Published at Uncertainty in AI (UAI) 2007

Collaborative Filtering and the Missing at Random Assumption

Benjamin M. Marlin
Yahoo! Research and
Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5

Richard S. Zemel
Department of
Computer Science
University of Toronto
Toronto, ON M5S 3H5

Sam Roweis
Department of
Computer Science
University of Toronto
Toronto, ON M5S 3H5

Malcolm Slaney
Yahoo! Research
Sunnyvale, CA 94089

Abstract

Rating prediction is an important application and a popular research topic in collaborative filtering. However, both the validity of learning algorithms, and the validity of standard testing procedures rests on the assumption that missing rating data is missing at random (MAR); this is often violated for real data. In this paper we present the results of a user survey and study, in which we collect a random sample of ratings from current users of an online radio service. In the survey, a large number of users report they believe their opinion of a song does affect whether they choose to rate that song, a violation of the MAR condition. We collected a true random sample of more than 300,000 song ratings from more than 30,000 users. An analysis of this data shows that the sample of random ratings has markedly different properties than ratings of user-selected songs. Finally, we present experimental results which show that learning an explicit model of the missing data mechanism with an informative prior can lead to a large improvement in prediction performance on the random sample of ratings.

1 Introduction

In a collaborative filtering system users assign ratings to items, and the system uses information from all users to predict previously unseen items that each user might like or find useful. The two main tasks within collaborative filtering are recommendation and rating prediction. A rating prediction method can be used to predict the rating for a given item, or as part of a recommendation method based on estimating all missing ratings, and then recommending the items with the highest predicted rating. Collaborative filtering research within the machine learning community has focused almost exclusively on developing new models and new learning procedures to improve rating prediction performance [2, 4, 5, 6, 8].

A critical assumption behind both learning methods and testing procedures is that the missing ratings are *missing at random* [7, p. 89]. One way to violate the missing at random condition in the collaborative filtering setting is for the probability of observing a rating to depend on the value of that rating. In an internet-based movie recommendation system, for example, a user may be much more likely to see movies that they think they will like, and to enter ratings for movies that they see. This would create a systematic bias towards observing ratings with higher values.

Consider how this bias in the observed data impacts learning and prediction. In a nearest neighbour method it is still possible to accurately identify the neighbours of a given user [5]. However, the prediction for a particular item is based only on the available ratings of neighbours who rated the item in question. Conditioning on the set of users who rated the item can introduce bias into the predicted rating. The presence of non-random missing data can similarly introduce a systematic bias into the learned parameters of parametric and semi-parametric models including mixture models [1], customized probabilistic models [8], and matrix factorization models [2].

It is important to note that the presence of non-random missing data introduces a complementary bias into the standard testing procedure for rating prediction experiments [1] [5] [8, p.90]. Models are usually learned on one subset of the observed data, and tested on a different subset of the observed data. If the distribution of the observed data is different from the distribution of the fully compltede data for any reason, the estimated error on the test data can be an arbitrarily poor estimate of the error on the fully completed data. Marlin, Roweis, and Zemel confirm this using experiments on synthetic data [9].

In this paper we present the results of the first study to analyze the the impact of the missing at random assumption on collaborative filtering using data collected from real users. The study is based on more than 30,000 current users of Yahoo! Music's LaunchCast radio service. We begin with a review of the theory of missing data due to Little and Rubin [7]. We analyze the data that was gathered during the study, which included a user survey, and collecting ratings for *randomly* chosen songs. We describe models for learning and prediction with non-random missing data, and introduce a new experimental protocol for rating prediction based on training using user-selected items, and testing using randomly selected items. Experimental results show that incorporating a simple, explicit model of the missing data mechanism can lead to significant improvements in test error compared to naively treating the data as missing at random.

2 Missing Data Theory

A collaborative filtering data set can be thought of as a rectangular array $\boldsymbol{x}$ where each row in the array represents a user, and each column in the array represents an item. x_{im} denotes the rating of user i for item m. Let N be the number of users in the data set, M be the number of items, and V be the number of rating values. We introduce a companion matrix of response indicators $\boldsymbol{r}$ where $r_{im} = 1$ if x_{im} is observed, and $r_{im} = 0$ if x_{im} is not observed. We denote any latent values associated with data case i by $\boldsymbol{z}_i$. The corresponding random variables are denoted with capital letters.

