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 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 $r_{ij}$ is an $A$ dimensional vector denoting the rating of user $j$ for item $i$ on each of its aspects. Each $r_{ij}$ is a discrete value between 1 and $K$ corresponding to a $K$-level scale \textit{(poor to excellent)}. We assume that $r_{ij}$ arises from a latent multivariate continuous response $v_{ij}$ 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 $z_i$ is an $A$ dimensional vector, drawn from a multivariate normal distribution, with mean $\mu$ and covariance matrix $\Sigma$. 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 \textit{room} quality but very poor \textit{cleanliness}, but it is possible to have a good \textit{location} and average food choices. Such correlations among subset of aspects are captured by the covariance matrix. The parameters $(\mu, \Sigma)$ are given a conjugate normal-inverse Wishart (NIW) prior.

The preference of a user for an aspect is captured by a bias vector $m_g$ of dimension $A$. If a user places great importance on a particular aspect (e.g. \textit{cleanliness}), this will be reflected in his ratings across all hotels. In other words, his rating on \textit{cleanliness} aspect will tend to be lower than the majority’s rating for \textit{cleanliness} on the same hotel. We cluster users with similar preferences into different user groups and associate a bias vector $m_g$ 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 $\theta$ with Dirichlet prior parameter $\alpha$.

Given the intrinsic quality $z_i$ and bias $m_g$, the latent response $v_{ij}$ is drawn from a multivariate Gaussian distribution with $z_i + m_g$ as mean and a hyper-parameter $B$ 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 $v_{ij}$, we sample the observed rating vector $r_{ij}$. Note that since the observed ratings are ordered and discrete, they should be drawn from a categorical distribution. However, the latent response $v_{ij}$ 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:

1. Draw a multinomial group distribution $\theta$ from Dirichlet $(\alpha)$.
2. For each group $g \in \{1, \cdots, G\}$ draw a bias offset $m_g$ from $\mathcal{N}(0, \Lambda)$.
3. For each user $j \in \{1, \cdots, J\}$, sample a group $s_j$ from Cat $(\theta)$.
4. For each item $i \in \{1, \cdots, I\}$, sample an intrinsic rating $z_i$ from $\mathcal{N}(\mu, \Sigma)$.
5. For each rating $r_{ij} \in R$
   \begin{enumerate}
   \item draw latent continuous rating $v_{ij}$ from $\mathcal{N}(z_i + m_{s_j}, B)$
   \item draw observed ordinal rating $r_{ij}$ from Cat $(\mathcal{S}(v_{ij}, c))$
   \end{enumerate}

where $\mathcal{S}(v_{ij}, c)$ refers to the stick-breaking parameterization of the continuous response $v_{ij}$ using cut-points $c$. 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 $v_{ij}$, denoted as $\text{Lik}(v_{ij})$, 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 $c = \{c_1, \cdots, c_{K-1}\}$ 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 $\eta_{ija} = c_k - v_{ija}$. Then the probability of observing the vector of ratings $r_{ij}$ is a product of probabilities of observing each of the aspect ratings $r_{ija}$ given the values of $\eta_{ija}$. Hence the likelihood of $v_{ij}$ is:

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

To squash $\eta_{ija}$ within $[0,1]$ we use a sigmoid function on it, denoted by $f(x) = \frac{1}{1 + e^{-x}}$. 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:

$$
P(r_{ija} = k) = \prod_{k' < k} (1 - f(\eta_{ija}^{k'})) f(\eta_{ija}^k)
$$
By encoding $r_{ija}$ with a 1-of-$K$ vector $x_{ija}$ where
\[
 x_{ija}^k = \begin{cases} 
 1 & \text{if } r_{ija} = k \\
 0 & \text{otherwise} 
\end{cases}
\]
we now rewrite the likelihood of $v_{ij}$ in binomial terms:
\[
P(r_{ija} | \eta_{ija}) = P(x_{ija}^k | \eta_{ija}) = \prod_{k=1}^{K-1} \text{Binom}(x_{ija}^k | N_{ija}^k, f(\eta_{ij}))
\]
where
\[
N_{ija}^k = 1 - \sum_{k'<k} x_{ija}^{k'}
\]

### 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:
\[
\frac{(e^\varphi)^a}{(1 + e^\varphi)^b} = 2^{-b} e^{\varphi \omega^2 / 2} \int_0^\infty e^{-\varphi \omega^2 / 2} p(\omega) d\omega
\]
where $\kappa = a - b/2$, $b > 0$ and $\omega \sim \mathcal{G}(b, 0)$.

