Bayesian Binomial Mixture Model for Collaborative Prediction With Non-Random Missing Data

Yong-Deok Kim and Seungjin Choi
Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
{karma13,seungjin}@postech.ac.kr

ABSTRACT

Collaborative prediction involves filling in missing entries of a user-item matrix to predict preferences of users based on their observed preferences. Most of existing models assume that the data is *missing at random* (MAR), which is often violated in recommender systems in practice. Incorrect assumption on missing data ignores the missing data mechanism, leading to biased inferences and prediction. In this paper we present a Bayesian binomial mixture model for collaborative prediction, where the generative process for data and missing data mechanism are jointly modeled to handle *non-random missing data*. Missing data mechanism is modeled by three factors, each of which is related to users, items, and rating values. Each factor is modeled by Bernoulli random variable, and the observation of rating value is determined by the Boolean OR operation of three binary variables. We develop computationally-efficient variational inference algorithms, where variational parameters have closed-form update rules and the computational complexity depends on the number of observed ratings, instead of the size of the rating data matrix. We also discuss implementation issues on hyperparameter tuning and estimation based on empirical Bayes. Experiments on Yahoo! Music and MovieLens datasets confirm the useful behavior of our model by demonstrating that: (1) it outperforms state-of-the-art methods in yielding higher predictive performance; (2) it finds meaningful solutions instead of undesirable boundary solutions; (3) it provides rating trend analysis on why ratings are observed.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; I.2.6 [**Artificial Intelligence**]: Parameter learning

Keywords

Collaborative filtering; non-random missing data; probabilistic models; recommender systems; variational Bayesian inference

RecSys'14, October 6âĂŞ10, 2014, San Jose or vicinity, CA, USA

http://dx.doi.org/10.1145/2645710.2645754.

1. INTRODUCTION

Collaborative prediction is a popular technique used in recommender systems, the task of which is to fill in missing entries of a user-item matrix given a small set of observed entries, to predict preferences of users based on their observed preferences. Suppose that we are given a set of observed entries, $\boldsymbol{X}_\Omega = \{X_{ij} \,|\, (i,j) \in \Omega, 1 \le i \le I, 1 \le j \le J\}$, where $X_{ij} \in \{1, \ldots, V\}$ represents a value of rating by user i on item j, and Ω is the set of (i,j) pairs for which the corresponding X_{ij} is observed. Then, collaborative prediction involves inferring $\boldsymbol{X}_{\Omega^c}$ given $\boldsymbol{X}_\Omega$, where $\boldsymbol{X}_\Omega \cup \boldsymbol{X}_{\Omega^c} = \boldsymbol{X} \in \mathbb{R}^{I \times J}$.

The real-world collaborative prediction problem is challenging computationally as well as statistically, since the user-item matrix is extremely large and sparse. For example, in the Yahoo! Music dataset [2], there are 1000990 users and 624961 items, but only 262810175 observed ratings are available, which constitutes 0.0042 percent of $\boldsymbol{X}$. In addition to size and sparseness problems, the collaborative prediction is more than completing a matrix, because of *non-ignorable missing data mechanism*, which is a main concern of this paper.

It follows from the theory of missing data that the mechanism for missing data are divided into three cases: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR) [7]. MCAR is the most restrictive assumption, where the probability that X_{ij} is observed does not depend on the data in any way. Most of theories developed for matrix completion assume MCAR, where missing or non-missing is decided by coin-flips, regardless of index (i,j) and rating value [1]. However, the MCAR assumption is easily violated in collaborative prediction problems. For example, some users rate items more actively than others and some items are rated by many users while others are rarely rated. MAR is the case where missing or non-missing depends on the index (i,j) but does not depend on rating value X_{ij}. More precisely, in the MAR assumption, the probability that a data point is observed is related to particular variables but is not related to the values of those variables. Most of previous methods for collaborative prediction have been developed, assuming MAR. They include theory on weighted trace norm regularization for non-uniform sampling [18, 3], probabilistic models such as URP [8] and RBM [17], and Bayesian matrix factorization (BMF) [5, 15, 16, 21, 14, 4].

When the data is MAR, the missing data mechanism can be ignored and parameter estimation or inferences are unbiased. However there is a strong evidence that MAR is an incorrect assumption for collaborative prediction [11]. Intuitively, the MAR assumption is easily violated in recommendation system if, for instance: (1) users rate items they like more than ones they dislike; (2) users provides extreme ratings, that is, they only provide feedback particu-

larly good or bad items. In these examples, dependencies between the missing data values and observation process exist, hence the data is not MAR. When the MAR assumption does not hold, inferences are biased, leading to the degraded prediction performance.

An approach to dealing with MNAR data is to learn jointly a *complete data model* that explain how the data is generated, and a *missing data model* that explains the observation process for the data [7]. In this paper, we choose a binomial mixture model as the complete data model for the ordinal rating data matrix and propose a novel missing data model that explains the reason of observation as the function of three factors, each of which related with users, items, and rating values. Each factor is modelled by Bernoulli random variable, and the observation of rating value is determined by the Boolean OR operation of three binary variables. Because of the property of OR operation, one of three factors can override others, hence it naturally models the 80-20 rule, which commonly arise in the recommendation systems.

The combination of binomial mixture model and our novel missing data model makes possible to develop computationally efficient variational inference algorithms, where variational parameters have closed-form update rules and the computational complexity depends on the number of observed ratings, instead the size of the rating data matrix. In fact for this reason, we select the binomial mixture model as the complete data model instead of widely used matrix factorization model. We also discuss implementation issues on hyperparameter tuning and estimation based on empirical Bayes.

Experiments on *Yahoo! Musing ratings for User Selected and Randomly Selected Songs* dataset show that our model outperforms the state-of-the-art method for MNAR data, which is a combination of multinomial mixture model and CPT-v+ missing data model [11, 9, 10]. Our model finds meaningful solution instead of undesirable boundary solution (see Figure 2-(d)), if hyperparameters are carefully estimated by empirical Bayes method. In addition to higher predictive performance on MNAR data, our model provides rating trend analysis on why ratings are observed. Preliminary experiments on Yahoo! Music and MovieLens datasets show that our model can capture the different rating trend between two domains: song and movie. Throughout the paper, we use notations described in Table 1.

