Development and Derivation of the Psychological Model

August 31, 2018

This document provides development logs of the psychological model, including intuitions, derivations, and result summaries.

To begin with, we must outline several key concepts in building the blocks of our models.

- K aspects. We assume there are K aspects based on which a user measures the items. The number K is predefined. Consequently. we define the variables, U, V where $v_{i,k}$ denotes item i's feature on aspect k and $u_{p,k}$ denotes user p's preference on aspect k. The range of U, V is specified by the model, thus a probabilistic generative model and a numerical model might provide apply different range constraints.

- Prominent aspect(s). We assume that the K aspects are not treated equally by the users. One or some (depending on the model) of the aspects are most important.

- Distinguish prominent and non-prominent aspects. We adopt different scoring or ranking methods for prominent and non-prominent aspects. This part essentially makes our model non-compensate.

Part I

Ranking Based Models

This part devotes to models that observe rankings. In each session, we assume that the user is given a set of items, and the user generates a set of pairwise rankings. For example, if a user has access to items i, j, and he buys i, clicks on j. Then we can make a reasonable conclusion that to the user $i \succ j$. The ranking based models mimic the generation of pairwise rankings. The details of observing rankings, i.e. which actions to be taken into accounts and how to grade these actions, are out of the scope of this document (but should be considered in the experiments).

0.1 The BTL Model

0.1.1 Model

In the extended BTL model, a ranking $i \succ j$ is likely to happen if the latent utility score i is relevantly larger than j, $p(i \succ j) = \frac{i}{i+\theta j}$. The parameter $\theta > 1$ plays as a tolerance threshold. If the absolute difference between i and j is not significant $|i - j| \leq \theta$, the user will consider it to be a tie $i = j$, $p(i = j) = \frac{(\theta^2-1)ij}{[i+\theta j][\theta i+j]}$. Note that $p(i \succ j) + p(i \prec j) + p(i = j) = 1$.

Based on the extended BTL model, we propose our first attempt of the psychological model. The highlights are (1) we consider only one prominent aspect in each session within the K aspects. (2) The winning item must be significantly better than the losing item on the prominent aspect. (3) The winning item can be not as good as the losing product as long as $\theta v_{i,k} > v_{j,k}$. The usage of parameter θ can be interpreted as the consumer sets a minimally acceptable level of performance. The cutoff point is controlled by both θ and how well other items perform.

Therefore the probability of w winning in the session d is defined in Equ. 1 as the product of the probability that w outranks other items l on the most pertinent aspect $p(w_k \succ l_k)$ and the probability that w at least ties with other items on other aspects $p(w_{k'} \succeq l_{k'})$.

$$p(< w, l > |\Theta, g) = \Pi_{k=1}^{K} [\frac{w_k}{w_k + \theta l_k}^{g_k} \frac{\theta w_k}{l_k + \theta w_k}^{1-g_k}] \tag{1}$$

Thus we present the likelihood of purchase model under Non-Compensatory Rules (NCR) as follows. The model parameters are denoted as $\Theta = \{\theta, v \in V, u \in U)\}$.

$$p(D|\Theta) = \Pi_{d \in D} \Sigma_g \{\Pi_{w \in W^d, v \in L^d} p(< w, l > |g, \Theta) p(g|u)\} \tag{2}$$

The inference is implemented by EM. As in [**?**], in the M-step of EM, instead of maximizing $Q(\Theta) = E_g \ln p(D, G|\Theta)$, we compute a minorization

function of Q . For the limited space, we skip the derivation here and present the updates of NCR in Equ. 3.

$$
\begin{aligned}
\gamma(d,k,\Theta^t) =& \frac{u_k \Pi_{w\in W^d, v\in L^d} \frac{w_k}{w_k+\theta^t v_k} \Pi_{k'\neq k} \frac{\theta^t w_{k'}}{v_{k'}+\theta^t w_{k'}}}{\Sigma_{k=1}^K u_k \Pi_{w\in W^d, v\in L^d} \frac{w_k}{w_k+\theta^t v_k} \Pi_{k'\neq k} \frac{\theta^t w_{k'}}{v_{k'}+\theta^t w_{k'}}} \qquad (3)\\
\alpha(v,v',k,\Theta^t) =& v_k^t + \theta^t {v'}_k^t \\
u_k =& \frac{\Sigma_{u(d)=u}\gamma(d,k,\Theta^t)}{\Sigma_{s=1}^K \Sigma_{u(d)=u}\gamma(d,s,\Theta^t)} \\
\frac{1}{v_k} =& \frac{\Sigma_{d\in W(v)}\Sigma_{v'\in L_d}[\frac{\gamma(d,k,\Theta^t)}{\alpha(v,v',k,\Theta^t)} + \Sigma_{k'\neq k}\frac{\theta^t\gamma(d,k',\Theta^t)}{\alpha(v',v,k,\Theta^t)}]}{\Sigma_{d\in W(v)}|L_d|} \\
&+ \frac{\Sigma_{d\in L(v)}\Sigma_{v'\in W_d}[\frac{\theta^t\gamma(d,k,\Theta^t)}{\alpha(v',v,k,\Theta^t)} + \Sigma_{k'\neq k}\frac{\gamma(d,k',\Theta^t)}{\alpha(v,v',k,\Theta^t)}]}{\Sigma_{d\in W(v)}|L_d|} \\
\theta =& \frac{(K-1)\Sigma_d|W_d||L_d|}{\Sigma_d\Sigma_k\gamma(d,k,\Theta^t)\Sigma_{w,v}[\frac{v_k}{\alpha(w,v,k,\Theta^t)} + \Sigma_{k'\neq k}\frac{w_{k'}}{\alpha(v,w,k',\Theta^t)}]}
\end{aligned}
$$

