Quantifying Aspect Bias in Ordinal Ratings using a Bayesian Approach

Lahari Poddar, Wynne Hsu, Mong Li Lee
School of Computing, National University of Singapore
{lahari, whsu, leeml}@comp.nus.edu.sg

Abstract

User opinions expressed in the form of ratings can influence an individual's view of an item. However, the true quality of an item is often obfuscated by user biases, and it is not obvious from the observed ratings the importance different users place on different aspects of an item. We propose a probabilistic modeling of the observed aspect ratings to infer (i) each user's aspect bias and (ii) latent intrinsic quality of an item. We model multi-aspect ratings as ordered discrete data and encode the dependency between different aspects by using a latent Gaussian structure. We handle the Gaussian-Categorical non-conjugacy using a stick-breaking formulation coupled with Pólya-Gamma auxiliary variable augmentation for a simple, fully Bayesian inference. We demonstrate the predictive ability of our model and its effectiveness in learning explainable user biases on two real world datasets.

1 Introduction

With easy availability of information on the web, user ratings have become increasingly important in molding people's perception of an item. However, an item typically has many aspects and not all aspects are equally important to all users. To some user, the *cleanliness* of a hotel is most important and he/she tends to rate this aspect stringently, but is lenient when rating *food* or *amenities*. Other users may have a different set of preferences and their aspect ratings for the same item could be vastly different. Hence, it is difficult to interpret conflicting ratings without knowing the underlying user biases. For an item with only few ratings this is aggravated, since even its average ratings are highly susceptible to the users' biases.

To enable proper interpretation of ratings, we propose a unified probabilistic model to quantify the underlying user biases for different aspects that lead to the observed ratings. We model the correlation between aspects by allowing a covariance structure among them. This is realistic since a user's bias, and in turn his rating, of one aspect may be correlated with another aspect. We detect the aspect preferences of individual users that are consistent across their ratings on different items. We can learn the aspect bias of users even with few ratings, by introducing latent user groups, based on the similarity of users' rating behavior on various aspects. For example, one user group might generally give low ratings for *ambience* while another user group gives high ratings for *food*.

Figure 1: Sample restaurant's ratings with user aspect bias.

Figure 1 shows an example application of the model where the learned user aspect bias is displayed beside the ratings. People with a negative bias tend to be more critical about the aspect and generally underrate the aspect than other users, whereas people with a positive bias for an aspect tend to overrate it. Knowing the aspect biases of individuals, users can better interpret their ratings. Furthermore this is beneficial for service providers to focus on improving the aspects of an item that consumers *truly* care about.

While existing works assume ratings to be continuous, in reality most observed ratings in e-commerce websites are ordinal in nature. Our model incorporates the ordinal nature of observed ratings through proper statistical formulation. However, modeling the ordinal nature of observed ratings as well the correlation between aspects introduce non-conjugacy into our model, making Bayesian inference very challenging.

To eliminate the non-conjugacy of Gaussian prior-Categorical likelihood, we utilize stick-breaking formulation with Pólya-Gamma auxiliary variable augmentation. The construction proposed in the paper is efficient and generic. It will help developing inference mechanisms for various applications that need to model ordinal data in terms of continuous latent variables with a correlation structure.

Experiments on two real world datasets from TripAdvisor and OpenTable demonstrate that the proposed model provides new insights in users' rating patterns, and outperforms state-of-the-art methods for aspect rating prediction.

To the best of our knowledge, this is the first work to model ordinal aspect ratings parameterized by latent multivariate continuous responses, with a simple, scalable and fully Bayesian inference.

2 Ordinal Aspect Bias Model

In this section, we describe the design of our Ordinal Aspect Bias model and present a Bayesian approach for inference.

Suppose we have J users and I items. Let R be the set of observed ratings where $\mathbf{r_{ij}}$ is an A dimensional vector denoting the rating of user j for item i on each of its aspects. Each $\mathbf{r_{ij}}$ is a discrete value between 1 and K corresponding to a K-level scale (*poor* to *excellent*). We assume that $\mathbf{r_{ij}}$ arises from a latent multivariate continuous response $\mathbf{v_{ij}}$ which is dependent on (i) the intrinsic quality of the item on the aspect and (ii) the bias of the user for the aspect.

The intrinsic quality of an item $\mathbf{z_i}$ is an A dimensional vector, drawn from a multivariate normal distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. We use multivariate normal distribution to account for the correlation among the subsets of aspects of an item. For example, it is highly unlikely for a hotel to have excellent *room* quality but very poor *cleanliness*, but it is possible to have a good *location* and average *food* choices. Such correlations among subset of aspects are captured by the covariance matrix. The parameters $(\boldsymbol{\mu}, \mathbf{\Sigma})$ are given a conjugate normal-inverse Wishart (NIW) prior.

The preference of a user for an aspect is captured by a bias vector $\mathbf{m_g}$ of dimension A. If a user places great importance on a particular aspect (e.g. *cleanliness*), this will be reflected in his ratings across all hotels. In other words, his rating on the *cleanliness* aspect will tend to be lower than the majority's rating for *cleanliness* on the same hotel. We cluster users with similar preferences into different user groups and associate a bias vector $\mathbf{m_g}$ with each group. The membership of a user j in a user group is denoted as s_j where s_j is drawn from a categorical distribution θ with a Dirichlet prior parameter α.