Table 1: Notation

Notation	Description
Ω	Set of indices (i,j) for which X_{ij} is observed.
Ω_i	Set of indices j for which X_{ij} is observed.
Ω_j	Set of indices i for which X_{ij} is observed.
Ω_v	Set of indices (i,j) for which X_{ij} is v.
Ω^c	Complementary set of Ω.
$\boldsymbol{z}_i$	The ith column vector of $\boldsymbol{Z}$.
$\boldsymbol{X}_\Omega$	$\{X_{ij}\|(i,j) \in \Omega\}$.
$\boldsymbol{X}_{\Omega^c}$	$\{X_{ij}\|(i,j) \in \Omega^c\}$.
H	The number of observed ratings.
H_i	The number of ratings by user i.
H_j	The number of ratings for item j.
H_v	The number of ratings with value v.
$\mathbf{1}[s]$	Indicator function that takes 1 if s is true.
$\text{Bin}(x\|p,N)$	Binomial distribution, p : success probability, N : and the number of trial.
$\text{Dir}(\boldsymbol{\theta}\|\alpha,K)$	Symmetric Dirichlet distribution, α : concentration parameter, K : the number of category .
$\text{Beta}(x\|a,b)$	Beta distribution, a,b : shape parameters.
$\text{Bern}(x\|p)$	Bernoulli distribution, p : success probability.

2. RELATED WORK

We introduce a companion matrix of response indicator $\boldsymbol{R}$, where $R_{ij} = 1$ if X_{ij} is observed, and $R_{ij} = 0$ if X_{ij} is missing. Following the missing data theory [7], we formulate the non-ignorable missing data mechanism as a two-step procedure. First a complete data model with parameter Θ, $p(\boldsymbol{X}|\Theta)$, generates the complete data matrix $\boldsymbol{X} = \boldsymbol{X}_\Omega \cup \boldsymbol{X}_{\Omega^c}$. Then a missing data model with parameter Υ, $p(\boldsymbol{R}|\boldsymbol{X},\Upsilon)$, determines which elements in $\boldsymbol{X}$ are observed. Hence, we can take a factorized joint distribution for $\boldsymbol{X}$ and $\boldsymbol{R}$ given Θ and Υ:

$$p(\boldsymbol{R},\boldsymbol{X}|\Theta,\Upsilon) = p(\boldsymbol{R}|\boldsymbol{X},\Upsilon)p(\boldsymbol{X}|\Theta). \quad (1)$$

Most prior researches on collaborative prediction focus on the estimation of the complete data model $p(\boldsymbol{X}|\Theta)$, and usually ignore the missing data model $p(\boldsymbol{R}|\boldsymbol{X},\Upsilon)$. Important exceptions are Marlin's works, where the *multinomial mixture* (MM) model is combined with two different missing data models called *CPT-v* and *Logit-vd* [11, 9, 10]. The first, CPT-v is a simple missing data model where the probability of observing a rating depends only on the underlying rating value. This model can capture the intuition that a user's preference for a particular item can influence whether the user rates that item. The CPT-v missing data model is parameterized by using a conditional probability table consisting of V Bernoulli parameters $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_V]$ (hence the name CPT-v). The parameter γ_v gives the probability that an item will be rated if its true rating value is v: $p(R_{ij} = 1|X_{ij} = v) = \gamma_v$.

The CPT-v model makes the very strong assumption that a single value-based selection effect is responsible for generating all missing data. The second, Logit-vd is more flexible missing data model because it allows the probability of observation to depend both on the underlying rating value and the identity of the item. The Logit-vd model specifies a logistic form for this relationship as:

$$p(R_{ij} = 1|X_{ij} = v) = \frac{1}{1 + \exp(-(\sigma_v + \omega_j))}, \quad (2)$$

where σ_v models a non-random missing data effect that depend on the underlying rating value, and w_j is a per-item bias. This parameterization is more suitable than simple factorizations, such as product of Bernoulli probabilities

$$p(R_{ij} = 1|X_{ij} = v) = \gamma_v\, \nu_j \leq \min(\gamma_v, \nu_j), \quad (3)$$

because (2) can model *80-20 rule* which commonly arises in the recommendation system. In practice, some items are rated by many users, while others are rarely rated. The Logit-vd model can account for this by setting to ω_j to a large value. The logistic combination rule (2) allows ω_j to override σ_v, and produces a high selection probability for these extremely popular items.

Compare to the widely used matrix factorization model, the MM model is less flexible as the complete data model. However this simplicity makes possible to develop computationally efficient learning algorithm based on the expectation maximization (EM), when it is combined with CPT-v or Logit-vd missing data model. In addition, experimental results on Yahoo! music dataset showed that the MM model with these missing data models significantly outperforms other collaborative prediction models based on MAR assumption.

Although Marlin's works opened the door for collaborative prediction with non-random missing data and showed promising results, several problems still remain unsolved. At first, when MM model and missing data model are jointly learned, the EM algorithm converges to undesirable boundary solution, where almost all of the missing rating data are predicted to the value 2, as shown in Fig. 2-(d). In fact the best performing model, referred to *MM/CPT-v+* in [9], was obtained by an unnatural way, where only parameters for the MM model is learned by EM algorithm, while parameters for the CPT-v model is fixed to optimal values

$$\hat{\boldsymbol{\gamma}} = [0.014, 0.011, 0.027, 0.063, 0.225], \tag{4}$$

which is estimated by using a held-out dataset. Interestingly joint learning of complete data model and missing data model was converged to undesirable boundary solution even strong informative priors on $\boldsymbol{\gamma}$ were given by using $\hat{\boldsymbol{\gamma}}$ [11, 9]. Secondly generalized EM algorithm for Logit-vd does not have close-form update rules, hence requires careful tuning of learning rate for gradient assent updates and a large number of iteration for convergence (more than 5000 iteration). In addition, although Logit-vd showed better predictive performance than CPT-v [10], it is still significantly worse than CPT-v+[9].

In this paper, we thoroughly try to solve the aforementioned problems. As a part of solution, we choose the binomial mixture model for modelling rating data because it takes account of the ordinal nature of rating values, which is ignored in the MM model. The idea of using the binomial distribution for modelling ordinal nature of rating values can also be found in binomial matrix factorization model [20], but it is based on the MAR assumption.

There are also several works which consider MNAR rating data in the widely used matrix factorization framework. Instead of explicitly modelling missing data model, AllRank model optimizes a ranking-based metric that is robust to MNAR data [19]. The resulting method can be applied to recommendation, but not to rating prediction task.

Response aware probabilistic matrix factorization model [6] combined probabilistic matrix factorization with missing data model which is similar to [11, 10]. Although this model is based on powerful matrix factorization, it has several weak points: (1) there is no closed form update rules; (2) careful tuning of learning rate and other meta parameters is required, hence reproducibility is bad; (3) the computational complexity is high, it depends on data matrix size instead of the number of observed ratings. In our empirical study, reported RMSE of this model on the randomly selected Yahoo! music dataset is significantly worse than MM/CPT-v+.