We adopt the factorization of the joint distribution of the data $\boldsymbol{X}$, response indicators $\boldsymbol{R}$, and latent variables $\boldsymbol{X}$ shown in equation 2.1.

$$P(\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{Z}|\mu, \theta) = P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}, \mu)P(\boldsymbol{X}, \boldsymbol{Z}|\theta) \quad (2.1)$$

We refer to $P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}, \mu)$ as the missing data model or missing data mechanism, and $P(\boldsymbol{X}, \boldsymbol{Z}|\theta)$ as the data model. The intuition behind this factorization is that a complete data case is first generated according to the data model, and the missing data model is then used to select the elements of the data matrix that will not be observed.

2.1 Classification Of Missing Data

Little and Rubin classify missing data into several types including missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [7, p. 14]. The MCAR condition is defined in equation 2.2, and the MAR condition is defined in equation 2.3. Under MCAR the response probability for an item or set of items can not depend on the data values in any way. Under the MAR condition, the data vector is divided into a missing and an observed part according to the value of $\boldsymbol{r}$ in question: $\boldsymbol{x} = [\boldsymbol{x}^{mis}, \boldsymbol{x}^{obs}]$. The intuition is that the probability of observing a particular response pattern can only depend on the elements of the data vector that are observed under that pattern [10]. In addition, both MCAR and MAR require that the parameters μ and θ be distinct, and that they have independent priors.

$$P_{mcar}(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}, \mu) = P(\boldsymbol{R}|\mu) \quad (2.2)$$

$$P_{mar}(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}, \mu) = P(\boldsymbol{R}|\boldsymbol{X}^{obs}, \mu) \quad (2.3)$$

Missing data is NMAR when the MAR condition fails to hold. The simplest reason for MAR to fail is that the probability of observing a particular element of the data vector depends on the value of that element. In the collaborative filtering case this corresponds to the idea that the probability of observing the rating for a particular item depends on the user's rating for that item. When that rating is not observed, the missing data are not missing at random.

2.2 Impact Of Missing Data

When missing data is missing at random, maximum likelihood inference based on the observed data only is unbiased. We demonstrate this result in equation 2.7. The key property of the MAR condition is that the response probabilities are independent of the missing data, allowing the complete data likelihood to be marginalized independently of the missing data model. However, when missing data is not missing at random, this important property fails to hold, and it is not possible to simplify the likelihood beyond equation 2.4 [7, p. 219]. Ignoring the missing data mechanism will clearly lead to biased parameter estimates since the incorrect likelihood function is being optimized. For non-identifiable models such as mixtures, we will use the terms "biased" and "unbiased" in a more general sense to indicate whether the parameters are optimized with respect to the correct likelihood function.

$$\mathcal{L}_{mar}(\theta|\boldsymbol{x}^{obs}, \boldsymbol{r})$$

$$= \int_{\boldsymbol{x}^{mis}} \int_{\boldsymbol{z}} P(\boldsymbol{X}, \boldsymbol{Z}|\theta)P(\boldsymbol{R}|\boldsymbol{X}, \mu)dZd\boldsymbol{X}^{mis} \quad (2.4)$$

$$= P(\boldsymbol{R}|\boldsymbol{X}^{obs}, \mu) \int_{\boldsymbol{x}^{mis}} \int_{\boldsymbol{z}} P(\boldsymbol{X}, \boldsymbol{Z}|\theta)dZd\boldsymbol{X}^{mis} \quad (2.5)$$