Given the intrinsic quality $\mathbf{z_i}$ and bias $\mathbf{m_g}$, the latent response $\mathbf{v_{ij}}$ is drawn from a multivariate Gaussian distribution with $\mathbf{z_i} + \mathbf{m_{s_j}}$ as mean and a hyper-parameter $\mathbf{B}$ as covariance. This is intuitive as a user's response depends on the item's intrinsic quality for an aspect as well as his own bias.

With the latent response $\mathbf{v_{ij}}$, we sample the observed rating vector $\mathbf{r_{ij}}$. Note that since the observed ratings are ordered and discrete, they should be drawn from a categorical distribution. However, the latent response $\mathbf{v_{ij}}$ is given a multivariate Gaussian prior. In order to have a fully Bayesian inference, we need to *transform* this categorical distribution to a Gaussian form to exploit conjugacy. This is the central technical challenge for our proposed model.

We develop a stick-breaking mechanism with logit function to map the categorical likelihood to a binomial form. Thereafter, leveraging the recently developed Pólya-Gamma auxiliary variable augmentation scheme [Polson *et al.*, 2013], the binomial likelihood is transformed to Gaussian, thus establishing conjugacy and enabling an effective posterior inference. The generative process of the model is as follows:

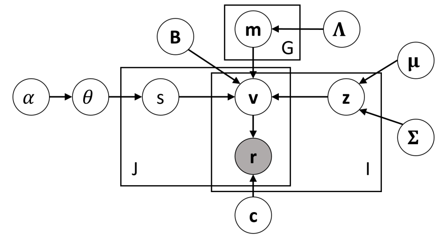


Figure 2: Ordinal Aspect Bias Model

1. Draw a multinomial group distribution θ from Dirichlet (α).
2. For each group $g \in 1, \cdots, G$ draw a bias offset $\mathbf{m_g}$ from $N_A(0, \Lambda)$
3. For each user $j \in 1, \cdots, J$, sample a group s_j from Cat (θ)
4. For each item $i \in 1, \cdots, I$, sample an intrinsic rating $\mathbf{z_i}$ from $N_A(\boldsymbol{\mu}, \mathbf{\Sigma})$
5. For each rating $\mathbf{r_{ij}} \in R$
   (a) draw latent continuous rating $\mathbf{v_{ij}}$ from $N_A(\mathbf{z_i} + \mathbf{m_{s_j}}, \mathbf{B})$
   (b) draw observed ordinal rating $\mathbf{r_{ij}}$ from Cat $(SB(\mathbf{v_{ij}}, \mathbf{c}))$

where $SB(\mathbf{v_{ij}}, \mathbf{c})$ refers to the stick-breaking parametrization of the continuous response $\mathbf{v_{ij}}$ using cut-points $\mathbf{c}$. Figure 2 shows the proposed graphical model using plate notation.

2.1 Stick-Breaking Likelihood

We first discuss how to map the categorical likelihood of $\mathbf{v_{ij}}$, denoted as $Lik(\mathbf{v_{ij}})$, to a binomial form.

Let r_{ija} denote the observed ordinal rating of item i, by user j on aspect a, and is drawn from a categorical distribution over K categories. Since the categories are ordered, we utilize a stick-breaking parameterization for the probabilities $P(r_{ija} = k)$ where $k \in \{1, \cdots, K\}$. Suppose we have a unit length stick where the continuum of points on this stick represents the probability of an event occurring. If we break this stick at some random point p, then we have a probability mass function over two outcomes (with probabilities p and $1 - p$). By breaking the stick multiple times, we obtain a probability mass function over multiple categories.

Let $\mathbf{c} = \{c_1, \cdots, c_{K-1}\}$ be a cut-point vector where $c_1 < c_2 < \cdots < c_{K-1}$ represent the boundaries between the ordered categories. The probability of each ordinal rating r_{ija} being assigned the categorical value k, is parametrized using a function of the covariate $\eta_{ija}^k = c_k - v_{ija}$. Then the probability of observing the vector of ratings $\mathbf{r_{ij}}$ is a product of probabilities of observing each of the aspect ratings r_{ija} given the values of $\boldsymbol{\eta_{ija}}$. Hence the likelihood of $\mathbf{v_{ij}}$ is:

$$Lik(\mathbf{v_{ij}}) = P(\mathbf{r_{ij}}|\mathbf{v_{ij}}, \mathbf{c}) = P(\mathbf{r_{ij}}|\boldsymbol{\eta_{ij}}) = \prod_{a=1}^{A} P(r_{ija}|\boldsymbol{\eta_{ija}}) \quad (1)$$

To squash $\boldsymbol{\eta_{ija}}$ within [0,1] we use a sigmoid function on it, denoted by $f(x) = \frac{e^x}{1+e^x}$. Sigmoid function enables us to use Pólya-Gamma augmentation scheme to handle the non-conjugacy subsequently. For identifiability, we set $f(\eta_{ija}^K) = 1$. The stick-breaking likelihood can be written as:

$$P(r_{ija} = k) = \prod_{k' < k} (1 - f(\eta_{ija}^{k'})) f(\eta_{ija}^{k}) \quad (2)$$

By encoding r_{ija} with a 1-of-K vector $\mathbf{x_{ija}}$ where

$$x_{ija}^k = \begin{cases} 1 & \text{if } r_{ija} = k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

we now rewrite the likelihood of $\mathbf{v_{ij}}$ in binomial terms:

$$P(r_{ija}|\boldsymbol{\eta_{ija}}) = P(\mathbf{x_{ija}}|\boldsymbol{\eta_{ija}}) = \prod_{k=1}^{K-1} Binom(x_{ija}^k|N_{ija}^k, f(\eta_{ija}^k))$$

$$\text{where} \quad N_{ija}^k = 1 - \sum_{k'<k} x_{ija}^{k'} \quad (4)$$

2.2 Pólya-Gamma Variable Augmentation

Next, we explain how to transform the binomial likelihood to a Gaussian form via Pólya-Gamma (PG) auxiliary variable augmentation scheme. The integral identity at the heart of the PG augmentation is:

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega \quad (5)$$

where $\kappa = a - b/2$, $b > 0$ and $\omega \sim PG(b, 0)$.