0.1.2 Results

Unfortunately, the BTL model does NOT outperform comparative models, e.g. BPR on our Tmall dataset. But it outranks BPR and other models (excluding classification and regression models) on the Yochoose dataset.

Nonetheless, we run experiments based on BTL for a series of evaluations. The first one is about different actions. Our conclusions include: (1) Ranking buy v.s. add to cart is slightly better than BPR. However, ranking solely on other types of actions suffer from the lack of data. (2) Ranking based on ensemble action pairs (3)

0.2 BTR

We want to make the smallest modification to the BPR model. We use the Bayesian maximum posterior estimator

$$BPR - Opt = \ln p(i \succ_d j|\Theta)p(\Theta) \qquad (4)$$

where $\ln p(i \succ_d j|\Theta) = \sigma(x) = \frac{1}{1+\exp -x_{u,i,j}}$, Θ is the set of all parameters. In the first type of BPR (with matrix factorization), $x_{u,i,j} = u^T(i-j)$, $\Theta = U, V$ is the set of all user and item latent vectors. Here, we introduce

a new parameter θ and define:

$$x_{u,i,j} = \sum_k \frac{\exp(u_k)}{\sum_{k'} \exp(u_{k'})} [\exp\theta(i_k - j_k) + \sum_{k'} (i_{k'} - j_{k'})] \tag{5}$$

As in [**?**], the gradient of BPR-Opt with respect to the model parameter is

$$\frac{\partial BPR - Opt}{\partial \theta} \propto \sum_{i \succ_d j} \frac{-e^{-x_{u,i,j}}}{1 + e^{-x_{u,i,j}}} \frac{\partial x_{u,i,j}}{\partial \Theta} - \lambda_\Theta \Theta \tag{6}$$

where

$$\frac{\partial x_{u,i,j}}{\partial \Theta} = \begin{cases} [\exp\theta(i_k - j_k) + \sum_{k' \neq k}(i_{k'} - j_{k'})][\frac{\exp u_k}{\sum_k \exp u_k} - \frac{\exp^2 u_k}{(\sum_k \exp u_k)^2}] & if \Theta = u_k \\ \frac{\exp u_k \exp\theta + \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = i_k \\ \frac{-\exp u_k \exp\theta - \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = i_k \\ \sum_k \frac{\exp(u_k)(i_k - j_k)}{\sum_{k'} \exp(u_{k'})} \exp\theta & if \Theta = \theta \end{cases} \tag{7}$$

0.3 AMF

Neighborhood information can also be incorporated in the definition of $x_{u,i,j}$. As in [**?**], in this paper, we adopt the asymmetric matrix factorization (AMF) [**?**], each item i is associated with two factor vectors q_i and p_i. The representation of user u is through the sum $\sum_{j \in R(u)} p_j / \sqrt{|R(u)|}$, where $R(u)$ is the set of positive items.

$$\hat{r_{ui}} = u + b_u + b_i + q_i(\sum_{s \in R(u)} p_s / \sqrt{|R(u)|}) \tag{8}$$

FSBPR [**?**] incorporates AMF in a pairwise BPR style ranking framework. First, the difference of prediction is defined as :

$$x_{u,i,j} = \hat{r_{ui}} - \hat{r_{uj}} = \sum_k [\frac{\sum_{s \in R(u)} p_s}{\sqrt{|R(u)|}}]_k (q_{i,k} - q_{j,k}) \tag{9}$$

