A Statistical Analysis of Internet Ratings Data

final report

1 Introduction

As the Internet population continue to grow, the ratings data generated by Internet users increase rapidly. The ratings data are typically ordinal measurements on the quality of all kinds of items such as movies, books, consumer products, etc. These data have provided much information for consumers to make product choices. For a given product, most of the current graphical displays provide the number of votes it receives and a number of stars to represent the mean rating. Some displays may also present the frequency plot of ratings.

However, typical ratings data exhibits systematic tendencies for some raters to give higher scores than others, and some may not discriminate very well between products. The current displays all ignore the systematic differences across raters.

Ho and Quinn [2] proposed to use a Bayesian item response theory (IRT) model for ratings data. Then, they fit the model using Markov Chain Monte Carlo (MCMC), and proposed graphical displays based on estimates of model parameters. This model can account for the rater bias, and the proposed graphical displays are easily interpretable and incorporate statistical uncertainty in the ratings. While the methodology is reasonable, it has not been used by any website. The main difficulty is that, as the Internet data grow rapidly and the new ratings continuously arrive, the MCMC methods may not be computationally feasible to adjust model parameters.

Some researchers have employed Bayesian IRT models in psychological and educational studies; for example, Patz and Junker [3], Fox and Glas [1], and Wang et al. [4]. They all rely on the MCMC approach.

The MCMC approach is a nondeterministic method for approximate Bayesian inference. This approach often gives more accurate inference, but requires considerably more computation. Alternatively, there are deterministic approaches such as Laplace method, variational Bayes, expectation propagation, among others. They are part of mainstream machine learning methodologies. For many applications, deterministic methods produce

solutions of comparable accuracy to MCMC at greater speed. In fact, Ho and Quinn [2, Section 5] pointed out that their model-fitting approach (MCMC) needs to be modified so as to work efficiently on an industrial scale or real-time data.

The present paper extends the online Bayesian method in Weng and Lin [5] to the IRT model for ratings data. First, we obtain an efficient online algorithm to adjust the parameters in real-time as new ratings arrive. Secondly, the proposed method provides a reasonable alternative to MCMC approach.

2 Preliminaries

For the sake of being self-contained, we review the model-based approach in Ho and Quinn [2]. Suppose that there are R raters and P products. Let y_{rp} be the rating of product p by rater r, and $Y = [y_{rp}]$ the $R \times P$ rater-by-product matrix. Assume that y_{rp} is ordinal and takes values in $\{1, 2, ..., C\}$, where larger numbers indicate higher preference. In many cases, y_{rp} is not observed. It is sensible to introduce a missingness indicator z_{rp} and assume that the data are generated according to

$$y_{rp}^{\text{obs}} = \begin{cases} c & \Leftrightarrow y_{rp}^* \in (\gamma_{c-1}, \gamma_c] \text{ and } z_{rp} = 0 \\ \text{missing} & \Leftrightarrow z_{rp} = 1 \end{cases} \tag{1}$$

where y_{rp}^* is a latent variable and $\gamma_0 < \gamma_1 < \cdots < \gamma_C$ are cutpoints. Assume that $\gamma_0 = -\infty$, $\gamma_1 = 0$, and $\gamma_C = \infty$.

The latent variable y_{rp}^* is parametrized as

$$y_{rp}^* = \alpha_r + \beta_r \theta_p + \epsilon_{rp},\ \epsilon_{rp} \overset{iid}{\sim} \mathcal{N}(0, 1),\ r \in R, p \in P. \tag{2}$$

The parameter α captures the center of rater r's internal scale. For a more "critical" rater r, the α_r tends to be smaller. The parameter β_r captures how well rater r discriminates between low and high quality products. Ho and Quinn [2] constrained $\beta_r \in \Re^+\ \forall r$ to identify the sign of θ_r; otherwise, two different sets of parameter values can give the same model. A value of β_r near 0 means that rater r is unable to distinguish between low and high quality; a larger β_r means that rater r is discriminating. The parameter θ_p captures the latent quality of product p. With the constraint that $\beta_r \in \Re^+\ \forall r$, the value of y_{rp}^* is increasing in θ_p; and the interpretation of θ_p is that quality is increasing in $\theta_p\ \forall p$.