By expanding the binomial likelihood in Eqn. 4, we get

$$\begin{aligned} P(\mathbf{x_{ija}}|\boldsymbol{\eta_{ija}}) &= \prod_{k=1}^{K-1} \binom{N_{ija}^k}{x_{ija}^k} (f(\eta_{ija}^k))^{x_{ija}^k} (1 - f(\eta_{ija}^k))^{N_{ija}^k - x_{ija}^k} \\ &= \prod_{k=1}^{K-1} \binom{N_{ija}^k}{x_{ija}^k} \frac{(e^{\eta_{ija}^k})^{x_{ija}^k}}{(1+e^{\eta_{ija}^k})^{N_{ija}^k}} \end{aligned} \quad (6)$$

Using the integral identity of PG augmentation, we can now rewrite the categorical likelihood of $\mathbf{v_{ij}}$ as:

$$\begin{aligned} Lik(\mathbf{v_{ij}}) &= \prod_{a=1}^{A} P(\mathbf{x_{ija}}|\boldsymbol{\eta_{ija}}) \quad (7) \\ &\propto \prod_{a=1}^{A} \prod_{k=1}^{K-1} e^{\kappa_{ija}^k \eta_{ija}^k} \int_0^{\infty} e^{-\omega_{ija}^k (\eta_{ija}^k)^2/2} p(\omega_{ija}^k) d\omega_{ija}^k \end{aligned}$$

where $\kappa_{ija}^k = x_{ija}^k - N_{ija}^k/2$, $\psi_{ija}^k = \eta_{ija}^k$ and $p(\omega_{ija}^k)$ is $PG(N_{ija}^k/2, 0)$ independent of ψ_{ija}^k.

By property of PG distribution, we can draw the auxiliary variable ω_{ija}^k from $PG(N_{ija}^k, \eta_{ija}^k)$. Conditioning on $\boldsymbol{\omega_{ij}}$, $Lik(\mathbf{v_{ij}})$ can be transformed to a Gaussian form:

$$\begin{aligned} Lik(\mathbf{v_{ij}}) &\propto \prod_{k=1}^{K-1} \prod_{a=1}^{A} e^{\kappa_{ija}^k \eta_{ija}^k} e^{-\omega_{ija}^k (\eta_{ija}^k)^2/2} \quad (8) \\ &\propto \prod_{k=1}^{K-1} \prod_{a=1}^{A} exp\{\kappa_{ija}^k (c_k - v_{ija}) - \omega_{ija}^k (c_k - v_{ija})^2/2\} \\ &\propto \prod_{k=1}^{K-1} \prod_{a=1}^{A} exp\{-\omega_{ija}^k ((c_k - v_{ija}) - \frac{\kappa_{ija}^k}{\omega_{ija}^k})^2\} \\ &\propto \prod_{k=1}^{K-1} exp\{-\frac{1}{2}(\frac{\boldsymbol{\kappa_{ij}^k}}{\boldsymbol{\omega_{ij}^k}} - (\mathbf{c_k} - \mathbf{v_{ij}}))^T \boldsymbol{\Omega_{ij}^k} (\frac{\boldsymbol{\kappa_{ij}^k}}{\boldsymbol{\omega_{ij}^k}} - (\mathbf{c_k} - \mathbf{v_{ij}}))\} \end{aligned}$$

where $\boldsymbol{\kappa_{ij}^k}, \boldsymbol{\omega_{ij}^k}$ are vectors of dimension A, $\boldsymbol{\Omega_{ij}^k}$ is a diagonal matrix of $(\omega_{ij1}^k, \omega_{ij2}^k, \cdots, \omega_{ijA}^k)$.

Here, we assume the values in the A-dimensional cut-point vector $\mathbf{c_k}$ are all equal to c_k. In practice, if we need different cut-points for different aspects, $\mathbf{c_k}$ can be set accordingly.

2.3 Bayesian Inference

Finally, we describe the sampling of user groups $\mathbf{s}$, bias offset of user groups $\mathbf{m}$, intrinsic ratings $\mathbf{z}$, cut-points $\mathbf{c}$ and latent continuous ratings $\mathbf{v}$ using fully Bayesian MCMC inference. We factor the joint probability of these variables as:

$$P(\mathbf{r}, \mathbf{v}, \mathbf{m}, \mathbf{z}, \mathbf{s}, \mathbf{c}) = P(\mathbf{r}|\mathbf{v}, \mathbf{c}) P(\mathbf{v}|\mathbf{m}, \mathbf{z}, \mathbf{s}) P(\mathbf{c}) P(\mathbf{z}) P(\mathbf{s}) P(\mathbf{m})$$

Sampling Bias Offset of User Groups. For each user group g, we sample its bias offset $\mathbf{m_g}$ from the Gaussian posterior:

$$P(\mathbf{m_g}|\boldsymbol{\Lambda}, \mathbf{v}, \mathbf{z}) \propto P(\mathbf{m_g}|\boldsymbol{\Lambda}) \prod_{j \in J[g]} \prod_{i \in I[j]} P(\mathbf{v_{ij}}|\mathbf{m_g}, \mathbf{z_i}, \mathbf{B})$$

where $J[g]$ is the set of users belonging to group g and $I[j]$ is the subset of items rated by user j.