Another variation of probabilistic matrix factorization for MNAR rating data is specialized to implicit (a.k.a binary or one-class) feedback case [13], where Bayesian generative process for implicit rating data is proposed. It models the *like* probability by interpreting the missing signal as a two-stage process: firstly, by modelling the odds of a user considering an item, and secondly, by eliciting a probability that the item will be view or liked. In this paper we model the explicit MNAR rating data, and it is harder than implicit one because common users usually rate on the good item but sometimes will rate on the bad item to express their strong dissatisfaction.

3. BAYESIAN-BM/OR

We present our Bayesian binomial mixture model for collaborative prediction with non-random missing data, which is referred to as Bayesian-BM/OR in this paper. The graphical model for Bayesian-BM/OR is shown in Figure 1.

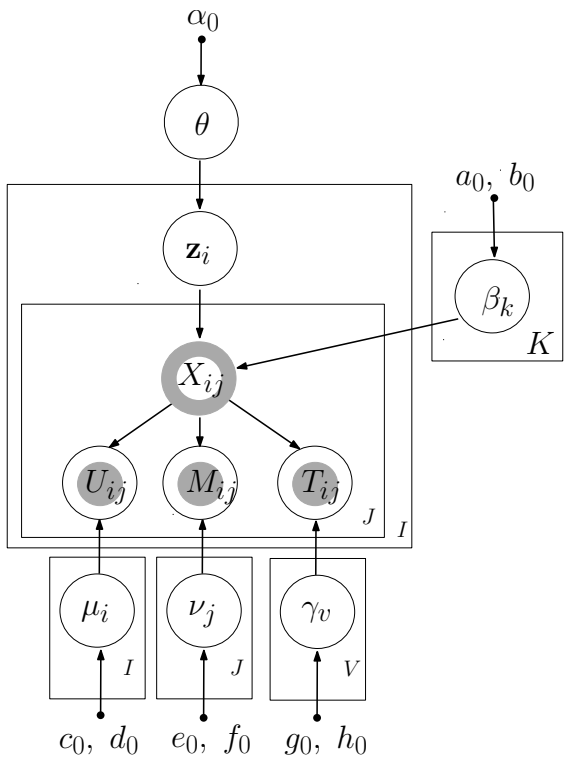


Figure 1: A graphical model for binomial mixture with our non-random missing data model. Blank and hollow circles denote unobserved and partially observed variables respectively. Note that if X_{ij} is observed, then U_{ij}, M_{ij}, T_{ij} are unobserved, and vice versa.

3.1 Model

The generative process of the rating data matrix and missing data mechanism are as follows.

1. Choose the number of clusters K.
2. Choose mixing proportions $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\theta}|\alpha_0, K)$,
3. Choose parameters for binomial distributions

$$\boldsymbol{\beta} \sim \prod_{k=1}^{K}\prod_{j=1}^{J} \text{Beta}(\beta_{kj}|a_0, b_0)$$

4. For each user $i \in \{1, \ldots, I\}$
   (a) Choose a cluster indicator
   $$\boldsymbol{z}_i|\boldsymbol{\theta} \sim \text{Mult}(\boldsymbol{z}_i|\boldsymbol{\theta}) = \textstyle\prod_{k=1}^{K} \theta_k{}^{Z_{ki}}.$$
   (b) For each item $j \in \{1, \ldots, J\}$
   Choose a rating
   $$X_{ij}|\boldsymbol{z}_i, \boldsymbol{\beta} \sim \textstyle\prod_{k=1}^{K} \text{Bin}(X_{ij} - 1|\beta_{kj}, V - 1)^{Z_{ki}}.$$
5. Choose per-user, per-item, and per-value parameters for missing data model
$$\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma} \sim \prod_{i=1}^{I} \text{Beta}(\mu_i|c_0, d_0) \prod_{j=1}^{J} \text{Beta}(\nu_j|e_0, f_0) \prod_{v=1}^{V} \text{Beta}(\gamma_v|g_0, h_0)$$
6. For each $(i, j) \in \{1, \ldots, I\} \times \{1, \ldots, J\}$
   (a) Choose $U_{ij}|X_{ij}, \boldsymbol{\mu} \sim \text{Bern}(U_{ij}|\mu_i)$

(b) Choose $M_{ij}|X_{ij}, \boldsymbol{\nu} \sim \text{Bern}(M_{ij}|\nu_j)$
(c) Choose $T_{ij}|X_{ij}, \boldsymbol{\gamma} \sim \text{Bern}(T_{ij}|\gamma_{X_{ij}})$
(d) $R_{ij} = 1 - (1 - U_{ij})(1 - M_{ij})(1 - T_{ij})$.
X_{ij} is observed if $R_{ij} = 1$, otherwise X_{ij} is missing.

We choose a binomial mixture model as the complete data model for the rating data matrix because (1) the mixture model can capture the simple intuition that users form clusters according to their preferences for items; (2) the binomial distribution is suitable for modelling discrete, finite, and ordered rating values; (3) combined with our missing data model, it allows the efficient learning algorithm, of which time and space complexity depend on the number of observed rating instead the data matrix size.

We consider following conditions for the missing data model: (1) user activity, item popularity, value-based selection effect should be modeled; (2) one of these factors can override other factors such that the 80-20 rule can be well modeled; (3) combined with binomial mixture model, the computationally efficient learning algorithm should be derived.

One possible solution will be a extending the Logit-vd model by simply adding a per-user bias τ_i:

$$p(R_{ij} = 1|X_{ij} = v) = \Lambda_{ijv} = \frac{1}{1 + \exp(-(\sigma_v + \tau_i + \omega_j))}. \quad (5)$$

However unlike CPT-v and Logit-vd, the parameterization (5) does not allow computationally efficient learning algorithm. For example, the complexity for computing a log-likelihood depends on the size of rating matrix IJ, not on the number of observed rating H:

$$\begin{aligned}
\log p(\boldsymbol{R}, \boldsymbol{X}_\Omega) &= \log p(\boldsymbol{R}_\Omega, \boldsymbol{X}_\Omega) + \log p(\boldsymbol{R}_{\Omega^c}), \quad (6)\\
\log p(\boldsymbol{R}_\Omega, \boldsymbol{X}_\Omega) &= \sum_{(i,j)\in\Omega} \log(\sum_{k=1}^{K} \theta_k P_{kjX_{ij}} \Lambda_{ijX_{ij}}),\\
\log p(\boldsymbol{R}_{\Omega^c}) &= \sum_{(i,j)\in\Omega^c} \log\left(\sum_{k=1}^{K} \theta_k \sum_{v=1}^{V} P_{kjv}(1 - \Lambda_{ijv})\right),
\end{aligned}$$

where $P_{kjv} = p(X_{ij} = v|Z_{ki} = 1)$. Note that sum operations are involved with all entry $(i, j) \in \Omega \cup \Omega^c$.

