Statistical Modeling for Review Ratings Data

Yeh Ching Low

***Abstract*—Consumer reviews and ratings of a product or service on websites is an important factor in a potential customer's decision-making process. The ratings given by the consumers or reviewers are also valuable to the product vendor or service provider for market segmentation or product and service evaluation and improvement. Since the ratings can be treated as discrete ordinal data, we apply the Combination of Uniform and shifted Binomial mixture (CUB) model to analyze review ratings. The CUB model gives an interesting perspective on review ratings modeling and analysis since the model provides a tool to interpret the ratings in terms of the reviewer's level of feelings and uncertainty towards the product or service. The CUB model is also able to incorporate covariates, such as the reviewer's gender, to explain the level of feelings and uncertainty. We illustrate with examples from two well-known datasets in the field of data mining.**

***Index Terms*—Mixture models, ordinal data, review ratings, statistical models.**

I. Introduction

With increased accessibility to the Internet and the rise of e-commerce, consumer reviews and ratings of a product or service on websites have become an important factor in a potential customer's decision-making process. Reference [1] revealed that about 82% of Americans consult online ratings and reviews for their first-time purchase of a product. The ratings given by the consumers along with the consumers' profile, when available, are valuable information to the product vendor or service provider for market segmentation, product or service evaluation and quality improvement. For market segmentation, it is of interest to explain the ratings by the reviewers' profile such as demographic variables. On the other hand, decision-making processes for product or service evaluation and quality improvement can be improvised through further analysis of the ratings awarded. An example of such an analysis would be to examine the ratings in terms of the reviewer's level of feelings and amount of uncertainty when selecting the rating. This type of analysis is useful since the uncertainty component is significant in the elicitation process, especially when the consumers do not give a rating towards a strongly positive or negative feeling [2]. In this paper, a statistical model that is able to account for these factors in the analysis of review ratings data is reviewed and studied.

Since the ratings are usually integers in a scale of 1 to 5 (or could be transformed as such), the ratings data can be treated as discrete ordinal data. The Conway-Maxwell-Poisson mixture model has been proposed to model aggregate survey data and ratings data [3] but it does not model the latent aspects of review ratings. The Inverse Hypergeometric model [4] has the constraint of an extreme mode thus is not appropriate for review ratings data. In view of this, we consider and apply the Combination of Uniform and shifted Binomial (CUB) model [2], [5] to analyze the review ratings. The CUB model is a statistical approach for modeling and analysis of ordinal data. Substantial work has been done on using the CUB model for analyzing preferences or ranks data [6]-[8].

We illustrate the practicality of the CUB model in this context by fitting the CUB model on real ratings data to model the distribution of the ratings and seek to analyze the ratings as a direct consequence of two latent aspects: the reviewers' feeling toward the product or service and the reviewers' uncertainty when choosing a rating. Furthermore, covariates such as the reviewers' gender can be incorporated into the CUB model to explain the reviewers' level of feeling and/or level of uncertainty.

The outline of the paper is as follows. In the next section, we briefly review the CUB model. We discuss the results of the application of the model on two review ratings dataset in Section III. Finally, Section IV concludes the paper and points out some future research direction.

II. The CUB Model

The CUB model was proposed in [2] to understand ranking or ratings data of products and services. The CUB model is based on the assumption that the choices of ranking or ratings of an item are a direct consequence of personal feelings toward the item and the uncertainty in choosing. In using the CUB model on ratings data, the probability distribution of the ratings is a mixture of the shifted binomial distribution and the uniform distribution. The level of feelings of the reviewer towards the item rated is modeled by a shifted binomial distribution and the degree of uncertainty inherent in the decision making process of selecting the rating is represented by the uniform distribution. The choice of the uniform distribution implies that the probability of choosing each item is the same (i.e., complete uncertainty). Consequently, each decision is characterized by a certain degree of feeling and a certain degree of uncertainty, resulting in a mixture of a shifted binomial and a uniform distribution.

In the context of review ratings, let R denote the rating assigned by a reviewer to a given product or service, such that R is a random variable taking the first m integers, m being the number of possible ratings. Without loss of generality, let 1 represent the worst rating and 5 the best. The probability that R takes the value r is given by:

Manuscript received August 5, 2017; revised November 15, 2017.