Since the prior is a multivariate Gaussian $N_A(0, \Lambda)$ and the observations $\mathbf{v_{ij}}$ are also drawn from a multivariate Gaussian $N_A(\mathbf{z_i} + \mathbf{m_g}, \mathbf{B})$, the posterior of $\mathbf{m_g}$ is given by a Gaussian $N_A(\hat{\mathbf{m_g}}, \hat{\boldsymbol{\Lambda_g}})$ with

$$\begin{aligned} \hat{\mathbf{m_g}} &= \hat{\boldsymbol{\Lambda_g}}(\mathbf{B}^{-1} \sum_{j \in J[g]} \sum_{i \in I[j]} (\mathbf{v_{ij}} - \mathbf{z_i})) \\ \hat{\boldsymbol{\Lambda_g}} &= (n_g \mathbf{B}^{-1} + \boldsymbol{\Lambda})^{-1} \end{aligned}$$

where n_g is the total number of ratings observed for users belonging to group g.

Sampling User Groups. We integrate out the group distribution θ by exploiting Dirichlet-Multinomial conjugacy, and sample the group of each user j as:

$$P(s_j|\alpha, \mathbf{m}, \mathbf{v}) \propto P(s_j|\alpha) \prod_{i \in I[j]} P(\mathbf{v_{ij}}|\mathbf{m_{s_j}}, \mathbf{z_i}, \mathbf{B})$$

where $I[j]$ are the subset of items rated by user j, the prior $P(s_j|\alpha)$ is given by the Dirichlet distribution. The likelihood is the multinomial distribution given by the probability of observing all the ratings of the user j given bias m_{s_j}.

Sampling Intrinsic Ratings. Similar to the bias offsets of user groups, we sample intrinsic rating $\mathbf{z_i}$ of each item i from a Gaussian distribution $N_A(\hat{\boldsymbol{\mu_i}}, \hat{\boldsymbol{\Sigma_i}})$ where

$$\begin{aligned} \hat{\boldsymbol{\mu_i}} &= \hat{\boldsymbol{\Sigma_i}}(\mathbf{B}^{-1} \sum_{j \in J[i]} (\mathbf{v_{ij}} - \mathbf{m_{s_j}}) + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) \\ \hat{\boldsymbol{\Sigma_i}} &= (n_i \mathbf{B}^{-1} + \boldsymbol{\Sigma})^{-1} \end{aligned}$$

where n_i is the total number of ratings observed for item i and $J[i]$ is the subset of users who have rated item i. The prior parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of the intrinsic ratings are given a conjugate Normal-Inverse Wishart (NIW) prior and sampled.

Sampling Latent Continuous Ratings. The latent continuous ratings, $\mathbf{v_{ij}}$ have a Gaussian prior $N_A((\mathbf{z_i} + \mathbf{m_{s_j}}), \mathbf{B})$ and a categorical likelihood $P(r_{ija}|\ \mathbf{v_{ij}}, \mathbf{c})$. We have transformed the categorical likelihood to the conditional Gaussian form (recall Eqn. 8). The posterior can be formulated as:

$$\begin{aligned} P(\mathbf{v_{ij}}) &\propto P(\mathbf{v_{ij}}|\mathbf{m_{s_j}}, \mathbf{z_i}, \mathbf{B}) * Lik(\mathbf{v_{ij}}|\boldsymbol{\omega}, \mathbf{r_{ij}}, \mathbf{c}) \\ &\propto exp\{-\frac{1}{2}(\mathbf{v_{ij}} - (\mathbf{z_i} + \mathbf{m_{s_j}}))^T \mathbf{B}^{-1} (\mathbf{v_{ij}} - (\mathbf{z_i} + \mathbf{m_{s_j}}))\} \\ &* \prod_{k=1}^{K-1} exp\{-\frac{1}{2}(\frac{\boldsymbol{\kappa_{ij}^k}}{\boldsymbol{\omega_{ij}^k}} - (\mathbf{c_k} - \mathbf{v_{ij}}))^T \boldsymbol{\Omega_{ij}^k} (\frac{\boldsymbol{\kappa_{ij}^k}}{\boldsymbol{\omega_{ij}^k}} - (\mathbf{c_k} - \mathbf{v_{ij}}))\} \end{aligned}$$

Since both the prior and likelihood are now Gaussian, we have the following Gibbs sampler:

$$\mathbf{v_{ij}} \sim N_A(\boldsymbol{\mu_{ij\omega}}, \boldsymbol{\Sigma_{ij\omega}})$$
$$\omega_{\mathbf{ija}} \sim PG(\mathbf{N_{ija}}, \mathbf{v_{ija}} - \mathbf{c})$$

where

$$\boldsymbol{\mu_{ij\omega}} = \mathbf{B^{-1}}(\mathbf{z_i} + \mathbf{m_{s_j}}) + \sum_{\mathbf{k=1}}^{\mathbf{K-1}} \boldsymbol{\Omega_{ij}^{k}}(\mathbf{c_k} - \frac{\boldsymbol{\kappa_{ij}^{k}}}{\boldsymbol{\omega_{ij}^{k}}})$$
$$\boldsymbol{\Sigma_{ij\omega}} = \mathbf{B^{-1}} + \sum_{\mathbf{k=1}}^{\mathbf{K-1}} \boldsymbol{\Omega_{ij}^{k}}$$

Sampling Cut-Points. Sigmoid function in the stick-breaking formulation allows us to sample cut-points while ensuring their relative order without additional constraints. Figure 3 shows probability distributions for simulated cut-points.