According to the chain rule $\frac{\partial BPR - Opt}{\partial \theta} \propto \sum_{i \succ_d j} \frac{-e^{-x_{u,i,j}}}{1+e^{-x_{u,i,j}}} \frac{\partial x_{u,i,j}}{\partial \Theta} - \lambda_\Theta \Theta$, gradient descent is applied to update the parameters $\Theta \leftarrow \Theta - \alpha(\frac{-e^{-x_{u,i,j}}}{1+e^{-x_{u,i,j}}} \frac{\partial x_{u,i,j}}{\partial \Theta} - \lambda_\Theta \Theta)$, where α is the learning rate, and $\frac{\partial x_{u,i,j}}{\partial \Theta}$ is:

$$\frac{\partial x_{u,i,j}}{\partial \Theta} = \begin{cases} \sum_{u|s\in R(u)} \frac{q_{i_k} - q_{j_k}}{\sqrt{|R(u)|}} & if \Theta = p_{s_k} \\ [\frac{\sum_{s\in R(u)} p_s}{\sqrt{|R(u)|}}]_k & if \Theta = q_{i_k} \\ -[\frac{\sum_{s\in R(u)} p_s}{\sqrt{|R(u)|}}]_k & if \Theta = q_{j_k} \end{cases} \tag{10}$$

To build the non-compensatory version of AMF, we set $u_k = [\frac{\sum_{s\in R(u)} p_s}{\sqrt{|R(u)|}}]_k$, and thus we have:

$$\frac{\partial x_{u,i,j}}{\partial \Theta} = \begin{cases} \sum_{u|s\in R(u)} [\exp\theta(i_k - j_k) + \sum_{k'\neq k}(i_{k'} - j_{k'})][\frac{\exp u_k}{\sum_k \exp u_k} - \frac{\exp^2 u_k}{(\sum_k \exp u_k)^2}] \frac{p_{s_k}}{\sqrt{|R(u)|}} & if \Theta = p_{s_k} \\ \frac{\exp u_k \exp\theta + \sum_{k'\neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = q_{i_k} \\ \frac{-\exp u_k \exp\theta - \sum_{k'\neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = q_{j_k} \\ \sum_k \frac{\exp(u_k)(i_k - j_k)}{\sum_{k'} \exp(u_{k'})} \exp\theta & if \Theta = \theta \end{cases} \tag{11}$$

0.4 LCR

Local **C**ollaborative **R**anking (LCR) [**?**] minimizes a pair-wise loss function, which is constructed using observed entries. In the experiments [**?**], the log-loss function is shown to achieve best results, thus it is adopted in this paper.

$$L = \sum_{u\in U} \frac{1}{s_u} \sum_{k=1}^{s_u} \Delta M_{u,i,j} \log(1 + \exp - g(u,i,j)), \tag{12}$$

where s_u is the number of ordered items rated by the user u. The loss function approximates the observed rating difference with predicted rating difference $g(u,i,j)$. LCR is inspired by LLORMA, which assumes that the rating matrix itself is not low-rank. Instead, M is locally low-rank where locality is defined by a neighborhood with respect to the given kernel K. Therefore, the prediction difference can be approximated by a set of low-rank factorizations around q anchor points i_s: $g(u,i,j) = \sum_{t=1}^{q} \sum_k u_{t,k} [\frac{K((u_t,i_t),(u,i))}{\sum_{s=1} qK((u_s,i_s),(u,i))} i_{t,k} - \frac{K((u_t,i_t),(u,j))}{\sum_{s=1}^{q} K((u_s,i_s),(u,j))} j_{t,k}]$. [1]

Again, we can modify the score difference $g(u,i,j) = \sum_{t=1}^{q} \sum_k \frac{\exp(u_{t,k})}{\sum_{k'} \exp(u_{t,k'})} [\exp\theta(s(u,t,i) i_{t,k} - s(u,t,j) j_{t,k}) + \sum_{k'} (s(t,u,i) i_{t,k'} - s(t,u,j) j_{t,k'})]$, where $s(t,u,i) = \frac{K((u_t,i_t),(u,i))}{\sum_{s=1}^{q} K((u_s,i_s),(u,i))}$, $s(t,u,j) =$

[1] This equation is not the same as Equ.15 in the paper

$\frac{K((u_t,j_t),(u,j))}{\sum_{s=1}^{q} K((u_s,i_s),(u,j))}$. Applying the chain rule, the gradient is $\frac{\partial L}{\partial \Theta} = \frac{\partial L}{\partial g(u,i,j)} \frac{\partial g(u,i,j)}{\theta}$, which is defined by Equ. 23. Thus, the update steps will be:

$$
\begin{aligned}
line12: \quad & s_{t,u,i} && \leftarrow \frac{K(u_t,u)K(i_t,i)}{w_{u,i}} \\
line15: \quad & g_{u,i,j} && = \sum_{t=1}^{q} \sum_{k} \frac{\exp(u_{t,k})}{\sum_{k'} \exp(u_{t,k'})} [\exp\theta(s(t,u,i)i_{t,k} - s(t,u,j)j_{t,k}) + \sum_{k'} (s(t,u,i)i_{t,k'} - s(t,u, \\
line15: \quad & l_{u,i,j} && = \frac{-\Delta M \exp -g_{u,i,j}}{1+\exp -g_{u,i,j}} \\
line23: \quad & [\Delta u]_k && \leftarrow [\Delta u]_k + [\exp\theta(s(t,u,i)i_{t,k} - s(t,u,j)j_{t,k}) + \sum_{k' \neq k} (s(t,u,i)i_{t,k'} - s(t,u,j)j_{t,k'})] \\
& && \quad [\frac{\exp u_{t,k}}{\sum_k \exp u_{t,k}} - \frac{\exp^2 u_{t,k}}{(\sum_k \exp u_{t,k})^2}] l_{u,i,j} \\
line25: \quad & [\Delta i]_k && \leftarrow [\Delta i]_k + \frac{\exp u_k \exp\theta + \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} s(t,u,i) l_{u,i,j} \\
line26: \quad & [\Delta i]_k && \leftarrow [\Delta i]_k + \frac{-\exp u_k \exp\theta - \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} s(t,u,i) l_{u,i,j} \\
line26+: \quad & [\Delta\theta] && \leftarrow \sum_k \frac{[s(u,t,i)i_{t,k} - s(u,t,j)j_{t,k}] \exp(u_k)}{\sum_{k'} \exp(u_{k'})} \exp\theta \\
line30+: \quad & \theta && = \theta - v[\Delta\theta]
\end{aligned}
$$

0.5 The Sigmoidal Model

As the section title indicates, we propose a set of models that relate pairwise rankings $i \succ j$ to logit functions. The optimization objective is in a general form of :

$$
\begin{aligned}
\min_{u,v} \quad & \textstyle\sum_{i \succ j} \frac{1}{2} (1 - \sum_k u_k \sigma(v_{i,k} - v_{j,k}))^2 - \frac{\lambda}{2} \sum_u \|u\|^2 \\
s.t. \quad & \forall u \forall k, u_k > 0 \\
& \forall u, \textstyle\sum_k u_k = 1
\end{aligned}
\tag{14}
$$

There are a few issues needed to be emphasized here. (1) Essentially this is a mixture model so only evaluation on prominent aspects is provided. We

could make explicit treatments for prominent and non-prominent aspects in the model by changing the constraints to $\forall u, \forall k, 0 < u_k < 1$. (2) With the negative regularization term (together with the constraints) we penalize sparse us, thus we allow more prominent aspects to be chosen. (3) We use the sum-of-square error function, because our prediction for a pair $< i, , j >$ $f(ij) = \sum_k u_k \sigma(v_{i,k} - v_{j,k})$ falls in the range of $(0, 1)$ and $f(i, j) = 1 - f(j, i)$. So we label $i \succ j$ to be 1. For symmetry, we don't have to label $j \succ i$. (4) We do not normalize v as $\sigma(v_{i,k}, v_{j,k})$ is already normalized.

The optimization algorithm is listed below. In each iteration, we first optimize over u when vs are fixed. We rewrite the objective as $\|Ax - b\|^2 - \lambda\|x\|^2$ with constraints $C_i^T x = d_i, \forall i = 1, \cdots, U$, where $x \in \mathcal{R}^{KU}$, U is the number of users, $x_i = u_{i,k}$, $A \in \mathcal{R}^D$, D is the number of observations and $D \simeq SL$, where S is the number of sessions, and L is the average number of pairs in a session, $C_i \in \mathcal{R}^{KU}$ where for $1 \leq i \leq U, iK + 1 \leq j \leq i(K + 1), C_{i,j} = 1, \forall j', C_{i,j'} = 0$, b is a vector of all ones, $d_i = 1$ is a scalar. Therefore we have

$$\begin{bmatrix} x^* \\ z^* \end{bmatrix} = \begin{bmatrix} 2A^T A - \lambda & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^T b \\ d \end{bmatrix} \tag{15}$$

Note that we have $S \geq U$ to filter cold start users, and generally $L \geq K - 1$, $K > 1$, therefore the matrix is invertible.

Next we optimize v while keeping u fixed. We adopt stochastic gradient. As shown below, the gradient is the addiction of two terms:

$$\begin{aligned} \frac{\delta L}{\delta v_{i,k}} &= \sum_{i \succ j} (1 - \sum_{k'} u_{k'} \sigma(v_{i,k'}, v_{j,k'})) \sigma(v_{i,k}, v_{j,k})(1 - \sigma(v_{i,k}, v_{j,k})) \\ &\quad + \sum_{w \succ i} (1 - \sum_{k'} u'_{k'} \sigma(v_{w,k'}, v_{i,k'})) \sigma(v_{w,k}, v_{i,k})(\sigma(v_{w,k}, v_{i,k}) - 1) \end{aligned}$$

We can sample an item j from $i \succ j$ and another item w from $w \succ i$, compute $S_1 = (1 - \sum_{k'} u_{k'} \sigma(v_{i,k'}, v_{j,k'})) \sigma(v_{i,k}, v_{j,k})(1 - \sigma(v_{i,k}, v_{j,k}))$ and $(1 - \sum_{k'} u'_{k'} \sigma(v_{w,k'}, v_{i,k'})) \sigma(v_{w,k}, v_{i,k})(\sigma(v_{w,k}, v_{i,k}) - 1)$, then we update

$$v_{i,k} = v_{i,k} + \eta |i \succ j| S_1 + \eta |w \succ i| S_2, \tag{16}$$

where η is the step length, $|\cdot|$ is the size of a set.