Y. C. Low is with the Department of Computing and Information Systems, Sunway University, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia (e-mail: yehchingl@sunway.edu.my).

doi: 10.18178/ijke.2017.3.2.086

$$\Pr(R=r)=\pi\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}+(1-\pi)\frac{1}{m}$$

where $\pi \in (0,1]$ and $\xi \in [0,1]$. The probability $(1-\xi)$ can be interpreted as a measure of the feeling toward the product or service whereas $(1-\pi)$ is the probability of uncertainty in choosing the rating. Using this definition of the model, the value of ξ decreases with the liking feeling toward the product or service. Assuming presence of heterogeneity in a group of reviewers, the coefficients π and $(1-\pi)$ can also be interpreted as the proportion of reviewers in the group who are thoughtful and instinctive, respectively.

The CUB model with covariates models the parameters π (uncertainty) and ξ (feeling) by the reviewers' characteristics, such as their demographic variables. If π is affected by *p* covariates and ξ is affected by *q* covariates, we have a CUB(*p*,*q*) model given as follows:

$$\Pr(R_i=r_i)=\pi_i\binom{m-1}{r_i-1}(1-\xi_i)^{r_i-1}\xi_i^{m-r_i}+(1-\pi_i)\frac{1}{m},$$

$$\pi_i=\frac{1}{1+e^{-\beta_0-\beta_1 x_{i1}-\ldots-\beta_p x_{ip}}},$$

$$\xi_i=\frac{1}{1+e^{-\gamma_0-\gamma_1 w_{i1}-\ldots-\gamma_q w_{iq}}},$$

where R_i represents the ratings of the *i*-th consumer; r_i is its observed value; $x_{i1},\ldots,x_{ip}$ are the covariates for the uncertainty of the *i*-th consumer; $w_{i1},..,w_{iq}$ the covariates for the feelings of the *i*-th consumer and β_j ($j = 1, \ldots, p$) and γ_j ($j = 1, \ldots, q$) represent the coefficients of the relation between the covariates and the parameters of uncertainty and feelings, respectively.

In this paper, all of the analysis is performed in *R* [9] and we use the CUB package [10] for model fitting. In the CUB R package, parameter estimation is performed via the Expectation-Maximization (EM) algorithm for maximum likelihood estimation.

III. Results

We apply the CUB model on two well-known datasets in the field of data mining and text analysis. The first dataset is the TripAdvisor dataset which is available for download at http://times.cs.uiuc.edu/~wang296/Data/ [11]. The second dataset is the MovieLens dataset, which is a benchmark dataset for ratings review prediction and is available for download at https://grouplens.org/datasets/movielens/1m/ [12]. Demographic information of the reviewers such as gender, age and occupation are included in the MovieLens dataset. We fit the CUB model without covariates on the TripAdvisor dataset, specifically on the ratings data of a selected hotel. For the purpose of comparison, we fit both the CUB model with and without covariates on the ratings data of the movie Titanic from the MovieLens dataset.

A. CUB Model without Covariates: TripAdvisor Dataset

The TripAdvisor website is a popular travel advisory website. The website's users can sign in to review travel-related services such as restaurants and hotel accommodation. The ratings available for a hotel review are on a five-star scale, with 1 indicating Terrible, 2 for Poor, 3 for Average, 4 for Very Good and 5 indicating Excellent. Reviewers may provide ratings for the following aspects of a hotel: Overall, Service, Value, Sleep Quality, Rooms, Location, Business Service and Cleanliness. For the TripAdvisor dataset, we analyze the ratings data on one randomly selected hotel which has a total of 316 reviewer ratings. Since a significant number of reviewers did not review the Business Service, Sleep Quality and Location of the hotel, we do not include these aspects in our analysis. Prior to model fitting, we also performed data cleaning and removed all reviews with missing values, resulting in 257 complete review ratings. A summary of the ratings for the remaining five aspects is given in Table I.

In general, the reviewers showed a good level of satisfaction towards this hotel. Satisfaction level were similar for the pair Service and Cleanliness, as well as for the pair Rooms and Value. The reviewers are more satisfied on Service and Cleanliness, while the satisfaction on Rooms and Value are slightly lower. More reviewers are not satisfied on Rooms and Cleanliness.