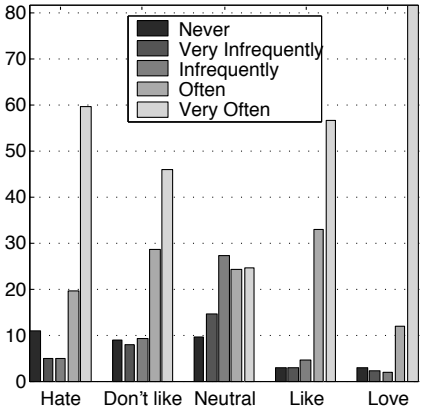
$$= P(\boldsymbol{R}|\boldsymbol{X}^{obs}, \mu)P(\boldsymbol{X}^{obs}|\theta) \quad (2.6)$$

$$\propto P(\boldsymbol{X}^{obs}|\theta) \quad (2.7)$$

Table 1: User reported frequency of rating songs as a function of preference level.

Rating Frequency	Preference Level				
	Hate	Don't Like	Neutral	Like	Love
Never	10.8%	8.8%	9.4%	3.0%	2.7%
Very Infrequently	5.0%	7.8%	14.5%	2.7%	2.1%
Infrequently	5.0%	9.1%	27.1%	4.4%	1.7%
Often	19.4%	28.5%	24.4%	33.1%	11.8%
Very Often	59.8%	45.9%	24.6%	56.8%	81.6%

Survey Results: Yahoo! LaunchCast users were asked to self report how often they thought they were likely to rate a song for which they had a given preference. The data above show the percentage of users selecting each frequency choice when asked about each preference level. Users could select only one frequency per preference were otherwise unconstrained.



From a statistical perspective, biased parameter estimates are a serious problem. From a machine learning perspective, the problem is only serious if it adversely affects the end use of a particular model. Using synthetic data experiments, Marlin, Zemel, and Roweis demonstrated that ignoring the missing data mechanism in a rating prediction setting can have a significant impact on predictive performance [9].

3 Yahoo! LaunchCast Case Study

To properly assess the impact of the missing at random assumption on rating prediction, we require a test set consisting of ratings that are a true random sample of the ratings contained in the complete data matrix. In this section we describe a study conducted in conjunction with Yahoo! Music's LaunchCast Radio service to collect such a data set.

LaunchCast radio is a customizable streaming music service where users can influence the music played on their personal station by supplying ratings for songs. The LaunchCast Radio player interface allows the user enter a rating for the currently playing song using a five star scale.[1]

Data was collected from LaunchCast Radio users between August 22, 2006 and September 12, 2006. Users based in the US were able to join the study by clicking on a link in the LaunchCast player. Both the survey and rating data were collected though the study's web site. A total of 35,786 users contributed useable data to the study.

[1]The Yahoo! Music LaunchCast web site is available at http://music.yahoo.com/launchcast/.

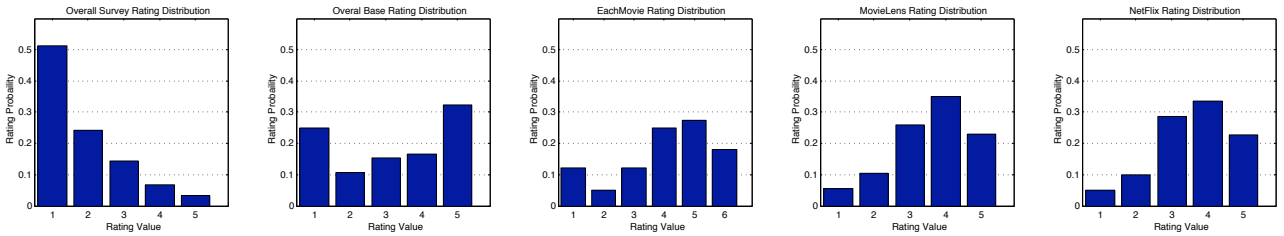
3.1 User Survey

The first part of the study consisted of a user survey containing of sixteen multiple choice questions. The questions relevant to this work asked users to report on how their preferences affect which songs they choose to rate. The question was broken down by asking users to estimate how often they rate a song given the degree to which they like it. The results are summarized in table 1, and represented graphically in the accompanying figure. Each column in the table gives the results for a single survey question. For example, the column labeled "neutral" corresponds to the question "If I hear a song I feel neutral about I choose to rate it:" with the possible answers being "never", "very infrequently", "infrequently", "often", and "very often".