Part II

Rating Based Models

In this section, we develop a series of ranking based models. These models have the following properties. (1) The models are based on rating observations, i.e. we assign a rating to every action. The actions include the old-fashioned preference ratings (in such case, each rating is a standalone session), actions in a session, positive and negative samples, etc. (2) The models are Bayesian generative models that have good explainability. (3) We control the assumptions by adjusting the hyper-parameters in these models.

0.6 Matrix Factorization

In conventional matrix factorization [**?**], the predicted rating can be computed as an inner product of user preferences and item features as follows.

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \mathbf{q}_k \mathbf{u}_k \tag{17}$$

For a regularized loss function $L = \sum_{(u,q)\in D}(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})^2 + \lambda(\sum_q \|\mathbf{q}\| + \sum_u \|\mathbf{u}\|)$ The update process

$$\begin{aligned}
\mathbf{u} &= \mathbf{u} - \alpha\frac{\partial L}{\partial \mathbf{u}} = \mathbf{u} + \alpha \sum_{q\in R(u)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{u}}] - \alpha\lambda\mathbf{u} \\
\mathbf{q} &= \mathbf{q} - \alpha\frac{\partial L}{\partial \mathbf{q}} = \mathbf{q} + \alpha \sum_{u\in R(q)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{q}}] - \alpha\lambda\mathbf{q}
\end{aligned} \tag{18}$$

where we can change the prediction to

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}} [\exp\theta \mathbf{q}_k + \sum_{k'\neq k} \mathbf{q}_{k'}]. \tag{19}$$

Thus in updating we use

$$\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \Theta} = \begin{cases} [(q_k)\exp\theta + \sum_{k'\neq k}(q_{k'})][\frac{\exp u_k}{\sum_k \exp u_k} - \frac{\exp^2 u_k}{(\sum_k \exp u_k)^2}] & if \Theta = u_k \\ \frac{\exp u_k \exp\theta + \sum_{k'\neq k}\exp u_{k'}}{\sum_k \exp u_k} & if \Theta = q_k \\ \sum_k \frac{q_k \exp(u_k)}{\sum_{k'}\exp(u_{k'})}\exp\theta & if \Theta = \theta \end{cases} \tag{20}$$

We further consider a model where the user set a cut-off point $b_{u,k}$ for every non-prominent aspect. Thus we can change the prediction to

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}} [\exp \theta \mathbf{q}_k + \sum_{k' \neq k} [\mathbf{q}_{k'} - b_{u,k'}]_+]. \tag{21}$$

The interpretation is that (1) the evaluation of u on q is mainly based on the prominent aspect k (2) the user u sets a cut-off point for non prominent aspects k'. If the item performance is lower than the threshold $b_{u,k'}$, it does not contribute to the overall evaluation as $[\mathbf{q}_{k'} - b_{u,k'}]_+ = 0$ if $\mathbf{q}_{k'} - b_{u,k'} \leq 0$

Thus

$$\mathbf{u} = \mathbf{u} - \alpha \frac{\partial L}{\partial \mathbf{u}} = \mathbf{u} + \alpha \sum_{q \in R(u)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q}) \frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{u}}] - \alpha \lambda \mathbf{u} \tag{22}$$

$$\mathbf{q} = \mathbf{q} - \alpha \frac{\partial L}{\partial \mathbf{q}} = \mathbf{q} + \alpha \sum_{u \in R(q)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q}) \frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{q}}] - \alpha \lambda \mathbf{q}$$

$$\mathbf{b}_u = \mathbf{b}_u - \alpha \frac{\partial L}{\partial \mathbf{b}_u} = \mathbf{b}_u + \alpha \sum_{q \in R(u)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q}) \frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{b}_u}]$$

$$\theta = \theta - \alpha \frac{\partial L}{\partial \theta} = \theta + \alpha \sum_{u,q} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q}) \frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \theta}]$$

In updating we use

$$\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \Theta} = \begin{cases} [(q_k) \exp \theta + \sum_{k' \neq k} ([\mathbf{q}_{k'} - b_{u,k'}]_+)][\frac{\exp u_k}{\sum_k \exp u_k} - \frac{\exp^2 u_k}{(\sum_k \exp u_k)^2}] & if \Theta = u_k \\ \frac{\exp u_k \exp \theta + \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = q_k \& q_k > b_{u,k} \\ \frac{\exp u_k \exp \theta}{\sum_k \exp u_k} & if \Theta = q_k \& q_k \leq b_{u,k} \\ -(1 - \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}}) & if \Theta = b_{u,k} \& q_k > b_{u,k} \\ 0 & if \Theta = b_{u,k} \& q_k \leq b_{u,k} \\ \sum_k \frac{q_k \exp(u_k)}{\sum_{k'} \exp(u_{k'})} \exp \theta & if \Theta = \theta \end{cases} \tag{23}$$