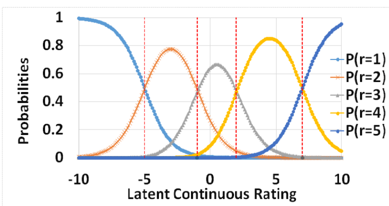


Figure 3: Category probabilities for cut-points (-5,-1,2,7)

The following lemma gives the relationship between cut-points, latent continuous ratings, and the observed ratings.

Lemma 2.1. *If* $v_{ija} > c_k - ln\ (1 - e^{-(c_{k+1} - c_k)})$*, then* $P(r_{ija} = k+1) > P(r_{ija} = k)$.

Proof. Let $\delta_k \geq -\ ln\ (1 - e^{-(c_{k+1} - c_k)})$. By replacing v_{ija} with $(c_k + \delta_k)$ in Eqn. 2, we have

$$P(r_{ija} = k) = \prod_{q<k}(1 - f(c_q - c_k - \delta_k))(f(c_k - c_k - \delta_k))$$
$$= \prod_{q<k}(1 - f(c_q - c_k - \delta_k))(f(-\delta_k))$$
$$P(r_{ija} = k+1) = \prod_{q<k}(1 - f(c_q - c_k - \delta_k))(1 - f(-\delta_k))(f(c_{k+1} - c_k - \delta_k))$$

Taking the ratio, we have

$$\frac{P(r_{ija} = k+1)}{P(r_{ija} = k)} = \frac{(1 - f(-\delta_k))(f(c_{k+1} - c_k - \delta_k))}{f(-\delta_k)}$$
$$= (\frac{e^{\delta_k}}{1 + e^{\delta_k}} * \frac{1}{1 + e^{c_k + \delta_k - c_{k+1}}})/(\frac{1}{1 + e^{\delta_k}})$$
$$= \frac{e^{\delta_k}}{1 + e^{c_k - c_{k+1} + \delta_k}}$$

Since $\delta_k \geq -ln(1 - e^{-(c_{k+1} - c_k)})$, we see that $\frac{e^{\delta_k}}{1 + e^{c_k - c_{k+1} + \delta_k}} > 1$. Hence, $P(r_{ija} = k+1) > P(r_{ija} = k)$. □

We have shown that $P(r_{ija} = k+1) > P(r_{ija} = k)$ when $v_{ija} \geq (c_k - ln(1 - e^{-(c_{k+1} - c_k)})$. Similarly, $P(r_{ija} = k) > P(r_{ija} = k-1)$ when $v_{ija} \geq (c_{k-1} - ln(1 - e^{-(c_k - c_{k-1})}))$. This implies that, when v_{ija} is within the range $(c_{k-1} - ln(1 - e^{-(c_k - c_{k-1})}), c_k - ln(1 - e^{-(c_{k+1} - c_k)})]$, then $P(r_{ija} = k)$ has the maximum probability over all other categories. In other words, for v_{ija} in the stated range, we have $\text{argmax}_{k'} P(r_{ija}|v_{ija}, k') = k$.

Hence, given the sampled values of v_{ija} we can constrain the possible set of values for the cut-points. We sample cut-point c_k from a uniform distribution within the range:

$$c_k \sim U[max\{v_{ija} | \underset{k'}{\text{argmax}}\, P(r_{ija}|v_{ija}, k') = k\} - ln(1 - e^{-(c_k - c_{k-1})}),$$
$$min\{v_{ija} | \underset{k'}{\text{argmax}}\, P(r_{ija}|v_{ija}, k') = k+1\} - ln(1 - e^{-(c_k - c_{k-1})})]$$

3 Experiments

For evaluation we use hotel ratings from TripAdvisor [Wang *et al.*, 2011] and restaurant ratings from Opentable.com. We crawled OpenTable.com for all the restaurant ratings in New York Tri-State area. Table 1 shows the details of the datasets.

Dataset	# Items	# Users	# Ratings	Aspects rated
TripAdvisor	12,773	781,403	1,621,956	Service, Value, Room, Location
OpenTable	2805	1997	73,469	Ambience, Food, Service, Value

Table 1: Statistics of experimental datasets.

3.1 Rating Prediction

One application of Ordinal Aspect Bias model is predicting observed aspect ratings. We perform five-fold cross validation on user-item pairs, and take expected value of an aspect rating as the predicted rating. Note that all the aspect ratings for the same user-item pair will be in the same training or test set. By default, the number of user groups are set to 10. For comparison, we also implemented the following models:

- **Continuous Aspect Bias model** is the continuous variant of our model where observed ratings are assumed to be continuous. Observed ratings are drawn from a (conjugate) multivariate Gaussian distribution, with mean as the true rating of the item offset with the bias of the user's group.
- **Ordinal** and **Continuous No Bias model** assume users are not biased. The observed ratings for an item are drawn from only the true rating of the item.
- **Ordinal** and **Continuous Global Bias model** assume all users have the same bias. All ratings for an item are drawn from the true rating of the item offset with a global bias.