We fit the CUB model without covariates to the ratings data and the estimated parameters with their standard errors and normalized dissimilarity index is given in Table II. The normalized dissimilarity index is used to measure how well the CUB model fits the ratings data and the fit is considered satisfactory when the index is ≤ 0.1 [13]. All the ratings showed a value of dissimilarity index of less than 0.05, confirming that the CUB model gives a very good fit.

In Fig. 1, the degree of uncertainty and the level of feelings are plotted as a function of the maximum likelihood estimates of the parameters π and ξ, respectively, for the five aspects. The estimated values of π imply that the uncertainty share is between 0.2 and 0.4 for all five aspects. Cleanliness has the lowest level of uncertainty whereas Service has the highest uncertainty. It appears that the reviewers may have some lingering doubts when rating on Service. The reviewers showed a high level of feelings in the overall rating of the hotel. Among the different aspects of the hotel, Service received the highest positive evaluation, followed by Cleanliness. These observations are consistent with the summary from Table I. Although the level of feelings for these two aspects are similar, it is interesting to note that the uncertainty of the reviewers when rating these two aspects differs to a certain degree. Compared to Cleanliness, the reviewers are less certain on Service, which may be more subjective to rate as compared to Cleanliness.

TABLE I: SUJMMARY OF RATINGS (%) FOR TRIP ADVISOR DATASET

Satisfaction toward:	**Ratings**				
	1	***2***	***3***	***4***	***5***
Overall	6.6	9.3	23.7	33.5	26.9
Service	7.4	6.6	15.2	30.7	40.1
Cleanliness	5.8	6.2	15.2	34.3	38.5
Rooms	6.2	11.7	24.9	32.7	24.5
Value	6.6	10.1	24.1	31.9	27.3

TABLE II: PARAMETER ESTIMATES AND DISSIMILARITY INDEX FOR TRIPADVISOR DATASET (STANDARD ERROR IN BRACKETS)

Satisfaction toward:	$\hat{\pi}$	$\hat{\xi}$	**Dissimilarity Index**
Overall	0.6946 (0.0680)	0.2682 (0.0225)	0.0254
Service	0.6588 (0.0582)	0.1599 (0.0201)	0.0218
Cleanliness	0.7289 (0.0549)	0.1789 (0.0187)	0.0065
Rooms	0.6893 (0.0739)	0.2947 (0.0236)	0.0255
Value	0.6756 (0.0709)	0.2697 (0.0235)	0.0344

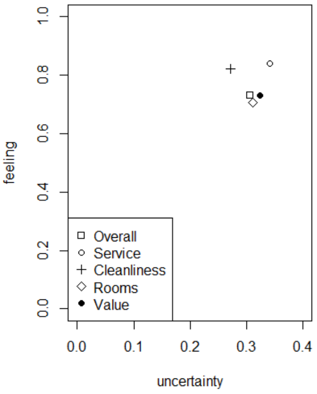

Fig. 1. CUB models of reviewers' ratings on the five aspects.

B. CUB Model with and without Covariates: MovieLens Dataset

The MovieLens website is a web-based movie recommender system. For each movie, the ratings available are on a five-star scale. We selected the movie Titanic for our analysis, which has a total of 115 ratings. The reviewers for this movie consist of 82 males and 33 females. In general, the movie received low ratings from MovieLens users with 32.2% of the reviewers giving a 1-star rating and only 7% of the reviewers give a 5-star rating, as shown in Table III. The low ratings received by this movie which is classified under the drama/romance genre in the dataset could be attributed to the gender distribution of the reviewers. Empirical studies have confirmed, to a certain degree, that men tend to dislike romantic and melodramatic movies [14], [15]. To illustrate the usefulness of incorporating covariates in explaining uncertainty and feelings, we model the feelings parameter and uncertainty parameter as a function of: (a) the reviewers' gender, or (b) the reviewers' age. In other words, we fit the CUB(0,1) model where the parameter for feeling ξ is further modeled by gender or age, and the CUB(1,0) model where the parameter π for uncertainty is further modeled by gender or age. We compare the model fit results of the CUB model without covariates, CUB(0,1) models and the CUB(1,0) models and the results are given in Table IV and Table V.