$$P(\mathbf{Y}^{\text{obs}}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) = \prod_{p,r:z_{rp}=0} \{\Phi(\gamma_{y_{rp}^{\text{obs}}} - \alpha_r - \beta_r\theta_p) - \Phi(\gamma_{y_{rp}^{\text{obs}}-1} - \alpha_r - \beta_r\theta_p)\}, \quad (3)$$

where $\gamma_{y_{rp}^{obs}} = \gamma_c \Leftrightarrow y_{rp}^{obs} = c$. The prior distribution for the parameters are assumed as follows: $\alpha_r \overset{iid}{\sim} \mathcal{N}(1,1)$, $\beta_r \overset{iid}{\sim} \mathcal{N}(-5,20)$ truncated to the positive half, $\gamma \overset{iid}{\sim}$ improper uniform, $\theta_p \overset{iid}{\sim} \mathcal{N}(0,1)$. Given M Markov Chain Monte Carlo samples $\{\alpha_r^{(m)}, \beta_r^{(m)}, \theta_p^{(m)}, \gamma^{(m)}\}_{m=1}^{M}$ from the posterior distribution $p(\alpha_r, \beta_r, \theta_p, \gamma|\mathbf{Y}^{obs})$, the posterior predictive density for y_{rp} can be approximated with:

$$\begin{aligned} & \mathrm{P}(y_{rp}^{rep} = c|\mathbf{Y}^{obs}) \\ \approx \quad & \frac{1}{M}\sum_{m=1}^{M}\{\Phi(\gamma_c^{(m)} - \alpha_r^{(m)} - \beta_r^{(m)}\theta_p^{(m)}) - \Phi(\gamma_{c-1}^{(m)} - \alpha_r^{(m)} - \beta_r^{(m)}\theta_p^{(m)})\} \end{aligned} \quad (4)$$

for $c = 1, ..., C$. Their proposed graphical displays depend on the posterior predictive probabilities for product p over all raters:

$$\tau_{pc} = \frac{1}{|R|}\sum_{r\in R} P(y_{rp}^{rep} = c|\mathbf{Y}^{obs}) \quad (5)$$

3 Main results

We propose to estimate τ_{pc} by

$$P(y_{pr} = c|Y^{\text{obs}}) \quad \approx \quad \Phi(\gamma_{c-1} - \mu_{\alpha r}^* - \mu_{\beta r}^*\mu_{\theta p}^*) - \Phi(\gamma_{c-1} - \mu_{\alpha r}^* - \mu_{\beta r}^*\mu_{\theta p}^*),$$

where $\mu_{\alpha r}^*$, $\mu_{\beta r}^*$, and $\mu_{\theta p}^*$ are current estimates of posterior means.

For the prior distributions, it is assumed that each α_r follows $\mathcal{N}(\mu_{\alpha r}, \sigma_{\alpha r}^2)$, each β_r follows $\mathcal{N}(\mu_{\beta r}, \sigma_{\beta r}^2)$, and each θ_p follows $\mathcal{N}(\mu_{\theta p}, \sigma_{\theta p}^2)$ with all parameters mutually independent. As γ values are not of primal interest, we suggest to set these values by observed proportions of $\{y^{\text{obs}} = c\}$ rather than treating them as unknown parameters. Priors such as improper uniform and normal may be used; however, they result in more complicated algorithms.

Now define

$$\alpha_r^* = \frac{\alpha_r - \mu_{\alpha r}}{\sigma_{\alpha r}}, \beta_r^* = \frac{\beta_r - \mu_{\beta r}}{\sigma_{\beta r}}, \theta_p^* = \frac{\theta_p - \mu_{\theta p}}{\sigma_{\theta p}}. \quad (6)$$

Denote $\alpha = (\alpha_1, ..., \alpha_R)^T$, $\beta = (\beta_1, ..., \beta_R)^T$, $\theta = (\theta_1, ..., \theta_P)^T$, and similarly for $\alpha^*, \beta^*, \theta^*$. Then, the posterior distribution of $(\alpha^*, \beta^*, \theta^*)$ given data Y^{obs} is

$$\begin{aligned} & p(\alpha^*, \beta^*, \theta^* | Y^{\text{obs}}) \\ \propto \quad & \phi(\alpha^*, \beta^*, \theta^*) \prod_{p,r:z_{pr}=0} \{\Phi(\gamma_{y^o_{pr}} - \alpha_r - \beta_r\theta_p) - \Phi(\gamma_{y^o_{pr}-1} - \alpha_r - \beta_r\theta_p)\}. \end{aligned}$$

In particular, if there is only one new observation $y^o_{pr} = c$, then the corresponding likelihood is

$$L(\alpha_r, \beta_r, \theta_p; y^o_{pr}) = \Phi(\gamma_c - \alpha_r - \beta_r\theta_p) - \Phi(\gamma_{c-1} - \alpha_r - \beta_r\theta_p). \tag{7}$$