Instead of the logistic combination rules, we present a novel missing data model which fulfils all above-mentioned conditions. We assume that for each entry (i, j), the observation of rating value X_{ij} is determined by the Boolean OR operation of three binary random variables U_{ij}, M_{ij}, and T_{ij}, each of which is related with users, items, and rating values:

$$R_{ij} = U_{ij} \vee M_{ij} \vee T_{ij} = 1 - (1 - U_{ij})(1 - M_{ij})(1 - T_{ij}).$$

In addition, we assume that these binary random variables follow the Bernoulli distribution conditioned on X_{ij}:

$$\begin{aligned}
&p(U_{ij}, M_{ij}, T_{ij}|X_{ij} = v, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma})\\
&\quad = \text{Bern}(U_{ij}|\mu_i)\text{Bern}(M_{ij}|\nu_j)\text{Bern}(T_{ij}|\gamma_v). \quad (7)
\end{aligned}$$

More precisely speaking, U_{ij}, M_{ij}, and T_{ij} are conditioned on the row index, column index, and value of X_{ij}. With our missing data model, the probability of observation is computed by

$$\begin{aligned}
&p(R_{ij} = 1|X_{ij} = v)\\
&\quad = 1 - p(R_{ij} = 0|X_{ij} = v)\\
&\quad = 1 - p(U_{ij} = 0, M_{ij} = 0, T_{ij} = 0|X_{ij} = v)\\
&\quad = 1 - (1 - \mu_i)(1 - \nu_j)(1 - \gamma_v). \quad (8)
\end{aligned}$$

Our missing data model naturally allows that one of three factors can override the rest, because of following inequality:

$$1 - (1 - \mu_i)(1 - \nu_j)(1 - \gamma_v) \geq \max(\mu_i, \nu_j, \gamma_v). \quad (9)$$

Table 2: Examples of the probability of observation (8)

μ_i	ν_j	γ_v	$p(R_{ij} = 1|X_{ij} = v)$
0.01	0.01	0.01	0.029
0.01	0.01	0.1	0.118
0.01	0.1	0.1	0.198
0.1	0.1	0.1	0.271
0.9	0.1	0.1	0.919
0.9	0.5	0.1	0.955

A behaviour of probability of observation under various combination of (μ_i, ν_j, γ_v) is illustrated in Table 2.

3.2 Variational Inference

Note that there is a interesting reverse observed/missing relationship between X_{ij} and (U_{ij}, V_{ij}, R_{ij}). If X_{ij} is observed (i.e. $R_{ij} = 1$), we only know that at least one of U_{ij}, M_{ij}, T_{ij} is one, hence they are latent variables. However if X_{ij} is missing (i.e. $R_{ij} = 0$), then $U_{ij} = M_{ij} = T_{ij} = 0$, consequently they are observed variables. Hence the set of of observed variables is defined by $\mathcal{D} = \{\boldsymbol{X}_\Omega, \boldsymbol{U}_{\Omega^c}, \boldsymbol{M}_{\Omega^c}, \boldsymbol{T}_{\Omega^c}\}$ and the set of latent variable is defined by $\mathcal{Z} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{X}_{\Omega^c}, \boldsymbol{Z}, \boldsymbol{U}_\Omega, \boldsymbol{M}_\Omega, \boldsymbol{T}_\Omega\}$.

We approximately compute posterior distributions over latent variables by maximizing a lower-bound on the marginal log-likelihood. The marginal log-likelihood $\log p(\mathcal{D})$ is given by

$$\begin{aligned}
\log p(\mathcal{D}) &= \log \int p(\mathcal{D}, \mathcal{Z}) d\mathcal{Z}\\
&\geq \int q(\mathcal{Z}) \log \frac{p(\mathcal{D}, \mathcal{Z})}{q(\mathcal{Z})} d\mathcal{Z} \equiv \mathcal{F}(q), \quad (10)
\end{aligned}$$

where Jensen's inequality was used to obtain the *variational lower-bound* $\mathcal{F}(q)$ and $q(\mathcal{Z})$ is *variational distribution*. We assume that the variational distribution factorizes as

$$q(\mathcal{Z}) = q(\boldsymbol{\theta})q(\boldsymbol{\beta})q(\boldsymbol{\mu})q(\boldsymbol{\nu})q(\boldsymbol{\gamma})q(\boldsymbol{X}_{\Omega^c}, \boldsymbol{Z})q(\boldsymbol{U}_\Omega, \boldsymbol{M}_\Omega, \boldsymbol{T}_\Omega).$$

Due to the space limitation, we omit the detailed derivation of variational inference algorithms and only present the final results in Table 3. But we emphasize that auxiliary latent variables $\boldsymbol{U}, \boldsymbol{M}, \boldsymbol{T}$ play key role for deriving closed-form update rules. Without them, we have to compute following expectation of log-likelihood

$$\langle \log p(R_{ij} = 1|X_{ij} = v, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}) \rangle = \langle \log(1 - \mu_i \nu_j \gamma_v) \rangle, \quad (11)$$

which is intractable because it is no more conjugate to Beta distribution. But the free-form optimization can be easily solved with auxiliary latent variables. For example, optimal $q(\mu_i)$ is given by

$$\begin{aligned}
&\log q(\mu_i)\\
&\propto \log p(\mu_i) + \sum_{j\in\Omega_i} \langle \log p(U_{ij}|X_{ij}, \mu_i) \rangle_{q(U_{ij})}\\
&\quad + \sum_{j\in\Omega_i^c} \sum_{v=1}^{V} q(X_{ij} = v) \log p(U_{ij} = 0|X_{ij} = v, \mu_i),\\
&\propto (c0 - 1) \log \mu_i + (d0 - 1) \log(1 - \mu_i)\\
&\quad + \sum_{j\in\Omega_i} (\langle U_{ij} \rangle \log \mu_i + \langle 1 - U_{ij} \rangle \log(1 - \mu_i))\\
&\quad + \sum_{j\in\Omega_i^c} \log(1 - \mu_i).
\end{aligned}$$

Taking the exponential on both sizes, we recognize that $q(\mu_i)$ is $\text{Beta}(\mu_i|c_i, d_i)$, where c_i and d_i are defined in Table 3.

Table 3: Variational distributions and corresponding updating rules and sufficient statistics are summarized.

