level i, $\mathbf{Z}_i$ is a $n_i \times q$ matrix of covariates which shows the correlation between responses $\mathbf{y}_i$ and random effects b_i, ε_i is the vector of residual errors for level i. What should be emphasized is that $\mathbf{Z}_i$ contains known values of q covariates corresponding to q random effects chosen from its distribution [10]:

$$\mathbf{Z}_i = (\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, \cdots, \mathbf{z}_i^{(q)}). \tag{2}$$

Moreover, $\mathbf{b}_i$ is unobservable. It delivers the information that random effect lacks pattern, causing difficulty being calculated.

In LMM, observations are considered not necessarily independent and have heteroscedasticity. The correlation between observations in the same level are reflected in the distribution of $\mathbf{b}_i$ and ε_i. Since they are in the same level, they are able to follow bivariate normal distribution:

$$\mathbf{b}_i \sim N(0, \mathbf{G}) \qquad \varepsilon_i \sim N(0, \mathbf{R}_i) \tag{3}$$

where $\mathbf{b}_i$ is independent of ε_i. Moreover, $\mathbf{G}$ and $\mathbf{R}_i$ can be specified that:

$$\mathbf{G} = \sigma^2 G \qquad \mathbf{R}_i = \sigma^2 R_i \tag{4}$$

where G and R_i both are variance function which represents the weight of observation's variance decided by parameter θ_G and θ_{R_i} respectively. Therefore, when random effects $\mathbf{b}_i$ is known, then the conditional distribution of $\mathbf{y}_i$ can be formulated:

$$E[\mathbf{y}_i|\mathbf{b}_i] = \mu_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i \tag{5}$$

$$Var[\mathbf{y}_i|\mathbf{b}_i] = \sigma^2 R_i \tag{6}$$

When $\mathbf{b}_i$ is not given, the unconditional distribution of $\mathbf{y}_i$ can be defined as:

$$E[\mathbf{y}_i] = \mathbf{X}_i\beta$$

$$Var[\mathbf{y}_i] = \sigma^2[\mathbf{Z}_i G \mathbf{Z}_i' + R_i]$$

Combing data sets of all levels of a grouping factor, we can get the classical formula of LMM for all data:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{7}$$

where $\mathbf{Y} = (\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_N')'$ is the $n \times 1$ vector of responses, where $n_1 + n_2 + \cdots + n_N = n$, β is the $p \times 1$ vector of fixed effects, $\mathbf{X}$ is the $n \times p$ design matrix for fixed effects, $\mathbf{Z}$ is the $n \times Nq$ matrix of random effects, $\mathbf{b}$ is the $Nq \times 1$ vector of random effects, where $\mathrm{b} = (\mathbf{b}_1', \mathbf{b}_2', \cdots \mathbf{b}_N')$, ε 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

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2(\mathbf{Z}G\mathbf{Z}' + R)) \qquad \mathbf{Y}|_{\mathbf{b}} \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{b}, \sigma^2 R) \tag{8}$$

III. Parameter estimation

A. Fixed and Random Effects Estimation

Once the variance parameter $\sigma; \theta = (\theta_G, \theta_R)$ are known, the LMM can be reduced to a general linear mixed model $y = X\beta + \epsilon$, with general variance-variance matrix V. The fixed effected can be estimated by the standard generalized least square (GLS),

$$\hat{\beta} = (\sum_{i=1}^{n} X_i' V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i. \tag{9}$$

If we want to estimate the random effect at the same time, we can solve the Mixed Effects Equations (MEE):

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}. \tag{10}$$

The estimation of the fixed effects in (9) and (10) are both *Best Linear Unbiased Estimations* (BLUE). The random effects in (10) are *Best Linear Unbiased Predictions*(BLUP) (the estimation for random effects are often referred as prediction).

B. Variance Parameter Estimation

In many practical cases, the variance parameters σ, θ are unknown. In those case, we first have to use a *Maximum Likelihood* or *Restricted Maximum Likelihood* method to estimate the variance parameters. In general REML is preferred for the cases that each subgroup have small samples.