Since (7) only involves $(\alpha_r, \beta_r, \theta_p)$, we need only to adjust posterior distribution of the three parameters $\alpha^*_r, \beta^*_r, \theta^*_p$. The posterior density of $(\alpha^*_r, \beta^*_r, \theta^*_p)$ given y^o_{pr} is

$$p(\alpha^*_r, \beta^*_r, \theta^*_p | y^o_{pr}) \propto \phi(\alpha^*_r, \beta^*_r, \theta^*_p) L(\alpha_r, \beta_r, \theta_p), \tag{8}$$

which of the form for a version of Stein's identity. Therefore, we apply it to obtain the posterior means and variances of $\alpha^*_r, \beta^*_r, \theta^*_p$. Then, by (6) we can easily convert these moments to that of $\alpha_r, \beta_r, \theta_p$; for instance,

$$\begin{aligned} E(\alpha_r | y^o_{pr}) &= \mu_{\alpha r} + \sigma_{\alpha r} E(\alpha^*_r | y^o_{pr}) \\ Var(\alpha_r | y^o_{pr}) &= \sigma^2_{\alpha r} Var(\alpha^*_r | y^o_{pr}). \end{aligned} \tag{9}$$

The following functions are needed when taking derivatives of L in (8). Let

$$\begin{aligned} M_a(x) &= \frac{\phi(x) - \phi(x-a)}{\Phi(x) - \Phi(x-a)} \\ N_a(x) &= \frac{x\phi(x) - (x-a)\phi(x-a)}{\Phi(x) - \Phi(x-a)} + \left(\frac{\phi(x) - \phi(x-a)}{\Phi(x) - \Phi(x-a)} \right)^2. \end{aligned} \tag{10}$$

For posterior means, applying a version of Stein's identity gives

$$E(\alpha^*_r | y^o_{pr}) = E\left(\frac{\partial L / \partial \alpha^*_r}{L} \middle| y^o_{pr} \right) = -\sigma_{\alpha r} E\left(M_a(x) | y^o_{pr} \right), \tag{11}$$

where

$$x = \gamma_c - \alpha_r - \beta_r\theta_p; \qquad a = \gamma_c - \gamma_{c-1} \tag{12}$$

and hence

$$M_a(x) = \frac{\phi(\gamma_c - \alpha_r - \beta_r\theta_p) - \phi(\gamma_{c-1} - \alpha_r - \beta_r\theta_p)}{\Phi(\gamma_c - \alpha_r - \beta_r\theta_p) - \Phi(\gamma_{c-1} - \alpha_r - \beta_r\theta_p)}.$$

Similarly, we have

$$E(\beta_r^*|y_{pr}^o) = E\left(\frac{\partial L/\partial \beta_r^*}{L}\Big| y_{pr}^o\right) = -\sigma_{\beta r} E\left(\theta_p M_a(x)\Big| y_{pr}^o\right) \tag{13}$$

$$E(\theta_p^*|y_{pr}^o) = E\left(\frac{\partial L/\partial \theta_p^*}{L}\Big| y_{pr}^o\right) = -\sigma_{\theta p} E\left(\beta_r M_a(x)\Big| y_{pr}^o\right). \tag{14}$$

The expressions of the posterior variance can be obtained similarly. We have

$$\begin{aligned} Var(\alpha_r^*|y_{pr}^o) =& 1 - (\sigma_{\alpha r})^2 E\left[N_a(x)\Big| y_{pr}^o\right], \\ Var(\beta_r^*|y_{pr}^o) =& 1 - (\sigma_{\beta r}\theta_p)^2 E\left[N_a(x)\Big| y_{pr}^o\right], \\ Var(\theta_r^*|y_{pr}^o) =& 1 - (\sigma_{\theta p}\beta_r)^2 E\left[N_a(x)\Big| y_{pr}^o\right]. \end{aligned}$$

The approximation of expectations are as in Weng and Lin [5], where the error incurred by this approximation is reduced by a scaling factor. Take $E(\alpha_r^*|y_{pr}^o)$ in (11) for illustration:

$$\begin{aligned} E(\alpha_r^*|y_{pr}^o) =& -\sigma_{\alpha r} E\left[M_a(x)\Big| y_{pr}^o\right] \\ \approx& -\frac{\sigma_{\alpha r}}{\nu} M_{a_\nu}(x_\nu), \end{aligned} \tag{15}$$

where a and x are as in (12), $\nu = \sqrt{1 + \sigma_{\alpha r}^2 + \sigma_{\beta r}^2 \mu_{\theta p}^2 + \sigma_{\theta p}^2 \mu_{\beta r}^2}$, and

$$a_\nu = \frac{\gamma_c - \gamma_{c-1}}{\nu} \qquad \text{and} \qquad x_\nu = \frac{\gamma_c - \mu_{\alpha r} - \mu_{\beta r}\mu_{\theta p}}{\nu}. \tag{16}$$