Model	TripAdvisor Data		OpenTable Data	
	log LL	RMSE	log LL	RMSE
Ordinal Aspect Bias	**-557.08**	**1.00**	**-493.79**	**1.03**
Continuous Aspect Bias	-1050.32	3.13	-560.14	2.21
Ordinal No Bias	-689.76	1.47	-546.25	1.95
Continuous No Bias	-1904.64	3.52	-651.16	2.39
Ordinal Global Bias	-2438.52	2.85	-570.28	2.37
Continuous Global Bias	-2632.95	3.91	-595.62	2.41

Table 2: Log likelihood and RMSE results on the test set. All comparisons are statistically significant (paired *t-test* with $p < 0.0001$).

Table 2 shows mean log likelihood and RMSE (root mean square error) on test data. For both datasets Ordinal Aspect Bias model performs the best, demonstrating the need to consider both user bias and the proper ordinal nature of ratings.

Next, we compare our proposed model with state-of-the-art rating prediction models, namely, PMF [Salakhutdinov and Mnih, 2007], BPMF [Salakhutdinov and Mnih, 2008], URP [Marlin, 2003; Barbieri, 2011], SVD++ [Koren, 2008] and BHFree [Pierre, 2012]. We used the best parameter settings published on LibRec.net website. We also compare with OrdRec [Koren and Sill, 2013] which can wrap collaborating

Model	TripAdvisor Data								OpenTable Data							
	Service		Value		Room		Location		Ambience		Food		Service		Value	
	RMSE	FCP	RMSE	FCP	RMSE	FCP	RMSE	FCP	RMSE	FCP	RMSE	FCP	RMSE	FCP	RMSE	FCP
PMF	2.006	0.501	1.933	0.526	1.836	0.592	2.127	0.603	2.584	0.524	2.232	0.530	2.388	0.511	2.151	0.521
BPMF	1.414	0.586	1.373	0.571	1.314	0.614	1.209	0.651	1.154	0.490	0.992	0.532	1.426	0.498	1.302	0.519
URP	1.179	0.489	1.156	0.515	1.194	0.513	1.001	0.492	0.952	0.557	0.818	0.551	1.144	0.522	1.120	0.514
SVD++	**1.064**	0.578	1.079	0.562	1.093	0.639	0.894	0.665	**0.944***	0.525	0.831	0.544	**1.088**	0.544	1.131	0.517
BHFree	1.143	0.553	1.199	0.582	1.124	0.624	1.007	0.671	0.956	0.483	0.812	0.499	1.151	0.512	1.096	0.495
LARA	1.193	0.576	1.221	0.531	1.087	0.558	1.170	0.672	1.150	0.538	2.242	0.514	2.444	0.549	1.089	0.526
OrdRec + SVD++	1.348	0.619	1.344	0.613	1.359	0.654	1.173	0.702	1.337	0.672	1.121	0.613	1.533	0.618	1.521	0.623
AspectBias	1.067	**0.646***	**1.063***	**0.645***	**1.045**	**0.678***	**0.854***	**0.717**	0.953	**0.854***	**0.787***	**0.850***	1.134	**0.842***	**1.043***	**0.864***

Table 3: RMSE and FCP results for rating prediction. ”*” denotes statistical significance with the runner up for $p < 0.005$

filtering methods to tackle ordinal rating. Since these models cannot predict multiple aspect ratings for a user-item pair, we train them separately for each aspect. We further compare with LARA [Wang *et al.*, 2011] which models latent aspect ratings using review texts. Since RMSE cannot capture personalization or ordinal rating values, we use FCP [] to measure the fraction of correctly ranked pair of items for each user. Table 3 shows the results. We see that the proposed model outperforms state-of-the art methods in most cases.

Method	TripAdvisor Data	OpenTable Data
PMF	0.016	0.142
BPMF	0.219	0.133
URP	0.238	0.177
SVD++	0.364	0.201
BHFree	0.359	0.205
LARA	0.289	0.152
OrdRec + SVD++	0.148	0.262
OrdinalAspectBias	**0.404**	**0.298**

Table 4: Pearsons Correlation of aspect ranking

The relative ranking of aspects for a user-item pair is also important to understand which aspects of an item the user liked better. Table 4 shows the Pearson correlation coefficient of aspect ranking for a user-item pair, compared to its ground truth ranking. Clearly, Ordinal Aspect Bias model outperforms all other methods for the task of relative ranking of aspects. This validates that our model is able to learn aspect rating behavior of users accurately.

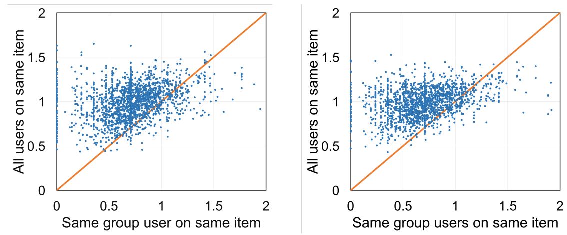
3.2 Evaluation of User Groups

A significant advantage of our model is that it can infer latent user groups depending on their rating behaviors across multiple items. In this set of experiments, we show that if users are assigned to the same group, then their ratings on the same items for the same aspects are indeed similar.