The results indicate that the choice to rate a song depends quite strongly the a user's opinion of that song. Most users tend to rate songs that they love much more often than songs they feel neutral about, and somewhat more often than songs that they hate. Users were next asked if they thought their preferences for a song *do not* affect whether they choose to rate it. 64.4% of users responded that their preferences *do* affect their choice to rate a song. This dependence indicates a violation of the missing at random assumption.

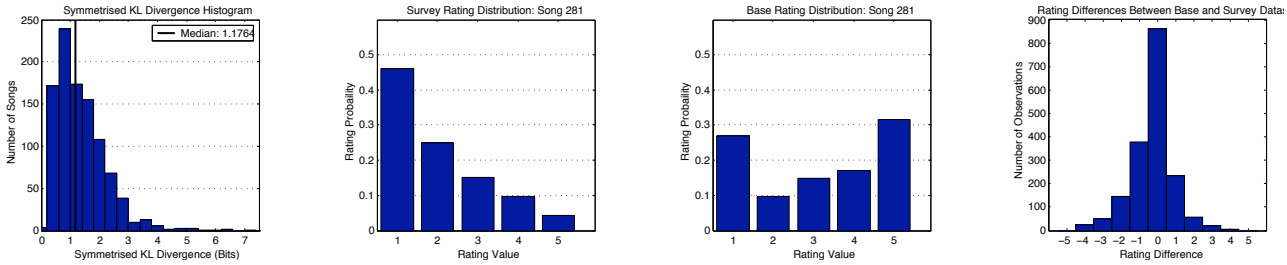
3.2 Rating Data Collection

Following the survey, users were presented with a set of ten songs to rate. The artist name and song title was given for each song, along with a thirty second audio clip. If users were not familiar with the song, they had the option of playing the clip before entering a rating. Ratings were entered on the standard five point scale used by Yahoo! Music. The set of ten songs presented to each user was chosen at random without replacement from a larger set of 1000 songs that we



(a) Yahoo! Survey Rating Distribution (b) Yahoo! Base Rating Distribution (c) EachMovie Rating Distribution (d) MovieLens Rating Distribution (e) NetFlix Rating Distribution

Figure 1: Distribution of rating values in the Yahoo! base set and survey set compared to several popular collaborative filtering data sets including EachMovie, MovieLens, and Netflix.



(a) Histogram of number of songs vs symmetrised KL divergence. The median value is 1.1764.

(b) Survey marginal distribution for song 281 with symmetrised KL divergence 1.1711.

(c) Base marginal distribution for song 281 with symmetrised KL divergence 1.1711.

(d) Histogram of rating difference between items rated in both survey and in base sets.

Figure 2: Panels (a) to (c) give an indication of the distribution of differences between base and survey marginal distributions for each song. Panel (d) shows the distribution of differences between items rated both in the survey set and the base set.

will refer to as the "survey songs". The survey songs were chosen at random from the songs in the LaunchCast playlist having at least 500 existing ratings in the LaunchCast rating database.

We will refer to the survey song ratings collected during the survey as the "survey set." We also extracted the complete set of ratings for the survey songs from the LaunchCast rating database. We will refer to this set of ratings as the "base set."

Figures 1(a) and 1(b) show the empirical distribution of ratings in the survey set and the base set. These figures show a dramatic difference between the two distributions. The number of five star rating values is many times lower in the survey set than the base set, and the two distributions exhibit opposite trends on rating values two to five. Figures 1(c) to 1(e) give the rating distributions for several other collaborative filtering data sets including EachMovie, MovieLens, and Netflix. All show a skew toward high rating values.