Together with (9), we have

$$E(\alpha_r|y_{pr}^o) \approx \mu_{\alpha r} - \frac{\sigma_{\alpha r}^2}{\nu}\left[\frac{\phi(\frac{\gamma_c - \mu_{\alpha r} - \mu_{\beta r}\mu_{\theta p}}{\nu}) - \phi(\frac{\gamma_{c-1} - \mu_{\alpha r} - \mu_{\beta r}\mu_{\theta p}}{\nu})}{\Phi(\frac{\gamma_c - \mu_{\alpha r} - \mu_{\beta r}\mu_{\theta p}}{\nu}) - \Phi(\frac{\gamma_{c-1} - \mu_{\alpha r} - \mu_{\beta r}\mu_{\theta p}}{\nu})}\right]. \tag{17}$$

Similar approximations give

$$Var(\alpha_r|y_{pr}^o) \approx \sigma_{\alpha r}^2\left\{1 - \left(\frac{\sigma_{\alpha r}}{\nu}\right)^2 N_{a_\nu}(x_\nu)\right\}. \tag{18}$$

The proposed algorithm is described below:

Step 1. Given current estimates $\mu_\alpha^{(t)}$, $\mu_\beta^{(t)}$, $\mu_\theta^{(t)}$, $\sigma_\alpha^{(t)}$, $\sigma_\beta^{(t)}$, $\sigma_\theta^{(t)}$, where $\mu_\alpha^{(t)} = (\mu_{\alpha 1}^{(t)}, \cdots, \mu_{\alpha R}^{(t)})^T$, $\mu_\beta^{(t)} = (\mu_{\beta 1}^{(t)}, \cdots, \mu_{\beta R}^{(t)})^T$, $\mu_\theta^{(t)} = (\mu_{\theta 1}^{(t)}, \cdots, \mu_{\theta P}^{(t)})^T$, and similarly for $\sigma_\alpha^{(t)}$, $\sigma_\beta^{(t)}$, $\sigma_\theta^{(t)}$.

Step 2. Given the $(t+1)$st observation $y_{pr} = c$. Calculate

$$\nu^{(t)} = \sqrt{1 + (\sigma_{\alpha r}^{(t)})^2 + (\sigma_{\beta r}^{(t)})^2(\mu_{\theta p}^{(t)})^2 + (\sigma_{\theta p}^{(t)})^2(\mu_{\beta r}^{(t)})^2}, \tag{19}$$

$$\omega^{(t)} = -\frac{1}{\nu^{(t)}} M_{a_\nu^{(t)}}(x_\nu^{(t)}), \tag{20}$$

$$\delta^{(t)} = \frac{1}{\left(\nu^{(t)}\right)^2} N_{a_\nu^{(t)}}(x_\nu^{(t)}), \tag{21}$$

where

$$a_\nu^{(t)} = \frac{\gamma_c - \gamma_{c-1}}{\nu^{(t)}} \qquad \text{and} \qquad x_\nu^{(t)} = \frac{\gamma_{c-1} - \mu_{\alpha r}^{(t)} - \mu_{\beta r}^{(t)}\mu_{\theta p}^{(t)}}{\nu^{(t)}}.$$

Step 3. Update parameters as below:

$$\begin{aligned}
\mu_{\alpha r}^{(t+1)} &= \mu_{\alpha r}^{(t)} + \left(\sigma_{\alpha r}^{(t)}\right)^2 \omega^{(t)} \\
\mu_{\beta r}^{(t+1)} &= \mu_{\beta r}^{(t)} + \left(\sigma_{\beta r}^{(t)}\right)^2 \mu_{\theta p}^{(t)} \omega^{(t)} \\
\mu_{\theta p}^{(t+1)} &= \mu_{\theta p}^{(t)} + \left(\sigma_{\theta p}^{(t)}\right)^2 \mu_{\beta r}^{(t)} \omega^{(t)} \qquad (22) \\
(\sigma_{\alpha r}^{(t+1)})^2 &= (\sigma_{\alpha r}^{(t)})^2 \max\left(1 - (\sigma_{\alpha r}^{(t)})^2 \delta^{(t)}, \kappa\right) \qquad (23) \\
(\sigma_{\beta r}^{(t+1)})^2 &= (\sigma_{\beta r}^{(t)})^2 \max\left(1 - (\sigma_{\beta r}^{(t)}\mu_{\theta p}^{(t)})^2 \delta^{(t)}, \kappa\right) \qquad (24) \\
(\sigma_{\theta p}^{(t+1)})^2 &= (\sigma_{\theta p}^{(t)})^2 \max\left(1 - (\sigma_{\theta r}^{(t)}\mu_{\beta r}^{(t)})^2 \delta^{(t)}, \kappa\right). \qquad (25)
\end{aligned}$$