By expanding the binomial likelihood in Eqn. 4, we get
\[
P(x_{ija} | \eta_{ija}) = \prod_{k=1}^{K-1} \left( N_{ija}^{k(0)} + \sum_{k'=1}^{k} (1 - f(\eta_{ij})) \right)^{N_{ija}^k - x_{ija}^k}
\]
\[
= \prod_{k=1}^{K-1} \left( N_{ija}^{k(0)} + \sum_{k'=1}^{k} (1 - f(\eta_{ij})) \right)^{N_{ija}^k - x_{ija}^k}
\]
\[
= \prod_{k=1}^{K-1} \left( N_{ija}^{k(0)} + \sum_{k'=1}^{k} (1 - f(\eta_{ij})) \right)^{N_{ija}^k - x_{ija}^k}
\]

Using the integral identity of PG augmentation, we can now rewrite the categorical likelihood of $v_{ij}$ as:
\[
\text{Lik}(v_{ij}) = \prod_{a=1}^{A} P(x_{ija} | \eta_{ija})
\]
\[
\times \prod_{k=1}^{K-1} \prod_{a=1}^{A} \int_0^\infty e^{-\varphi \omega^2 / 2} p(\omega_{ija}^k) d\omega_{ija}^k
\]
where $\kappa_0 = x_{ija}^k - N_{ija}^k/2$, $\psi_0 = \eta_{ija}^k$ and $p(\omega_{ija}^k)$ is
\[
\mathcal{G}(N_{ija}^k/2, 0)
\]

By property of PG distribution, we can draw the auxiliary variable $\omega_{ija}^k$ from $\mathcal{G}(N_{ija}^k, \eta_{ija}^k)$.

### 2.3 Bayesian Inference

Finally, we describe the sampling of user groups $s$, bias offset of user groups $m$, intrinsic ratings $z$, cut-points $c$ and latent continuous ratings $v$ using fully Bayesian MCMC inference.

We factor the joint probability of these variables as:
\[
P(r, v, m, z, s, c) = P(r|v, c) P(v|m, z, s) P(c) P(z) P(s|P(m))
\]

#### Sampling Bias Offset of User Groups

For each user group $g$, we sample its bias offset $m_g$ from the Gaussian posterior:
\[
P(m_g | A, v, z) \propto P(m_g | A) \prod_{j \in J[g]} P(v_{ij}|m_g, z_i, B)
\]
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 $v_{ij}$ are also drawn from a multivariate Gaussian $N_A(z_i + m_g, B)$, the posterior of $m_g$ is given by a Gaussian $N_A(m_{\tilde{g}}, \tilde{A}_g)$ with
\[
m_{\tilde{g}} = \tilde{A}_g^{-1} \sum_{j \in J[g]} (v_{ij} - z_i)
\]
\[
\tilde{A}_g = (n_g \Sigma^{-1} + \Lambda)^{-1}
\]
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 $\theta$ by exploiting Dirichlet-Multinomial conjugacy, and sample the group of each user $j$ as:
\[
P(s_{ij} | s, m, v) \propto P(s_{ij} | s) \prod_{i \in I[j]} P(v_{ij}|m_{s_{ij}}, z_i, B)
\]
where $I[j]$ are the subset of items rated by user $j$, the prior $P(s_{ij} | s)$ 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 $m_{s_{ij}}$.

#### Sampling Intrinsic Ratings

Similar to the bias offsets of user groups, we sample intrinsic rating $z_i$ of each item $i$ from a Gaussian distribution $N_A(\tilde{p_i}, \tilde{\Sigma}_i)$ where
\[
\tilde{p}_i = \tilde{\Sigma}_i(B^{-1} \sum_{j \in I[i]} (v_{ij} - m_{s_{ij}}) + \Sigma^{-1} \mu)
\]
\[
\tilde{\Sigma}_i = (n_i \Sigma^{-1} + \Sigma)^{-1}
\]
where $n_i$ is the total number of ratings observed for item $i$ and $I[i]$ is the subset of users who have rated item $i$. The prior parameters $\mu, \Sigma$ of the intrinsic ratings are given as a conjugate Normal-Inverse Wishart (NIW) prior and sampled.