To further analyze the difference between the base set and the survey set, we computed the distribution over ratings for each item. For a particular item m let $P^S(X_m = v)$ be the empirical probability of rating value v in the survey set, and $P^B(X_m = v)$ be the empirical probability of rating value v in the base set. We smooth the empirical probabilities by one count per rating value to avoid zeros. We use the symmetrised Kullback−Leibler divergence (SKL) shown in equation 3.8 to measure the difference between the $P^S(X_m = v)$ and $P^B(X_m = v)$ distributions for each item m.

$$SKL_m = \sum_{v=1}^{V} P^S(X_m = v) \log \left(\frac{P^S(X_m = v)}{P^B(X_m = v)} \right) + P^B(X_m = v) \log \left(\frac{P^B(X_m = v)}{P^S(X_m = v)} \right) \quad (3.8)$$

Figure 2(a) shows a histogram of the symmetrised Kullback−Leibler divergence values. The thick vertical line in the plot indicates the median SKL value of 1.1764 bits. Song 281 has an SKL value of 1.1711 bits, the largest SKL value less than the median. Figures 2(b) and 2(c) illustrate the marginal rating distributions for song 281. These distributions are qualitatively quite different, and half of the songs in the survey set exhibit a more extreme difference according to the SKL measure.

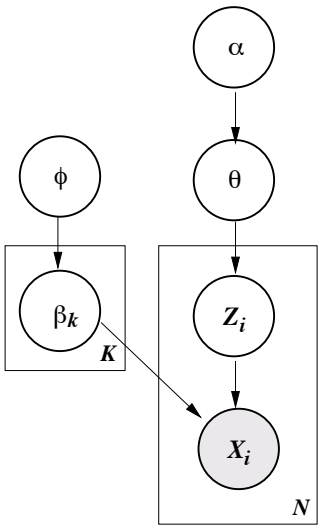


Figure 3: Bayesian multinomial mixture model.

A pertinent question is whether users were truthfully reporting ratings during the survey? To help answer this question we extracted the set of ratings that were observed both in the survey set, and in the base set. Figure 2(d) shows a histogram of the differences $x^B_{im} - x^S_{im}$ where the user-song pair (i, m) is observed in both the survey S and base sets B. We can see from figure 2(d) that the agreement between the two sets of ratings is quite good.

It is important to note that the observed discrepancy between the survey set marginal distributions and the base set marginal distributions is not conclusive evidence that the missing data in the base set is NMAR. This is due to the fact that the MAR assumption can hold for the true underlying data model, but not for more simplistic models like the marginal model used in the present analysis. Nevertheless, we believe that the results of the present analysis combined with the results of the user survey provide compelling evidence against the MAR assumption.

4 Modeling Non-Random Missing Data

Many probabilistic models have the property that missing data can be analytically integrated away under the missing at random assumption. This allows for computationally efficient, unbiased parameter estimation. The multinomial mixture model has this convenient property, and it has been well studied in the collaborative filtering domain [8].

When the missing at random assumption is not believed to hold, equation 2.4 shows that parameter estimation will be biased unless the true missing data mechanism is known. In a domain as complex and high dimensional as collaborative filtering, a more realistic goal is to formulate models of the missing data mechanism that capture some of its key properties.

Algorithm 1 MAP EM Algorithm for the Bayesian multinomial mixture model.

E-Step:

$$q_{zi} \leftarrow \frac{\theta_z \prod_{m=1}^{M} \prod_{v=1}^{V} \beta_{vmz}^{r_{im}[x_{im}=v]}}{\sum_{z=1}^{K} \theta_z \prod_{m=1}^{M} \prod_{v=1}^{V} \beta_{vmz}^{r_{im}[x_{im}=v]}}$$

M-Step:

$$\theta_z \leftarrow \frac{\alpha_z - 1 + \sum_{i=1}^{N} q_{zi}}{\sum_{z=1}^{K} (\alpha_z + \sum_{i=1}^{N} q_{zi}) - K}$$

$$\beta_{vmz} \leftarrow \frac{\phi_{vmk} - 1 + \sum_{i=1}^{N} q_{zi} r_{im}[x_{im}=v]}{\sum_{v=1}^{V} \phi_{vmk} - V + \sum_{i=1}^{N} q_{zi} r_{im}}$$

In this section we present the basic multinomial mixture model, and give learning and prediction methods under the MAR assumption. We extend the mixture model by combining it with a Bayesian variant of the *CPT-v* missing data model [9], which captures a key property of the non-random missing data mechanism implied by the user survey results. We give learning and prediction methods for the combined mixture/*CPT-v* model.