1) Profile log-likelihood for ML: In the ML approach, we first derive the the likelihood expression:

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

where $V_i = Z_i G(\theta_G) Z_i' + R(\theta_R)$. Ignore constant and take log operation, we get log-likelihood function:

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

Assume that θ is known, then maximizing (12) with respect to β for every value of θ leads to an estimation of β given by

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

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

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

where $r_i = r_i(\theta) = y_i - X_i\hat{\beta(\theta)}$. Maximizing $l^*_{ML}(\sigma^2, \theta)$ with respect to σ^2 for every known value of θ leads to the estimation of σ^2:

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

By plugging (15) into (14), we get a log-profile likelihood function for θ:

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

Then maximization of $l^*_{ML}(\theta)$ can yield an estimator $\hat{\theta}_{ML}$ of θ. Plugging $\hat{\theta}_{ML}$ into (13), (10) and (15) produces estimator $\hat{\beta}_{ML}$, **b** of β and $\hat{\sigma}^2_{ML}$ of σ^2.

However, there is a significant limitation on maximum likelihood estimation that ML estimators $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ are both biased because they don't adjust for the uncertainty in estimation of β which means $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ will change value following β's altering and cannot estimate σ^2 and θ accurately. However, σ^2 and β can be better estimated by restricted maximum likelihood estimation.

2) Profile log-likelihood for REML: In the REML approach, we have to use an error transform techniques to make the estimation of fixed effects β and variance parameters independent. And then we use the likelihood function of a set of $n - p$ independent contrasts of y to do it where p is the dimension of β. After obtaining $\hat{\beta}(\theta)$, the log-restricted-likelihood function is given by:

$$\begin{aligned} l^*_{REML}(\sigma^2, \theta) = &-\frac{n-p}{2} log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n} log[det(V_i)] \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i' V_i^{-1} r_i - \frac{1}{2} log[det(\sum_{i=1}^{n} X_i' V_i^{-1} X_i)] \end{aligned} \tag{17}$$

Maximizing $l^*_{REML}(\sigma^2, \theta)$ with respect to σ^2 leads to estimator of σ^2 that:

$$\hat{\sigma^2}_{REML} = \sum_{i=1}^{n} r_i' V_i^{-1} r_i / (n - p). \tag{18}$$

Plugging (18) into (17), we get a function with respect to θ only:

$$\begin{aligned} l^*_{REML}(\theta) = &-\frac{n-p}{2}[log(\sum_{i=1}^{n} r_i' V_i^{-1} r_i/(n-p)) + 1] \\ &-\frac{1}{2}\sum_{i=1}^{n} log[det(V_i)] - \frac{1}{2} log[det(\sum_{i=1}^{n} X_i' V_i^{-1} X_i)] \end{aligned} \tag{19}$$

Estimator of θ can be obtained from maximization from (19), which can be applied to get estimators of β and σ^2, respectively. For maximizing the function (19), the reader is directed to [11], [12], [13], [14], [15], for more technical details.

IV. Criterion for model selection

Different models possess different focus and purposes which result in different values. One benefit for using linear mixed model in recommendation systems lies in the fact that there are sophisticated hypothesis test techniques for model selection. The commonly used criterions include, the log-likelihood, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and p-value.

A. Log-likelihood

It is the most simplest criterion which has the expression that

$$l(\Theta) = logL(\Theta) = log(f(Y|\Theta)) \tag{20}$$

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 $\beta = (\beta_1, \beta_2, \cdot \cdot, \beta_k)$ 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.

According to this expression, we can see clearly the design of this criterion: Θ is the key component of our model so that it can represent our model, Y is the data set observed. Therefore, the meaning of f is that how much the data fits our candidate models and adding log is to avoid zero value in the likelihood[16]. The higher value of the log-likelihood achieves, the better the data fit our models.

However, log-likelihood criterion has a huge problem that it doesn't consider the number of parameters because we prefer a model with high log-likelihood and small number of parameters which can not only guarantee high level of fitting but also require less calculations and operations. Normally, it is widely considered that more 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, every more parameters added in model would have highly little influence to log-likelihood value.

That is, it makes little sense to having too many parameters. On the contrary, we prefer the model with less parameters when its log-likelihood value are just little 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.