Variational distributions	Updating rules and sufficient statistics	
$q(\boldsymbol{\theta}) = \text{Dir}(\boldsymbol{\theta}\|\boldsymbol{\alpha})$	$\alpha_k = \alpha_0 + \sum_{i=1}^{I}\langle Z_{ki}\rangle,$	$\langle \log \alpha_k\rangle = \psi(\alpha_k) - \psi(\sum_{k=1}^{K}\alpha_k).$
$q(\boldsymbol{\beta}) = \prod_{k,j}\text{Beta}(\beta_{kj}\|a_{kj}, b_{kj})$	$\xi_{kj} = \sum_{i\in\Omega_j} X_{ij}\langle Z_{ik}\rangle + \sum_{i\in\Omega_j^c}\langle X_{ij}Z_{ik}\rangle,$	$\langle\beta_{kj}\rangle = a_{kj}/(a_{kj}+a_{kj}),$
	$a_{kj} = a_0 - \sum_{i=1}^{I}\langle Z_{ik}\rangle + \xi_{kj},$	$\langle\log\beta_{kj}\rangle = \psi(a_{kj}) - \psi(a_{kj}+a_{kj}),$
	$b_{kj} = b_0 + V\sum_{i=1}^{I}\langle Z_{ik}\rangle - \xi_{kj},$	$\langle\log(1-\beta_{kj})\rangle = \psi(b_{kj}) - \psi(a_{kj}+b_{kj}).$
$q(\boldsymbol{\mu}) = \prod_{i=1}^{I}\text{Beta}(\mu_i\|c_i, d_i)$	$\xi_i = \sum_{j\in\Omega_i}\langle U_{ij}\rangle,$	$\langle\mu_i\rangle = c_i/(c_i+d_i),$
	$c_i = c_0 + \xi_i,$	$\langle\log\mu_i\rangle = \psi(c_i) - \psi(c_i+d_i),$
	$d_i = d_0 + J - \xi_i,$	$\langle\log(1-\mu_i)\rangle = \psi(d_i) - \psi(c_i+d_i).$
$q(\boldsymbol{\nu}) = \prod_{j=1}^{I}\text{Beta}(\nu_j\|e_j, f_j)$	$\xi_j = \sum_{i\in\Omega_j}\langle M_{ij}\rangle,$	$\langle\nu_j\rangle = e_j/(e_j+f_j),$
	$e_j = e_0 + \xi_j,$	$\langle\log\nu_j\rangle = \psi(e_j) - \psi(e_j+f_j),$
	$f_j = f_0 + I - \xi_j,$	$\langle\log(1-\nu_j)\rangle = \psi(f_j) - \psi(e_j+f_j).$
$q(\boldsymbol{\gamma}) = \prod_{v=1}^{V}\text{Beta}(\gamma_v\|g_v, h_v)$	$\xi_v = \sum_{(i,j)\in\Omega_v}\langle T_{ij}\rangle,$	$\langle\gamma_v\rangle = g_v/(g_v+h_v),$
	$g_v = g_0 + \xi_v,$	$\langle\log\gamma_v\rangle = \psi(g_v) - \psi(g_v+h_v).$
	$h_v = h_0 + H_v + \sum_{(i,j)\in\Omega^c} q(X_{ij}=v) - \xi_v,$	$\langle\log(1-\gamma_v)\rangle = \psi(h_v) - \psi(g_v+h_v).$
$q(\boldsymbol{X}_{\Omega^c}, \boldsymbol{Z})$ $= \prod_{i=1}^{I} q(X_{\Omega_i^c}\|\boldsymbol{z}_i)q(\boldsymbol{z}_i),$	$\tilde{\lambda}_{kjv} = \langle\log(1-\gamma_v)\rangle + \log\binom{V-1}{v-1} + v\langle\log\beta_{kj}\rangle + (V-v)\langle\log(1-\beta_{kj})\rangle,$	
	$\lambda_{kjv} = \exp(\tilde{\lambda}_{kjv})/\sum_{v'=1}^{V}\exp(\tilde{\lambda}_{kjv'}),$	
	$\phi_{kj} = \sum_{v=1}^{V}(v-1)\,\lambda_{kjv},$	
	$\bar{\phi}_{kj} = \sum_{v=1}^{V}\langle\log(1-\gamma_v)\rangle\lambda_{kjv},$	
$q(X_{\Omega_i^c}\|\boldsymbol{z}_i)$ $= \prod_{j\in\Omega_i^c} q(X_{ij}\|\boldsymbol{z}_i),$	$\hat{\phi}_{kj} = \sum_{v=1}^{V}\lambda_{kjv}\log\lambda_{kjv},$	
	$\tilde{\phi}_{kj} = \phi_{kj}\langle\log\beta_{kj}\rangle + (V-1-\phi_{kj})\langle\log(1-\beta_{kj})\rangle + \sum_{v=1}^{V}\lambda_{kjv}\log\binom{V-1}{v-1},$	
	$\tilde{\rho}_{ki} = \langle\log\theta_k\rangle + \sum_{j\in\Omega_i}[(X_{ij}\langle\log\beta_{kj}\rangle + (V-X_{ij})\langle\log(1-\beta_{kj})\rangle]$ $+ \sum_{j\in\Omega_i^c}(\tilde{\phi}_{kj} + \bar{\phi}_{kj} - \hat{\phi}_{kj}),$	
$q(X_{ij}=v\|\boldsymbol{z}_i) = \prod_{k=1}^{K}(\lambda_{kjv})^{Z_{ki}},$	$\rho_{ki} = \exp(\tilde{\rho}_{ki})/\sum_{k'=1}^{K}\exp(\tilde{\rho}_{k'i}),$	
	$\langle Z_{ki}\rangle = \rho_{ki},$	
$q(\boldsymbol{z}_i) = \prod_{k=1}^{K}(\rho_{ki})^{Z_{ki}}.$	$\langle X_{ij}Z_{ki}\rangle = \sum_{v=1}^{V}(v-1)\,q(X_{ij}=v\|Z_{ki}=1)q(Z_{ki}=1) = \sum_{v=1}^{V}(v-1)\,\lambda_{kjv}\rho_{ki} = \phi_{kj}\rho_{ki},$	
	$q(X_{ij}=v) = \sum_{k=1}^{K}q(X_{ij}=v\|Z_{ki}=1)q(Z_{ki}=1) = \sum_{k=1}^{K}\lambda_{kjv}\rho_{ki}.$	
$q(\boldsymbol{U}_\Omega, \boldsymbol{M}_\Omega, \boldsymbol{T}_\Omega)$ $= \prod_{(i,j)\in\Omega} q(U_{ij}, M_{ij}, T_{ij})$	For $(U_{ij}, M_{ij}, T_{ij}) \in \{0,1\}^3 - (0,0,0)$ $q(U_{ij}, M_{ij}, T_{ij}) = \frac{(\tilde{\mu}_i^1)^{U_{ij}}(\tilde{\mu}_i^0)^{1-U_{ij}}(\tilde{\nu}_j^1)^{M_{ij}}(\tilde{\nu}_j^0)^{1-M_{ij}}(\tilde{\gamma}_{X_{ij}}^1)^{T_{ij}}(\tilde{\gamma}_{X_{ij}}^0)^{1-T_{ij}}}{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)(\tilde{\nu}_j^1+\tilde{\nu}_j^0)(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)-\tilde{\mu}_i^0\tilde{\gamma}_j^0\tilde{\gamma}_{X_{ij}}^0},$ where	
	$\tilde{\mu}_i^1 = \exp(\langle\log\mu_i\rangle),$	$\tilde{\mu}_i^0 = \exp(\langle\log(1-\mu_i)\rangle),$
	$\tilde{\nu}_j^1 = \exp(\langle\log\nu_j\rangle),$	$\tilde{\nu}_j^0 = \exp(\langle\log(1-\nu_j)\rangle),$
	$\tilde{\gamma}_v^1 = \exp(\langle\log\gamma_v\rangle),$	$\tilde{\gamma}_v^0 = \exp(\langle\log(1-\gamma_v)\rangle).$
	$\langle U_{ij}\rangle = q(U_{ij}=1) = \frac{\tilde{\mu}_i^1(\tilde{\nu}_j^1+\tilde{\nu}_j^0)(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)}{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)(\tilde{\nu}_j^1+\tilde{\nu}_j^0)(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)-\tilde{\mu}_i^1\tilde{\gamma}_j^1\tilde{\gamma}_{X_{ij}}^1},$	
	$\langle M_{ij}\rangle = q(M_{ij}=1) = \frac{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)\tilde{\nu}_j^1(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)}{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)(\tilde{\nu}_j^1+\tilde{\nu}_j^0)(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)-\tilde{\mu}_i^1\tilde{\gamma}_j^1\tilde{\gamma}_{X_{ij}}^1},$	
	$\langle T_{ij}\rangle = q(T_{ij}=1) = \frac{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)(\tilde{\nu}_j^1+\tilde{\nu}_j^0)\tilde{\gamma}_{X_{ij}}^1}{(\tilde{\mu}_i^1+\tilde{\mu}_i^0)(\tilde{\nu}_j^1+\tilde{\nu}_j^0)(\tilde{\gamma}_{X_{ij}}^1+\tilde{\gamma}_{X_{ij}}^0)-\tilde{\mu}_i^0\tilde{\gamma}_j^0\tilde{\gamma}_{X_{ij}}^0}.$	