We look at the standard deviation of the set of users belonging to a group who have rated the same item [Wang *et al.*, 2011]. For each aspect of each item, we compare the standard deviation of ratings of each user group with that of a control group of all users who have rated the item.

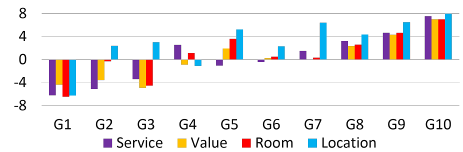
Figure 4 shows the scatter plots of the standard deviations for both datasets. We observe that most of the points lie above the line $y = x$, indicating that users who belong to the same group have smaller standard deviation compared to the control group. This implies that the latent user groups obtained by the proposed model can effectively cluster users who give similar aspect ratings to the same item.

Figure 5 shows the mean ratings of 10 user groups after scaling the ratings to the range [-10, 10]. For the TripAdvisor dataset, we observe that the first user group seems to be quite critical whereas the last three user groups are positive. We also see correlation of aspect biases for different groups. For example, group 5 and 7 seem to have similar biases for *Room* and *Value* whereas group 4 is demanding about *Value* and *Location*. Considering all the ratings of the users belonging to group 5 and 7, we see that their ratings for *Value* are indeed most correlated with their ratings for *Room* than other aspects. On the other hand for group 4 their ratings for *Value* are highly correlated with their ratings for *Location*. This suggests that for good *Value* for money, some users prefer good *Location* while some prioritize better *Room* quality and by modeling the covariance structure among aspects we are able to uncover such dependencies. For the OpenTable dataset, we see that users in group 4 who are particular about *Ambience* are also demanding about *Service* and *Value*.

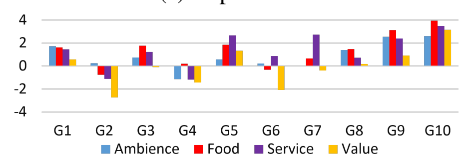


(a) TripAdvisor (b) OpenTable

Figure 4: Scatter plot of standard deviations of aspect ratings.



(a) TripAdvisor

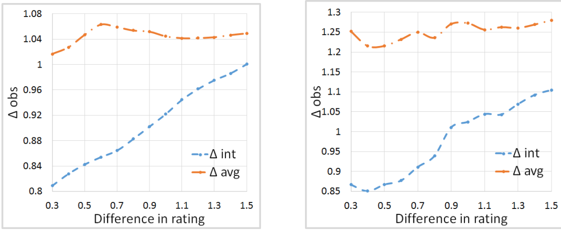


(b) OpenTable

Figure 5: Mean bias value of user groups.

3.3 Intrinsic Quality of Items

Often one forms a judgment about the quality of an item by the average rating it has received. However, if an item has received only a few ratings, it is difficult to form an accurate opinion concerning its quality. In this set of experiments, we show that the intrinsic quality, learned by the proposed model,



(a) OpenTable (b) TripAdvisor

Figure 6: Correlation with Δobs

is correlated with users' perception of the item's true quality, even for items with few ratings.

We focus on items with less than 30 ratings and whose intrinsic quality and average rating for an aspect differ by at least 0.5. Since an item's true quality is unknown, we estimate it by the relative difference in the observed ratings of the same user on a pair of items. This is because if the qualities of two items are similar, a user will rate them similarly.

For each pair of items rated by the same user on the same aspect, let their difference in observed ratings be Δobs, difference between their average ratings be Δavg and difference between the learned intrinsic ratings be Δint. Figure 6 shows the correlation between Δobs and Δint, as well as the correlation between Δobs and Δavg aggregated over all aspects. We observe that for both datasets, as Δint increases, Δobs also increases. However, Δavg remains almost constant. This indicates that Δobs is closely correlated with Δint, whereas Δavg appears to be independent of Δobs. This confirms that the learned intrinsic rating is better able to reflect users' perception of the true quality of an item compared to using average ratings of the items.

3.4 Case Study

Finally, we present the reviews of a user from OpenTable to demonstrate that the aspect bias learned by our model correlates with their review texts (see Figure 7). The user is from group G2 in Figure 5 that is particularly critical about *Value*.

From the reviews of this user, as well as the reviews of randomly selected users from other groups for the same item, we see that the user from group 2 is indeed critical. We further confirm this observation by manually going through 100 randomly sampled reviews and tabulate the sentiment distribution of each item. We observe that the user is consistently critical even though the majority opinion is positive. This strengthens the fact that the group bias captured by our model is accurate and can help us better interpret a users' rating.

4 Related Work

Existing works on aspect rating prediction use reviews to analyze latent aspect ratings [Wang and Ester, 2014; Wang *et al.*, 2011] and ignore the *explicit* aspect ratings provided by users. While the widely used CF approaches for rating prediction view ratings as *continuous* values and do not encode aspect dependencies [Lee and Seung, 2001; Koren, 2008; Hofmann, 2003; Marlin, 2003; Pierre, 2012; Salakhutdinov and Mnih, 2008; 2007; Pennock *et al.*, 2000].