TABLE III: RATINGS SUMMARY FOR MOVIELENS DATASET

	Ratings				
	1	***2***	***3***	***4***	***5***
Percentage of reviewers	32.2	26.1	21.7	13.0	7.0

TABLE IV: MODEL FITTING OF MOVIELENS DATASET WITH GENDER AS COVARIATE

Model	**Parameter Estimates (Standard Error)**	**Log-likelihood**	**Deviance**
Without covariates	$\hat{\pi}$ = 0.5284 (0.1190) $\hat{\xi}$ = 0.7822 (0.0476)	-173.9347	3.2712
CUB(0,1)	$\hat{\pi}$ = 0.5141 (0.1148) $\hat{\gamma}_0$ = 1.6249 (0.4055) $\hat{\gamma}_1$ = -0.8394 (0.5258)	-172.4934	0.3885
CUB(1,0)	$\hat{\beta}_0$ = 0.1618 (0.5134) $\hat{\beta}_1$ = -0.2637 (0.9841) $\hat{\xi}$ = 0.7865 (0.0516)	-173.9022	3.2062

TABLE V: MODEL FITTING OF MOVIELENS DATASET WITH AGE AS COVARIATE

Model	**Parameter Estimates (Standard Error)**	**Log-likelihood**	**Deviance**
Without covariates	$\hat{\pi}$ = 0.5284 (0.1190) $\hat{\xi}$ = 0.7822 (0.0476)	-173.9347	3.2712
CUB(0,1)	$\hat{\pi}$ = 0.5955 (0.1094) $\hat{\gamma}_0$ = 2.2834 (1.0645) $\hat{\gamma}_1$ = -1.5230 (1.2228) $\hat{\gamma}_2$ = -0.8483 (1.1277) $\hat{\gamma}_3$ = -0.4671 (1.1842) $\hat{\gamma}_4$ = -1.9502 (1.2944) $\hat{\gamma}_5$ = -2.2901 (1.2205) $\hat{\gamma}_6$ = -1.3795 (1.2492)	-169.3757	-5.8468
CUB(1,0)	$\hat{\beta}_0$ = 0.8852 (1.3031) $\hat{\beta}_1$ = -0.9737 (1.6794) $\hat{\beta}_2$ = -1.1719 (1.4858) $\hat{\beta}_3$ = 0.0291 (1.6030) $\hat{\beta}_4$ = -2.1162 (2.9637) $\hat{\beta}_5$ = -6.9163 (28.0762) $\hat{\beta}_6$ = -1.2807 (2.0172) $\hat{\xi}$ = 0.8131 (0.0506)	-172.2285	-0.1412

From the CUB model without covariates, it can be seen that the level of uncertainty is very high among the reviewers. As expected, the estimated value of the parameter ξ indicates that the level of feelings is very low. Comparing the CUB models with and without covariates, the log-likelihood and deviance values infer that the CUB(0,1) model gives the best fit to the ratings data in both cases where age or gender is included as a covariate. From the parameter estimates of the CUB(0,1) model with gender as covariate in Table IV, males yield a higher value for the feeling parameter ξ, indicating a lower level of liking as compared to females. From the model of best fit in Table V, which is the CUB(0,1) model with age as the covariate, we observe that the level of feelings is not a monotonic function of age for this particular review ratings data.

IV. DISCUSSION AND CONCLUSION

This paper offers a statistical modeling approach to analyze reviews rating data using the CUB model. The CUB model allows the reviewers' ratings of a product or service to be interpreted in terms of the reviewers' level of feelings towards the product or service and the underlying uncertainty of the reviewers. The feelings and uncertainty components can be further modeled by the reviewers' characteristics such as gender or other factors. The modeling of the review ratings in terms of reviewers' characteristics such as demographic profile can offer insight to the product vendor or service provider for market segmentation purpose. For the purpose of product or service evaluation, an insight into the level of feelings and level of uncertainty of the reviewers when assigning the rating can be useful information to the product vendor or service provider. For example, through an analysis of online reviews using text mining and content analysis, a study concluded that customers paid more attention to bed, reception services and room size and decoration when reviewing star-rated hotels [16]. Combining this information with analysis from CUB results gives a hotel service provider an understanding of the determinants of customer satisfaction as well as the underlying degree of uncertainty and level of feelings when assessing these determinants.