In Marlin's works, hyperparameters for the complete data model and missing data model are set to non-informative values and do not updated [11, 10]. However we empirically observed that hyperparameter estimation is crucial for finding meaningful solutions instead of undesirable boundary solutions. We estimate hyperparameters $\{a_0, b_0, c_0, d_0, e_0, f_0, g_0, h_0\}$ for Beta distributions and α_0 for symmetric Dirichlet distribution by empirical Bayes method, which maximizes the marginal log-likelihood of observed variables

$$\log p(\mathcal{D}|\alpha_0, a_0, b_0, c_0, d_0, e_0, f_0, g_0, h_0). \quad (12)$$

Since we can not compute the marginal log-likelihood exactly, an approximating approach is to maximize the variational lower bound (10). Unfortunately close-form updated rules can not be derived for these hyperparameters, hence numerical optimization methods are required. Empirically we observed that Newton-Raphson method do not work well for our problem since the resulting objective function is not well-approximated by a quadratic. Instead we apply the method in [12], which estimates the Dirchlet mean and precision separately. Note that Beta distribution is special case of Dirchlet distribution ($K = 2$).

3.3 Complexity Analysis

We analyse the time and space complexity of variational inference algorithms summarized in Table 3. We emphasize that precomputing and caching intermediate factors are very important to implement computationally efficient learning algorithms. Table 4 summarizes the time and space complexity of our learning algorithms, of which complexity is depend on the number of observations, instead of the rating data matrix size.

In the perspective of the space complexity, main computational burdens are come from the saving $\langle X_{ij} Z_{ki} \rangle$ and $q(X_{ij} = v)$, for all $(i, j) \in \Omega^c$. A required memory for saving them is

$$O((IJ - H)(K + V))),$$

and it is depends on the the data matrix size ($\because H \ll IJ$). Instead of saving $\langle X_{ij} Z_{ki} \rangle$ and $q(X_{ij} = v)$, we save intermediate factors $\rho_{ki}, \phi_{kj}, \lambda_{kjv}$. This approach not only reduces the space complexity but also the time complexity. For example, in the update rule for $q(\gamma_v)$, a naive implementation requires $O(IJ - H)$ time complexity because of summing $q(X_{ij} = v)$ for all $(i, j) \in \Omega^c$. However it can be efficiently computed in $O(HK)$ by

$$\sum_{j=1}^{J} \sum_{i \in \Omega_i^c} \sum_{k=1}^{K} \lambda_{kjv} \rho_{ki} = \underbrace{\sum_{j=1}^{J} \overbrace{\sum_{k=1}^{K} \lambda_{kjv} \underbrace{\sum_{i \in \Omega_i} (\dot{\rho}_k - \rho_{ki})}_{O(H_i)}}^{O(H_i K)}}_{O(HK)}, \quad (13)$$

where $\dot{\rho}_k = \sum_{i=1}^{I} \rho_{ki}$ is pre-computed factor. Using the similar idea, other updating rules also can be efficiently implemented. [1]

4. NUMERICAL EXPERIMENTS

4.1 Dataset

We performed experiments on two datasets: *Yahoo! Musing ratings for User Selected and Randomly Selected Songs, version 1.0* (Yahoo! music dataset)[2] and MovieLens dataset [3].

[1] Source code: http://mlg.postech.ac.kr/∼karma13

[2] http://webscope.sandbox.yahoo.com

[3] http://wwww.grouplens.org/datasets/movielens

Table 4: Complexity of variational inference algorithms

Variational distributions	Space	Time
$q(\boldsymbol{\theta})$	K	IK
$q(\boldsymbol{\beta})$	JK	HK
$q(\boldsymbol{\mu})$	I	H
$q(\boldsymbol{\nu})$	J	H
$q(\boldsymbol{\gamma})$	V	HKV
$q(\boldsymbol{X}_{\Omega^c}, \boldsymbol{Z})$	$(I + J + JV)K$	$(H + JV)K$
$q(\boldsymbol{U}_\Omega, \boldsymbol{M}_\Omega, \boldsymbol{T}_\Omega)$	H	H

Yahoo! Music dataset : This dataset provides a unique opportunity to test collaborative prediction methods that incorporate missing data models. The dataset consists of ratings collected during normal user interaction with Yahoo's LaunchCast internet radio service, as well as ratings for items collected using an online survey. The set of ratings collected through normal interaction with recommendation system are referred to as *ratings for user-selected items* and denoted by $\boldsymbol{X}^u$. There are 15400 users, 1000 songs, and 311704 ratings in $\boldsymbol{X}^u$. The set of ratings collected through survey are referred to as *ratings for randomly selected items* and denoted by $\boldsymbol{X}^r$. During a survey conducted by Yahoo! Research, exactly 10 songs are randomly selected and presented to each survey user. The survey users are enforced to rate on these randomly selected songs. In total 5400 users participated in this survey, hence $\boldsymbol{X}^r$ contains 54000 ratings.