4 Experiments

We consider the ratings of news outlets from Mondo Times (`http://www.mondotimes.com/`) used in Ho and Quinn [2]. Mondo Times is an online company that disseminates information about media outlets such as newspapers, magazines, radio stations, and television stations in 211 countries. Raters submit five-point ratings of the content quality of news outlets from awful, poor, average, very good, to great. The dataset used in Ho and Quinn [2], which features 1,515 products (news outlets) and 946 raters, is available from their Ratings package (available at `http://cran.r-project.org/`). The average number of ratings for a product is 3.0 and the average number rated by a rater is 4.8.

As in Ratings of Ho and Quinn [2], we remove raters who rate less than five products and remove products that are only rated by these raters. This ends up with 3249 ratings from 232 raters on 1344 products.

News outlets	MCMC: sub-Mondo	online: sub-Mondo	online: whole-Mondo
US News & World Report	-0.215 (0.481)	0.259 (0.241)	-0.148 (0.320)
Toronto Sun	-1.027 (0.392)	-0.684 (0.160)	-0.743 (0.224)
Toronto Star	0.397 (0.400)	0.321 (0.148)	0.215 (0.199)
San Diego Union Tribune	0.089 (0.482)	0.124 (0.280)	0.426 (0.224)
People	-1.793 (0.593)	-1.970 (0.689)	-1.909 (0.413)
PBS	1.223 (0.393)	1.695 (0.258)	0.809 (0.182)
Montana Magazine	-0.233 (0.358)	-0.071 (0.171)	0.060 (0.199)
London Sun	-2.140 (0.569)	-1.477 (0.267)	-1.380 (0.312)
Great Falls Tribune	-2.839 (0.770)	-1.817 (0.399)	-2.353 (0.484)
Daily Utah Chronicle	0.054 (0.818)	-0.161 (0.256)	-0.686 (0.668)
Colorado Public Radio	1.431 (0.602)	2.572 (0.700)	1.617 (0.484)
CNN	0.038 (0.192)	-0.055 (0.078)	0.312 (0.074)

Table 1: Posterior means of θ for twelve outlets.

As in Ho and Quinn [2], we assume that each α_r follows $\mathcal{N}(1,1)$ and each θ_p follows $\mathcal{N}(0,1)$. Instead of letting each β_r follow $N(-5,20)$ truncated to positive, we assume that each β_r follows $\mathcal{N}(1,20)$. All parameters are assumed to be mutually independent. Since the γ values are not the main interest here, we set them by the following steps: first, calculate the observed proportions $\#\{y_{pr}=c\}/N$, for $c=1,...,5$, where N is the number of observed ratings; next, find the z-scores corresponding to these areas; then, obtain approximations of the mean and variance of y^* and convert the z-scores to y^*'s scale. In our scenario, the new ratings arrive sequentially. So, the empirical proportions are based on just part of the data, or in a pilot study.

The initial parameter values in Algorithm 1 are set to be $\mu_{\alpha r}^{(0)}=1$, $\mu_{\beta r}^{(0)}=1$, $\mu_{\theta p}^{(0)}=1$, $\sigma_{\alpha r}^{(0)}=1$, $\sigma_{\beta r}^{(0)}=1$, $\sigma_{\theta p}^{(0)}=1$, for $r=1,...,R$ and $p=1,...,P$; and the positive lower bound κ in (23)-(25) is set to be 0.0001.

5 Conclusions

Ho and Quinn [2] have demonstrated the advantages of fitting the IRT models to Internet ratings data. However, for a real-time data pipeline that continuously collects new ratings

on new items, their MCMC approach is not computationally viable. Our proposed online method can adjust the model parameters in a large-scale problem.

References

[1] J.-P. Fox and C. A. W. Glas. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66:269–286, 2001.

[2] D. E. Ho and K. M. Quinn. Improving the presentation and interpretation of online ratings data with model-based figures. *The American Statistician*, 62(4):279–288, 2008.

[3] R. J. Patz and B. Junker. A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24:146–178, 1999.

[4] X. Wang, J. O. Berger, and D. S. Burdick. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1):126–153, 2013.

[5] R. C. Weng and C.-J. Lin. A Bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300, 2011.