If we use a "softer" cut-off constraint for non-prominent aspect. We have the prediction according to

$$\hat{\mathbf{X}}_{u,q} = \sum_{k=1}^{K} \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}} [\exp \theta \mathbf{q}_k + \sum_{k' \neq k} [\mathbf{q}_{k'} - b_{u,k'}]]. \tag{24}$$

The interpretation is that (1) the evaluation of u on q is mainly based on the prominent aspect k (2) the user u sets a cut-off point for non prominent

aspects k'. The item performance on non-prominent aspect contributes a term of $[\mathbf{q}_{k'} - b_{u,k'}]$ to the overall evaluation.

Thus

$$\mathbf{u} = \mathbf{u} - \alpha\frac{\partial L}{\partial \mathbf{u}} = \mathbf{u} + \alpha \sum_{q \in R(u)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{u}}] - \alpha\lambda\mathbf{u} \tag{25}$$

$$\mathbf{q} = \mathbf{q} - \alpha\frac{\partial L}{\partial \mathbf{q}} = \mathbf{q} + \alpha \sum_{u \in R(q)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{q}}] - \alpha\lambda\mathbf{q}$$

$$\mathbf{b}_u = \mathbf{b}_u - \alpha\frac{\partial L}{\partial \mathbf{b}_u} = \mathbf{b}_u + \alpha \sum_{q \in R(u)} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \mathbf{b}_u}]$$

$$\theta = \theta - \alpha\frac{\partial L}{\partial \theta} = \theta + \alpha \sum_{u,q} [(\mathbf{X}_{u,q} - \hat{\mathbf{X}}_{u,q})\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \theta}]$$

In updating we use

$$\frac{\partial \hat{\mathbf{X}}_{u,q}}{\partial \Theta} = \begin{cases} [(q_k)\exp\theta + \sum_{k' \neq k}([\mathbf{q}_{k'} - b_{u,k'}]_+)][\frac{\exp u_k}{\sum_k \exp u_k} - \frac{\exp^2 u_k}{(\sum_k \exp u_k)^2}] & if \Theta = u_k \\ \frac{\exp u_k \exp\theta + \sum_{k' \neq k} \exp u_{k'}}{\sum_k \exp u_k} & if \Theta = q_k \\ -(1 - \frac{\exp \mathbf{u}_k}{\sum_{k'} \exp \mathbf{u}_{k'}}) & if \Theta = b_{u,k} \\ \sum_k \frac{q_k \exp(u_k)}{\sum_{k'} \exp(u_{k'})} \exp\theta & if \Theta = \theta \end{cases} \tag{26}$$

0.7 The Beta model

We are motivated to design this model by the following assumptions: (1) we allow multiple ($0\ K$) aspects in a session to be prominent. (2) We use positive and negative observations, e.g. a click is negative and a buy is positive. (3) For prominent aspects, we directly model the likelihood by the Beta function. (4) For non-prominent aspects, we assume that they do not contribute to the likelihood of being positive or negative, ie. $p(o = 1|g_k = 0) = 0.5$. (5) We assume that the final label is assigned by a very strict measurement, i.e. the item must be "good" on all aspects.

As such, in experiments, we should monitor model performances on the following issues. (1) A proper baseline would be to also treat the recommendation system as binary classification problems. For simplicity, I would recommend starting with those traditional MF or classifiers without any number, or time related features. (2) The binary classification framework is

easily extend to one class classification scheme for implicit feedback. However, I would recommend to first testify the performance for two classes. When we involve implicit feedback, a fair comparison is important. It would be best if we could verify the capacity of our model on two-class problems, before we use confidence levels or sampling techniques for the negative implicit feedback. (3) Shall we induce sparse constraints on the number of prominent aspects? This could be done by choosing proper values for a, b. (4) Tuning the parameter K is the last step in experimental studies.

For future extensions, we should keep in mind that this model might suffer from the following weaknesses. (1) The likelihood is obtained by a multiplication of all aspects. Although this is proved by preceding pilot studies, that a complete BTL models multiplying all aspects is better than a simplified BTL model on just one aspect, I am not very confident about the conclusion. We can easily modify the model to account for an opposite assumption, that the "good" performance on one aspect can override other aspects. (2) The non-prominent aspects are discarded. This could be fixed by introducing a step function, i.e. $p(o_v = 1|g_k = 0) = f(v-\theta)$. (3) Shall we consider ratings (normalized to $(0, 1)$), instead of positive and negative samples? A fair comparative study with MF is needed to answer this question. If the answer is, we can modify the generation of observations. All the above directions are possible and can be implemented without much difficulty.

First we generate the user preferences and item features.