Figure 2-(a) and 2-(b) show the marginal distribution of ratings for randomly selected items, $\boldsymbol{X}^r$, compared to the distribution of rating for user-selected items, $\boldsymbol{X}^u$. The two sets of ratings $\boldsymbol{X}^r$ and $\boldsymbol{X}^u$ exhibit significantly different marginal statistics. The most remarkable feature of the ratings in $\boldsymbol{X}^r$ is that they contain much fewer high ratings ($v = 5$) compared to the ratings for $\boldsymbol{X}^u$. This provides strong evidence on the violation of MAR condition in the Yahoo! ratings for user-selected items.

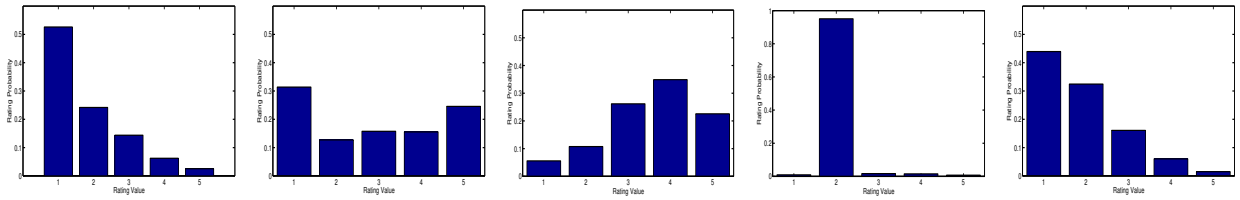
Our experimental protocol with the Yahoo! dataset is quite simple. We train the model on the ratings for user-selected items $\boldsymbol{X}^u$, and test on the ratings for randomly selected items $\boldsymbol{X}^r$. We used both root mean squared error (RMSE) and mean absolute error (MAE) to measure the predictive performance.

MovieLens dataset: We used MovieLens-1M dataset which consist of 1000209 ratings with 6040 users and 3706 movies. Although MovieLens dataset only consists of ratings for user-selected movies, it is well worth comparing two trained models from different domains (songs and movies), because the missing data mechanism for each domain can be different. The different behaviour of marginal statistics shown in Figure 2-(b) and 2-(c) support our assumption.

4.2 Results on Yahoo! Music Dataset

Firstly, we investigated on the benefit of binomial mixture model over multinomial mixture model. We combined binomial mixture model with CPT-v missing data model (referred to as BM/CPT-v). The number of cluster K is set to 10, and all hyperparameters for Beta distributions are set to 1 (i.e. uniform prior). Surprisingly BM/CPT-v model significantly outperformed the MM/CPT-v and it was comparable with MM/CPT-v+ as shown in Table 5.

However we observed that BM/CPT-v converge to a different kind of boundary solution. While MM/CPT-v found the boundary solution which predict almost all of the missing rating to the value 2 (see Figure 2-(d)), BM/CPT-v found the boundary solution which never predict value 5 for missing rating. The marginal statistics of



(a) Yahoo! Random (b) Yahoo! User (c) MovieLens (d) MM/CPT-v (e) Bayesian BM/OR

Figure 2: The sub-figure (a), (b), and (c) are distribution rating values for randomly selected Yahoo!, user-selected Yahoo!, and MovieLens dataset respectively. The sub-figure (d) and (e) are prediction results on the Yahoo! Random dataset by MM/CPT-V and Bayesian-BM/OR model respectively.

Table 5: RMSE, MAE, and trained parameters for missing data model (γ_v) of MM/CPT-v and BM/CPT-v on different hyper-parameter settings. Uniform denote uniform Beta prior ($g_0 = h_0 = 1$). In other case, $g_0/(g_0 + h_0)$ is fixed to 0.02, and scale of $(g_0 + h_0)$ is controlled from $H \times 10^{-3}$ to $H \times 10^{-1}$. RMSE and MAE of MM/CPT-v+ model are 0.989 and 0.727, respectively.

MM/CPT-v				
	Uniform	$H \times 10^{-3}$	$H \times 10^{-2}$	$H \times 10^{-1}$
RMSE	1.1057	1.054	1.057	1.070
MAE	0.869	0.868	0.867	0.873
γ_1	0.437	0.365	0.247	0.098
γ_2	0.003	0.003	0.003	0.003
γ_3	0.191	0.174	0.143	0.070
γ_4	0.207	0.195	0.153	0.073
γ_5	0.493	0.423	0.261	0.109
BM/CPT-v				
	Uniform	$H \times 10^{-3}$	$H \times 10^{-2}$	$H \times 10^{-1}$
RMSE	1.011	1.012	**0.997**	1.000
MAE	0.712	0.709	**0.704**	0.708
γ_1	0.011	0.011	0.011	0.017
γ_2	0.009	0.009	0.009	0.008
γ_3	0.034	0.035	0.030	0.018
γ_4	0.139	0.136	0.105	0.045
γ_5	0.999	0.943	0.640	0.186

predicted missing ratings was $[0.533, 0.3420.105, 0.002, 0.000]$. In addition we also observed that missing data model parameter γ_5 was also converged to boundary solution as shown in Table 5. This is quite undesirable result, because $\gamma_5 = 1$ implies that, all of the rating with the highest value 5 is already observed, consequently it implied that there is no more item with rating value 5.

We performed a second set of experiments where informative prior is given for missing data model parameter $\boldsymbol{\gamma}$ to suppress the over-estimated γ_5. Based on the observation that the sparsity of the rating data matrix is about 2 percent, hyperparameters of Beta prior are set to $g_0 = 0.02S$ and $h_0 = 0.98S$, such that mean of Beta prior is matched to 2 percent. S is the prior strength. We controlled the prior strength between $H \times 10^{-3}$ and $H \times 10^{-1}$. We observed that as the prior strength is increased, γ_5 is suppressed well. Especially in the case of $S = H \times 10^{-1}$, the prediction accuracy is good and trained missing data model parameters are similar to optimal ones used in CPT-v+ (4). We computed the variational lower bound for all trained models to check that we can select a meaningful model by empirical Bayes method. Against our expectations, the variational lower bound is maximized when uniform priors are set to missing data model parameters such that boundary solution is learned.