There have been very few attempts to address the ordinal nature of ratings. A model using regression is developed in [Stern *et al.*, 2009] to handle ordinal ratings as a special case. The work in [Koren and Sill, 2013] proposes a *wrapper* around a CF method for ordinal data. Both of these works use a logit model for ordinal regression. In contrast, most statistical approaches handle ordinals using ordinal probit model [Albert and Chib, 1993; Rossi *et al.*, 2001; Muthukumarana and Swartz, 2014]. Although they allow Bayesian inference it necessitates using truncated Gaussian distributions and forced ordering of cut-off points. This leads to complicated and even sub-optimal inference.

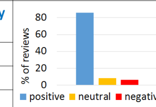
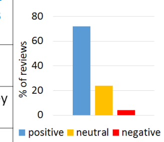
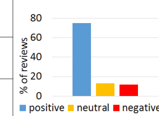
Item	Reviews	Sentiment Distribution of Reviews
Item 1	**... this place is a trap, very expensive for the poor food quality and lack of service...**	
	... very tasty Greek food. Service was very friendly...	
	...prices were pretty reasonable for the portion and quality...	
	...very good lunch spot.. prices were very reasonable....	
Item 2	**...it is a pity-current status deserves a total reconsideration of how to keep the good name it had... waiters are tired, food is far from acceptable...**	
	...very pleasant ambience, excellent food and service.. very friendly home like atmosphere...	
	...while restaurants come and go, with 50 years in business they are definitely doing it right.. never disappoints..	
	...superb menu with generous portions.. highly recommended..	
Item 3	**...It's a shame that such famous owner offers such inferior food. Amounts served are ridiculous, quality barely acceptable, exorbitant price...**	
	...what appeared to be a small portion was exactly well served and balanced	
	The portions were generous, food was tasty and splendidly presented..	
	Exquisite food, expert service and excellent advice on wine..	

Figure 7: Reviews of user belonging to "critical" group contrasted with other reviews on the same items

The authors of [Virtanen and Girolami, 2015] used stick-breaking formulation to parameterize the underlying continuous rating. However, since the non-conjugacy made an MCMC sampling non-trivial, they performed an approximate variational Bayesian inference. For correlated topic models [Chen *et al.*, 2013], Pólya-Gamma auxiliary variable augmentation is used with logistic-normal transformation, whereas the work in [Khan *et al.*, 2012] used stick-breaking likelihood for categorical data. None of these works use stick-breaking likelihood with Pólya-Gamma variable augmentation to exploit conjugacy to facilitate Gibbs sampling.

5 Conclusion

We have presented a novel approach to understand users' aspect bias, while capturing aspect dependencies as well as the proper ordinal nature of user responses. Our construction of the stick-breaking likelihood coupled with Pólya-Gamma auxiliary variable augmentation has resulted in an elegant Bayesian inference of the model. Empirical evaluation on two real world datasets demonstrates that through proper statistical modeling of data we are able to capture users' rating behavior and outperform state-of-the-art approaches. Furthermore, our model is effective in user modeling, analyzing users' aspect preferences and provides a better product quality estimation even when the product has received few ratings. Most importantly, the construction of the model described here is generic and presents new possibilities for modeling such data in a wide-range of domains.

References

[Albert and Chib, 1993] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 1993.

[Barbieri, 2011] Nicola Barbieri. Regularized gibbs sampling for user profiling with soft constraints. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2011.

[Chen *et al.*, 2013] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, 2013.

[Hofmann, 2003] Thomas Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[Khan *et al.*, 2012] Mohammad Emtiyaz Khan, Shakir Mohamed, Benjamin M Marlin, and Kevin P Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *AISTATS*, 2012.

[Koren and Sill, 2013] Yehuda Koren and Joseph Sill. Collaborative filtering on ordinal user feedback. In *IJCAI*, 2013.

[Koren, 2008] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *ACM SIGKDD*, 2008.

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 2001.

[Marlin, 2003] Benjamin M Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in neural information processing systems*, 2003.

[Muthukumarana and Swartz, 2014] Saman Muthukumarana and Tim B Swartz. Bayesian analysis of ordinal survey data using the dirichlet process to account for respondent personality traits. *Communications in Statistics-Simulation and Computation*, 2014.

[Pennock *et al.*, 2000] David M Pennock, Eric Horvitz, Steve Lawrence, and C Lee Giles. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In *Uncertainty in artificial intelligence*, 2000.

[Pierre, 2012] C Pierre. Balancing prediction and recommendation accuracy: hierarchical latent factors for preference data. 2012.

[Polson *et al.*, 2013] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 2013.

[Rossi *et al.*, 2001] Peter E Rossi, Zvi Gilula, and Greg M Allenby. Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Association*, 2001.

[Salakhutdinov and Mnih, 2007] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.

[Salakhutdinov and Mnih, 2008] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *International conference on Machine learning*, 2008.

[Stern *et al.*, 2009] David H Stern, Ralf Herbrich, and Thore Graepel. Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, 2009.

[Virtanen and Girolami, 2015] Seppo Virtanen and Mark Girolami. Ordinal mixed membership models. In *International Conference on Machine Learning*, 2015.

[Wang and Ester, 2014] Hao Wang and Martin Ester. A sentiment-aligned topic model for product aspect rating prediction. In *EMNLP*, 2014.

[Wang *et al.*, 2009] Pu Wang, Carlotta Domeniconi, and Kathryn Blackmond Laskey. Latent dirichlet bayesian co-clustering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2009.

[Wang *et al.*, 2011] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD*, 2011.