- For aspect $k = 1 : K$, for user $u = 1 : M$, sample user preference on aspect $k : u_k \sim Beta(a, b)$. Unlike most of previous research, the user preference vector u in our work is not the normalized weight over all aspects.
- For items $v = 1 : N$, sample item features for the item universe $v_k \sim Beta(\alpha, \beta)$. Again, unlike most of previous research, the user preference vector u in our work is not the normalized weight over all aspects.

Next, given S sessions of user u

- For each session s, for aspect $k = 1 : K$, generate aspect indicator $g_{s,k} \sim Bern(u_k)$. If $g_{s,k} = 1$, then k is a prominent aspect for user u in the session s.
- For each item v within the session s of length L
  - For each aspect k, generate the indicator $p(h_{s,v,k}|g_{s,k}, v_k) = [v_k^{h_{s,v,k}}(1-v_k)^{1-h_{s,v,k}}]^{g_{s,k}}[\frac{1}{2}^{h_{s,v,k}}\frac{1}{2}^{1-h_{s,v,k}}]^{1-g_{s,k}}$

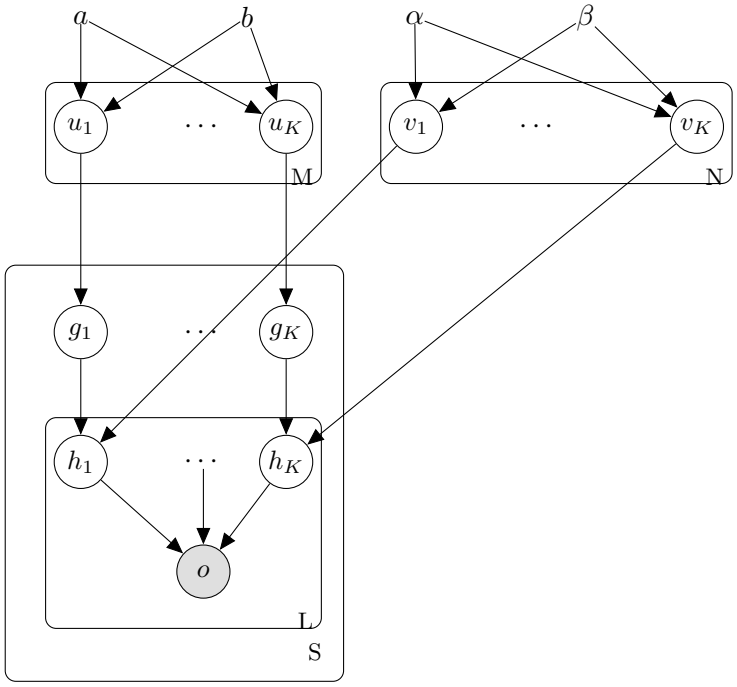

Figure 1: Plate notation of the proposed sigmoidal model

– For each observed rating o_v, generate label observation $o_{s,v} = 1 - \prod_k (1 - h_{s,v,k})$ where $o_v = 1$ if and only if $\forall k, h_{s,v,k} = 1$

The joint probability is given by

$$p(U, V, G, H, D|a, b, \Lambda) = \prod_k \prod_u p(u_k|a, b) \prod_s p(g_{s,k}|u_k) \prod_v \{p(v_k|0, \alpha, \beta) p(o_{s,v}|H) p(h_{s,v,k}|g_{s,k}, v_k)\}$$

We should notice that $o_{v,k} = 1$ indicates $\forall k, h_{v,k} = 1$. Thus the likelihood must be decomposed to a term with observations and a term with hidden variables. We have:

$$\begin{aligned} \ln p(U, V, G, H, D|a, b, \Lambda, \xi) &= \sum_s \sum_{o_{s,v}=1} \sum_k p(h_{s,v,k} = 1|g_{s,k}, v_k) \\ &+ \sum_s \sum_{o_{s,v}=0} p(o_{s,v} = 0|H) \sum_k p(h_{s,v,k}|g_{s,k}, v_k) \\ &+ \sum_u \sum_k p(u_k|a, b) + \sum_v \sum_k p(v_k|\alpha, \beta) + \sum_u \sum_s \sum_k p(g_{s,k}|u_k) \end{aligned} \tag{27}$$

By variational inference, we have factorize the posterior probability $p(U, V, G, H|D, a, b, \alpha, \beta) = q(U)q(V)q(G)q(H)$. We can see from the following derivations that $q(u_k) \sim Bern(a', b')$

$$\begin{aligned} \ln q(u_k) &= \mathbb{E}_{G,V,H}[\ln p(u_k|a, b) + \sum_s \ln p(g_{s,k}|u_k)] + const \\ &= (a - 1) \ln u_k + (b - 1) \ln(1 - u_k) + \sum_s \mathbb{E}[g_{s,k}] \ln u_k + \sum_s (1 - \mathbb{E}[g_{s,k}] \ln(1 - u_k) + const \\ \mathbb{E}[u_k] &= \frac{a + \sum_s \mathbb{E}[g_{s,k}]}{a + b + |S|} \end{aligned}$$