B. Akaike Information Criterion (AIC)

Akaike information criterion is a standard to measure the goodness of statistical model fitting. It was founded and developed by Japanese statistician Akaike. This criterion suffices to weigh 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.

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

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

Both expressions have the same structure with slight difference including sign and multiplies. To be clear, in this essay, we use formula

$$AIC = 2p - 2l(\hat{\Theta}) \tag{21}$$

where p is the number of parameters, $l(\hat{\Theta})$ is the maximum log-likelihood. The lower AIC value a model has, the better the model is.

C. Bayesian Information Criterion (BIC)

In statistics, there are two ways to optimize models. On the one hand, adding more parameters in models can increase model's complexity. On the other hand, collecting more observations or data suffices to facilitate model's 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 showed below:

$$BIC = p \cdot log(n) - 2l(\hat{\Theta}) \tag{22}$$

where p is the number of parameters n is the number of observations

D. F-test

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

$$D = 2[l(\hat{\Theta}_{max}; y) - l(\hat{\Theta}; y)] \tag{23}$$

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 freedom, p and q respectively where $p < q$ and the 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:

$$F = \frac{(D_0 - D_1)/(p-q)}{D_1/(N-q)} \tag{24}$$

After we obtain the F value, p-value also can be achieved by referring to F table. Next the process of making decision 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 extra parameters in $m1$, we tend to choose $m1$ which has more parameters. On the contrary, when the p-value is more than 5 percent, there is not significant difference existing so that the model with less parameters $m0$ is our choice.

E. Model comparison

To conclude, we can use Anova(model1, model2) which is an useful command in R software [17]. It can directly calculate log-likelihood, AIC, BIC of models respectively as well as p-value of F-test.

V. Numerical Results

A. Benchmark data set

The ***ml-1m*** dataset is used in this experiment. It provides us 6040 users, 3952 movies and movie rating made on a 5-star scale. Among these ratings, each user has at least 20 ratings. Furthermore, movie information including Movie ID, title and genres are listed. User information contains gender, age and occupation and they are all in the category type. Although some inaccurate information exists in the dataset, the huge volume allows us to ignore its effect to our data analysis.

Secondly, we are willing to see the relationship between the rating toward a certain genre movies (1064 movies from ***ml-1m***) and user's condition. Here the user's age is regarded as our interest, i.e. the fixed effect. Occupation and gender are considered as random effects. Therefore, the model in the *lme4* can be constructed as:

```
Model1 <- lmer(Rating  -1 + Age +
(1|Occupation) + (1|Gender), comedy).
```
(25)

In this model, age is consider to be the fixed effect with seven levels(groups), and `occupation` and `gender` are the random effects. `Comedy` is the data set. After summarizing the model, we can see the figure of the effects of each age group to rating on comedy movies. To be clear, the radar chart is able to show the comparison clearly.

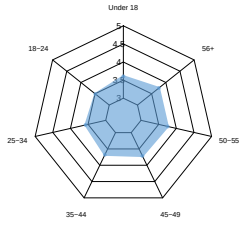

Fig. 1. Age Effects in comedy

At first cursory look, we can see the difference of effects to rating between ages are small. However, this deduction is unreasonable since we are not sure about whether these

differences are significant. To judge the significance of differences, we need to do hypothesis test. At first, we choose two effects having the largest difference i.e. in the model, they will share the same parameter and the model contains one less parameter. Then, we use **Anova()** to compare previous model and the model after change with 99 percent confidence level. If the result shows they are not significant, then all difference between ages are of no significance. On the contrary, when the result reflects significance, we need to make the value of 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 know which effects differences are worth being considered.

Of course, the analysis of comedy movie only is hard to support our opinion that the young are more censorious than senior people. Thus, by constructing the same model and operating the same hypothesis tests for science fiction, children, adventure and all movies, we can get radar charts below respectively and our standpoint can be more valid.