#### Sampling Latent Continuous Ratings

The latent continuous ratings, $v_{ij}$ have a Gaussian prior $N_A(z_i + m_g, B)$ and a categorical probability $P(r_{ija} | v_{ij}, c)$. We have transformed the categorical likelihood to the conditional Gaussian form (recall Eqn. 8). The posterior can be formulated as:
\[
P(v_{ij} | m_{s_{ij}}, z_i, B) \propto \text{Lik}(v_{ij} | v, r_{ij}, c)
\]
\[
\propto \exp \left\{ -\frac{1}{2} \left( v_{ij} - (z_i + m_{s_{ij}}) \right)^T \Sigma_{v}^{-1} \left( v_{ij} - (z_i + m_{s_{ij}}) \right) \right\}
\]
\[
\times \prod_{k=1}^{K-1} \exp \left\{ -\frac{1}{2} \left( \kappa_0^k \omega_{ij}^k - (c_k - v_{ij}) \right)^T \Omega_{ij}^k \right\}
\]
where $\kappa_0^k, \omega_{ij}^k$ are vectors of dimension $A$, $\Omega_{ij}^k$ is a diagonal matrix of $\omega_{ij}^{k_1}, \omega_{ij}^{k_2}, \ldots, \omega_{ij}^{k_d}$. Here, we assume the values in the A-dimensional cut-point vector $c_k$ are all equal to $c_k$. In practice, if we need different cut-points for different aspects, $c_k$ can be set accordingly.
Since both the prior and likelihood are now Gaussian, we have the following Gibbs sampler:
\[
\begin{align*}
\mathbf{v}_{ij} & \sim N_d(\mu_{ijw}, \Sigma_{ijw}) \\
\omega_{ija} & \sim PG(N_{ija}, \mathbf{v}_{ija} - \mathbf{c})
\end{align*}
\]
where
\[
\begin{align*}
\mu_{ijw} &= \mathbf{B}^{-1}(\mathbf{z}_i + \mathbf{m}_{ij}) + \sum_{k=1}^{K-1} \omega_{ij}^k (\mathbf{c}_k - \mathbf{K}_{ij}^k) \\
\Sigma_{ijw} &= \mathbf{B}^{-1} + \sum_{k=1}^{K-1} \Omega_{ij}^k
\end{align*}
\]

**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)](image)

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

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

Taking the ratio, we have

\[
P(r_{ija} = k + 1) = \frac{1 - f(\delta_k))}{P(r_{ija} = k)} \frac{(f(\mathbf{c}_{k+1} - \mathbf{c}_k - \mathbf{\delta}_k))}{f(\mathbf{c}_k - \mathbf{c}_k - \mathbf{\delta}_k)}
\]
\[
= \frac{1 + e^{\mathbf{c}_k}}{1 + e^{\mathbf{c}_k + \mathbf{\delta}_k + 1}}
\]
\[
\delta_k \geq -\ln (1 - e^{-(c_{k+1} - c_k)}), \text{ see that } \frac{1 + e^{\mathbf{c}_k}}{1 + e^{\mathbf{c}_k + \mathbf{\delta}_k + 1}} > 1. \text{ Hence, } P(r_{ija} = k + 1) > P(r_{ija} = k).
\]

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 \( v_{ija} \geq (c_{k-1} - \ln (1 - e^{-(c_{k-1} - c_k)})) \). This implies that, when \( v_{ija} \) is within the range \( (c_k - \ln (1 - e^{-(c_{k+1} - c_k)}), c_{k-1} - \ln (1 - e^{-(c_{k-1} - c_k)})) \), 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 \( \arg\max_{k'} P(r_{ija} | v_{ija}, k') = k \).

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:

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

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Items</th>
<th># Users</th>
<th># Ratings</th>
<th>Aspects rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>TripAdvisor</td>
<td>12,773</td>
<td>781,403</td>
<td>1,621,956</td>
<td>Service, Value, Room, Location</td>
</tr>
<tr>
<td>OpenTable</td>
<td>2805</td>
<td>1991</td>
<td>75,469</td>
<td>Ambience, Food, Service, Value</td>
</tr>
</tbody>
</table>

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 Comparison](image)

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], BPFM [Salakhutdinov and Mnih, 2008], URP [Marlin, 2003; Barbieri, 2011], SVD++ [Koren, 2008] and BHFRe [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
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.

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.

### 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,
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 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 $\Delta obs$, difference between their average ratings be $\Delta avg$ and difference between the learned intrinsic ratings be $\Delta int$. Figure 6 shows the correlation between $\Delta obs$ and $\Delta int$, as well as the correlation between $\Delta obs$ and $\Delta avg$ aggregated over all aspects. We observe that for both datasets, as $\Delta int$ increases, $\Delta obs$ also increases. However, $\Delta avg$ remains almost constant. This indicates that $\Delta obs$ is closely correlated with $\Delta int$, whereas $\Delta avg$ appears to be independent of $\Delta obs$. 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.

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], Polya-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 Polya-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 Polya-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