In the existing literature, a significant amount of work on review rating prediction and text review analysis uses text and sentiment analysis [17], [18]. Further work on our proposed statistical modeling approach can be done by incorporating results from text analysis into the CUB model, such as the sentiment of the text review, thus making full use of the text review and review ratings in analyzing online review data.

ACKNOWLEDGMENT

Y.C. Low thanks the anonymous reviewer for providing constructive comments which significantly improved this paper.

REFERENCES

[1] A. Smith and M. Anderson, "Online shopping and e-commerce," *Pew Research Center*, 2016.

[2] A. D. Elia and D. Piccolo, "A mixture model for preferences data analysis," *Comp. Stats. Data Analysis*, pp. 917-934, 2005.

[3] P. Sur, G. Shmueli, S. Bose, and P. Dubey, "Modeling bimodal discrete data using conway-maxwell-poisson mixture models," *Journal of Business & Economic Statistics*, vol. 33, no. 3, pp. 352-365, 2015.

[4] A. D'Elia, "Modelling ranks using inverse hypergeometric distribution," *Statistical Modelling: An International Journal*, vol. 3, pp. 65-78, 2003.

[5] M. Iannario and D. Piccolo, "A new statistical model for the analysis of customer satisfaction," *Quality Technology and Management*, vol. 7, no. 2, pp. 149-168, 2007.

[6] R. A. Giancristofaro, V. Boatto, T. Tempesta, and D. Vecchiato, "Statistical modeling for the evaluation of consumers' preferences," *Comm. Stats. Sim. Comp.*, vol. 45, no. 5, pp. 1569-1582, 2016.

[7] G. Cicia, M. Corduas, T. D. Giudice, and D. Piccolo, "Valuing consumer preferences with the CUB model: A case study of fair trade coffee," *International Journal on Food System Dynamics*, vol. 1, pp. 82-93, 2010.

[8] R. Gambacorta and M. Iannario, "Measuring job satisfaction with CUB models," *Labour*, vol. 27, pp. 198-224, 2013.

[9] R: A language and environment for statistical computing, R Core Team, 2013, R Foundation for Statistical Computing, Vienna, Austria, [Online]. Available: http://www.R-project.org/

[10] M. Iannario, D. Piccolo, and R. Simone, "CUB: A class of mixture models for ordinal data," *R Package, Version 1.0,* 2016.

[11] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," in *Proc. 17th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'2011)*, 2011, pp. 618-626.

[12] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 4, 2015.

[13] M. Iannario, "Fitting measures for ordinal data models," *Quaderni di Statistica*, vol. 11, pp. 39-72, 2009.

[14] M. B. Oliver, "The respondent gender gap," in *Media Entertainment: The Psychology of its Appeal*, D. Zillmann and P. Vorderer, Eds. NJ: Lawrence Erlbaum Associates, 2000.

[15] D. Greenwood and J. R. Lippmann, "Gender and media: Content, uses, and impact," in *Handbook of Gender Research in Psychology*, J. C. Chrisler and D. R. McCreary, Eds. New York: Springer, pp. 643-669, 2010.

[16] H. Li, Q. Ye, and R. Law, "Determinants of customer satisfaction in the hotel industry: An application of online review analysis," *Asia Pacific Journal of Tourism Research*, vo. 18, no. 7, pp. 784-802, 2013.

[17] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content," presented at the Twelfth International Workshop on the Web and Databases, Providence, Rhode Island, USA, 2009.

[18] B. Wang, Y. Huang, and X. Li, "Combining review text content and reviewer-item rating matrix to predict review rating," *Computational Intelligence and Neuroscience*, pp. 1-11, 2016.

Y. C. Low graduated from the University of Malaya, Kuala Lumpur, Malaysia with a B.Sc. in statistics 2004, an M.Sc in statistics 2007 and a Ph.D in the field of applied and computational statistics in year 2016.

She has more than seven years of working experience as a lecturer. Currently she is a lecturer at Sunway University, Bandar Sunway, Selangor, Malaysia. Her research interest is in the area of discrete data analysis, applications of probabilistic models and computation-intensive statistical inference methods.

Y. C. Low is a member of the association for computing machinery since 2017 and Institut Statistik Malaysia since 2016.