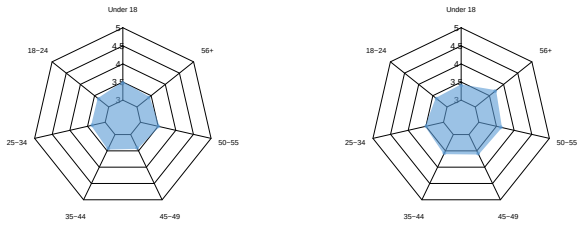

Fig. 2. Age Effects in science fiction (left) and children (right) movies

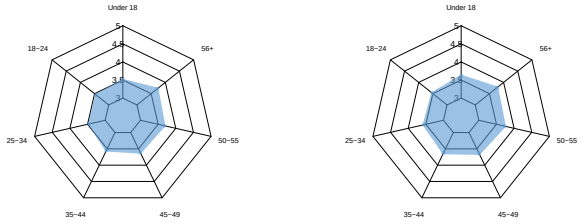

Fig. 3. Age Effects in adventure (left) and all movies (right)

VI. Conclusion

Linear mixed model provide a robust parametric learning approach which forms recommendation system. With some information of users or the movie genre available, it can solve the cold start problem. With abundant hypothesis testing procedures, we can analysis the data based on its solid mathematics foundations. The association between the rating score and the age categorical information provides qualitative supports for a social behavior that the young people are more censorious than the senior people for most of the movies or movies genres. Based on this fact the companies collecting continuous age information may help them to better analysis their data base and when evaluating the rating results, the effects of age should be carefully considered.

Acknowledgement

The research is supported by Natural Science Foundation of China (NSFC. 11501044), Jiangsu Science and Technology Basic Research Programme (BK20171237), Key Program Special Fund in XJTLU (KSF-E-21)), Research Development Fund of XJTLU (RDF-2017-02-23), Research Enhance Fund of XJTLU(REF-18-01-04) and partially supported by NSFC (No.11571002, 11571047, 11671049, 11671051, 61672003, 11871339).

References

[1] G. P.-S. Usama Fayyad and P. Smyth, "From data mining to knowledge discoveries in databases," *AI Magazine*, 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 Transactions on Knowledge and Data Engineering*, vol. 5, no. 6, pp. 903–913, Dec 1993.

[3] B. Sarwar, J. Konstan, A. Borchers, and n. Riedl, "Applying knowledge from kdd to recommender systems," *University of Minnesota*, 1999.

[4] J. Schafer, "The application of data-mining to recommender systems," 2009.

[5] S. Moon, P. K. Bergey, and D. Iacobucci, "Dynamic effects among movie ratings, movie revenues, and viewer satisfaction," *Journal of Marketing*, vol. 74, no. 1, pp. 108–121, 2010. [Online]. Available: https://doi.org/10.1509/jmkg.74.1.108

[6] D. J. Yang and X. Y. Zhong, "The perception of film attractiveness and its effect on the audience satisfaction, intention and investment." *Journal of Service Science & Management*, vol. 09, no. 1, pp. 21–27, 2016.

[7] B. Gao, G. Zhan, H. Wang, Y. Wang, and S. Zhu, "Learning with linear mixed models for group recommendations," in *Proceedings of the 11th Interational Conference on Machine Learning and Computing*, ser. ICMLC'19, 2019.

[8] 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 *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 363–372. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939684

[9] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software, Articles*, vol. 67, no. 1, pp. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/v067/i01

[10] A. Gaecki and T. Burzykowski, *Linear Mixed-Effects Models Using R*, 2013.

[11] S. Zhu, "Computing log-likelihood and its derivatives for restricted maximum likelihood methods," *arxiv:1608.07207*, 2016.

[12] S. Zhu, T. Gu, X. Xu, and Z. M, "Information splitting for big data analytics," in *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2016, Chengdu, China, October 13-15, 2016*, 2016, pp. 294–302. [Online]. Available: https://doi.org/10.1109/CyberC.2016.64

[13] S. Zhu, T. Gu, and X. Liu, "Information matrix splitting," *arxiv:1605.07646*.

[14] S. Zhu, "Fast calculation of restricted maximum likelihood methods for unstructured high-throughput data," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, March 2017, pp. 40–43.

[15] S. Zhu and A. Wathen, "Essential formulae for restricted maximum likelihood and its derivatives associated with the linear mixed models," *arxiv:1805.05188*, 2018.

[16] L. Wasserman, "Bayesian model selection and model averaging," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, 2000.

[17] A. Kuznetsova, P. Brockhoff, and R. Christensen, "lmerTest Package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, 2017.