Finally we trained our Bayesian-BM/OR model, which considers not only the value-based selection effect but also user activity and item popularity. All hyperparameters for Dirichlet and Beta priors are initialized to 1, and variational parameters are initialized as

$$\begin{aligned} a_{kj} &= 1 + \epsilon_{kj}, \quad & b_{kj} &= 4 + \epsilon_{kj}, \\ c_i &= H_i/J, & d_i &= 1 - H_i/J, \\ e_j &= H_j/I, & f_j &= 1 - H_j/I, \\ g_v &= H_v/H, & h_v &= 1 - H_v/H, \end{aligned}$$

where ϵ_{kj} are Gaussian random noise with standard deviation 0.01. Bayesian-BM/OR model showed best performing RMSE and MAE, which are 0.983 and 0.699 respectively. In addition the model did not converges to boundary solution. The marginal statistics of predicted missing ratings are shown in Figure 2-(e).

We observed that Bayesian-BM/OR model can also be converged to boundary solution if hyperparameters are not optimized. In this case, γ_2 was converged to zero. From these results, we concluded that both three factors are equally important to learn the meaningful model from the MNAR rating data:

- Modeling the ordered property of rating values.
- Modeling the activity of user, popularity of item, and value-based selection effect in the missing data model.
- Estimating hyper parameters via empirical Bayes method.

4.3 Rating Trend Analysis

In addition to outperforming predictive performance on MNAR data, Bayesin-BM/OR model has another advantage which other models do not have. For observed rating X_{ij}, by using variational posterior $q(U_{ij}, M_{ij}, T_{ij})$, our model can explain the reason of observation with user, item, and rating value factor. It can be used in analysis and comparison of rating trend between different recommendation systems.

Table 6 shows the our preliminary rating trend analysis on two datasets: Yahoo! music and MovieLens. Each value in the table is computed by summing and normalizing $q(U_{ij} = 1)$, $q(M_{ij} = 1)$, and $q(T_{ij} = 1)$ for $(i, j) \in \Omega_v$. For the ratings with value 5, the main reason of observation is rating value factor on both datasets, that means users rate on good items because they like them. However for the ratings with lower values 1 or 2, we can observe the difference between two datasets. In both datasets user and item factor are larger than rating value factor, that means some active users rate on bad items or popular items are rated even though their rating is low. However the rating value factor for lower ratings on in MovieLens (0.044, 0.075) is extremely smaller than that of Yahoo! music (0.213, 0.161). We guess the reason of this pattern following way: watching the movie require more money and time than

Table 6: Preliminary rating trend analysis on Yahoo! Music and MovienLens datasets.

Yahoo! music dataset				MovieLens dataset			
v	User	Item	Value	v	User	Item	Value
All	0.266	0.390	0.421	All	0.263	0.330	0.448
1	0.377	0.416	0.213	1	0.579	0.387	0.044
2	0.303	0.540	0.161	2	0.509	0.427	0.075
3	0.177	0.511	0.320	3	0.352	0.377	0.2943
4	0.084	0.424	0.506	4	0.178	0.291	0.580
5	0.035	0.180	0.832	5	0.098	0.274	0.701

listening the song in the internet radio, hence people may select the movie more conservative way than song.

5. CONCLUSIONS

We have presented a Bayesian binomial mixture model for collaborative prediction, where the generative process for the data and missing data mechanism are jointly modeled to handle non-random missing data. Missing data mechanism was modeled by three factors, each of which is related to users, items, and rating values. Each factor was modeled by Bernoulli random variable, and observation of rating value was determined by the Boolean OR operation. Computationally efficient variational inference algorithms were presented, where variational parameters have closed-form update rules and the computational complexity depend on the number of observed ratings, instead of the size of the rating data matrix. We also discussed implementation issues on hyperparameter tuning and estimation based on empirical Bayes. Finally, we presented experimental results demonstrating that (1) binomial mixture model is more suitable than multinomial mixture model for modellng discrete, finite, and ordered rating values; (2) our model finds meaningful solutions instead of undesirable boundary solutions, if hyperparameters are estimated by empirical Bayes; (3) our model can capture different rating trend between domain(e.g. songs and movies).

Acknowledgements

This work was supported by the IT R&D Program of MSIP/IITP (14-824-09-014, Machine Learning Center), National Research Foundation (NRF) of Korea (NRF-2013R1A2A2A01067464), and Samsung Electronics Co., Ltd.

6. REFERENCES

[1] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009.

[2] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The yahoo! music dataset and KDD-Cup'11. In *Proceedings of KDD Cup and Workshop*, 2011.

[3] Y.-D. Kim and S. Choi. Variational Bayesian view of weighted trace norm regularization for matrix factorization. *IEEE Signal Processing Letters*, 20(3):261–264, 2013.

[4] Y.-D. Kim and S. Choi. Scalable variational Bayesian matrix factorization with side information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Reykjavik, Iceland, 2014.

[5] Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, San Jose, CA, 2007.

[6] G. Ling, H. Yang, M. R. Lyu, and I. King. Response aware model-based collaborative filtering. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, USA, 2012.

[7] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Willey & Sons, Inc., 1987.

[8] B. M. Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, 2004.

[9] B. M. Marlin. *Missing Data Problems in Machine Learning*. PhD thesis, University of Toronto, 2008.

[10] B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the ACM International Conference on Recommender Systems (RecSys)*, New York, New York, USA, 2009.

[11] B. M. Marlin, R. S. Zemel, S. T. Roweis, and M. Slaney. Collaborative filtering and the missing at random assumption. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Vancouver, Canada, 2007.

[12] T. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2000.

[13] U. Paquet and N. Koenigstein. One-class collaborative filtering with random graphs. In *Proceedings of the International Conference on World Wide Web (WWW)*, Rio de Janeiro, Brazil, 2013.

[14] S. Park, Y.-D. Kim, and S. Choi. Hierarchical Bayesian matrix factorization with side information. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Beijing, China, 2013.

[15] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for large scale problems with lots of missing values. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 691–698, Warsaw, Poland, 2007.

[16] R. Salakhutdinov and A. Mnih. Bayesian probablistic matrix factorization using MCMC. In *Proceedings of the International Conference on Machine Learning (ICML)*, Helsinki, Finland, 2008.

[17] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the International Conference on Machine Learning (ICML)*, Corvallis, OR, USA, 2007.

[18] R. Salakhutdinov and N.Srebro. Collaborative filtering in a non-uniform world: Learning with weighted trace norm. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23. MIT Press, 2010.

[19] H. Steck. Training and testing of recommender systems on data missing not at random. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Washington, DC, USA, 2010.

[20] J. Wu. Binomial matrix factorization for discrete collaborative filtering. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Miami, Florida, USA, 2009.

[21] J. Yoo and S. Choi. Bayesian matrix co-factorization: Variational algorithm and Cramér-Rao bound. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Athens, Greece, 2011.