Because $p(h_{s,v,k}|g_{s,k}, v_k) = [v_k^{h_{s,v,k}}(1-v_k)^{1-h_{s,v,k}}]^{g_{s,k}}[\frac{1}{2}^{h_{s,v,k}} \frac{1}{2}^{1-h_{s,v,k}}]^{1-g_{s,k}}$, we have $\ln p(h_{s,v,k}|g_{s,k}, v_k) = g_{s,k}[h_{s,v,k} \ln v_k + (1 - h_{s,v,k} \ln(1 - v_k)] + (1 - g_{s,k} \ln \frac{1}{2}$

$$
\begin{aligned}
\ln q(v) &= \mathbb{E}_{G,U,H}[\ln p(H|v,G) + \ln(v|\alpha,\beta)] + const \\
&= \sum_s \sum_{o_{s,v}=1} \sum_k p(h_{s,v,k}=1|g_{s,k},v_k) + \sum_s \sum_{o_{s,v}=0} \sum_k p(h_{s,v,k}|g_{s,k},v_k) + \sum_v \sum_k p(v_k|\alpha,\beta) \\
\ln q(v_k) &= \{\sum_s \sum_{o_{s,v}=1} \mathbb{E}[g_{s,k}] + \sum_s \sum_{o_{s,v}=0} \mathbb{E}[g_{s,k}]\mathbb{E}[h_{s,v,k}]\} \ln v_k \\
&\quad + \sum_s \sum_{o_{s,v}=0} \mathbb{E}[g_{s,k}]\mathbb{E}[1-h_{s,v,k}] \ln(1-v_k) + (\alpha-1)\ln v_k + (\beta-1)\ln(1-v_k) + const \\
\mathbb{E}[v_k] &= \frac{\sum_s \sum_{o_{s,v}=1} \mathbb{E}[g_{s,k}] + \sum_s \sum_{o_{s,v}=0} \mathbb{E}[g_{s,k}]\mathbb{E}[h_{s,v,k}] + \alpha}{\sum_s \sum_{o_{s,v}=1} \mathbb{E}[g_{s,k}] + \sum_s \sum_{o_{s,v}=0} \mathbb{E}[g_{s,k}]\mathbb{E}[h_{s,v,k}] + \alpha + \beta + \sum_s \sum_{o_{s,v}=0} \mathbb{E}[g_{s,k}]\mathbb{E}[1-h_{s,v,k}]}
\end{aligned}
$$

$$
\begin{aligned}
\ln q(G) &= \mathbb{E}_{U,V,H}(\ln p(G|U) + \ln(H|G,U)) + const \\
\ln q(g_{s,k}) &= \ln p(g_{s,k}|u_k) + \sum_{o_{s,v}=1} p(h_{s,v,k}=1|g_{s,k},v_k) + \sum_{o_{s,v}=0} p(h_{s,v,k}|g_{s,k},v_k) \\
&= g_{s,k}\{\mathbb{E}[\ln u_k] + \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=1} \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=0} [\mathbb{E}[h_{s,v,k}]\mathbb{E}[\ln v_k] + \mathbb{E}[1-h_{s,v,k}]\mathbb{E}[1-\ln v_k] \\
&\quad + (1-g_{s,k})\{\mathbb{E}[\ln(1-u_k)] + \sum_{o_{s,v}} \ln\frac{1}{2}\} \\
\mathbb{E}[g_{s,k}] &= \frac{\mathbb{E}[\ln u_k] + \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=1} \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=0} [\mathbb{E}[h_{s,v,k}]\mathbb{E}[\ln v_k] + \mathbb{E}[}{\mathbb{E}[\ln u_k] + \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=1} \mathbb{E}[\ln v_k] + \sum_{o_{s,v}=0} [\mathbb{E}[h_{s,v,k}]\mathbb{E}[\ln v_k] + \mathbb{E}[1-h_{s,v,k}]\mathbb{E}[1-\ln v_k}
\end{aligned}
$$

where $\mathbb{E}[\ln(u_k) = \phi(a')\phi(a'+b'), \mathbb{E}[\ln(v_k) = \phi(\alpha')\phi(\beta')$.

$$
\begin{aligned}
\ln q(H) &= \mathbb{E}_{G,U,V}(\ln p(H|V,G) + \ln p(O|H)) \\
\ln q(h_{s,v,k}=1) &= \sum_{o_{s,v}=1} \sum_k p(h_{s,v,k}=1|g_{s,k},v_k) + \sum_{o_{s,v}=0} \sum_{h_{s,v,k'}} p(o_{s,v}=0|H)p(h_{s,v,k'}|g_{s,k'},v_k) + const \\
\mathbb{E}[h_{s,v,k}] &= q(h_{s,v,k}=1) + q(h_{s,v,k}=0)
\end{aligned}
$$