4.1 Multinomial Mixture Data Model

The multinomial mixture model is a generative probabilistic model. It captures the simple intuition that users form groups or clusters according to their preferences for items. We give a graphical depiction of the finite Bayesian mixture model in figure 3, and summarize the probabilistic model below.

$$P(\theta, \beta | \alpha, \phi) = D(\theta|\alpha) \prod_k \prod_m D(\beta_{mk}|\phi_{mk}) \tag{4.9}$$

$$P(Z_i = k|\theta) = \theta_k \tag{4.10}$$

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i | Z_i = k, \beta) = \prod_m \prod_v \beta_{vmk}^{[x_{im}=v]} \tag{4.11}$$

The main feature of the model is the variable Z_i, which indicates which of the K groups or clusters user i belongs to. To generate a complete data vector $\boldsymbol{X}_i$ for user i, a value k for Z_i is first sampled according to the discrete distribution $P(Z_i = k|\theta)$. A rating value v for each item m is then sampled independently from the discrete distribution $P(X_{im} = v|Z_i = k, \boldsymbol{\beta}_{mk})$. Importantly, all we observe is the final data vector $\boldsymbol{X}_i$. Z_i is considered a latent variable since its value is never observed.

In a Bayesian mixture model, the parameters θ and $\boldsymbol{\beta}_{mk}$ are also regarded as random variables. Before generating any data cases, the model parameters are

first sampled from their prior distributions. The same model parameters are assumed to generate all data cases. In the present case we choose conjugate Dirichlet priors for both θ, and $\boldsymbol{\beta}_{mk}$. We give the form of the Dirichlet priors for θ and $\boldsymbol{\beta}_{mk}$ in equations 4.12 and 4.13.

$$D(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \tag{4.12}$$

$$D(\beta_{mk}|\phi_{mk}) = \frac{\Gamma(\sum_{v=1}^{V}\phi_{vmk})}{\prod_{v=1}^{V}\Gamma(\phi_{vmk})}\prod_{v=1}^{V}\beta_{vmk}^{\phi_{vmk}-1} \tag{4.13}$$

The posterior log probability of the mixture model parameters θ and $\boldsymbol{\beta}_{mk}$ given a sample of incomplete data is shown below in equation 4.14.

$$\mathcal{L}_{mar} = \sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\theta_k\prod_{m=1}^{M}\prod_{v=1}^{V}\beta_{vmk}^{r_{im}[x_{im}=v]}\right) + \log D(\theta|\alpha) + \sum_{m=1}^{M}\sum_{k=1}^{K}\log D(\beta_{mk}|\phi_{mk}) \tag{4.14}$$

The Bayesian mixture model parameters are learned from incomplete data by maximizing the posterior log probability of the observed data. This optimization is efficiently performed using the Expectation Maximization (EM) algorithm of Dempster, Laird, and Rubin [3]. We give the maximum a posteriori (MAP) EM algorithm for the Bayesian multinomial mixture model in Algorithm 1. In the expectation step of the algorithm we compute posterior distribution on Z_i for each user i given the current values of the model parameters. This inference procedure is also important for prediction. We give it in equation 4.15.

$$P(Z_i = z|\boldsymbol{x}_i, \boldsymbol{r_i}, \theta, \beta) = \frac{\theta_z\prod_{m=1}^{M}\prod_{v=1}^{V}\beta_{vmz}^{r_{im}[x_{im}=v]}}{\sum_{z=1}^{K}\theta_z\prod_{m=1}^{M}\prod_{v=1}^{V}\beta_{vmz}^{r_{im}[x_{im}=v]}} \tag{4.15}$$

4.2 The *CPT-v* Missing Data Model

The *CPT-v* missing data model was proposed by Marlin, Roweis, and Zemel as the simplest non-random missing data model [9]. The *CPT-v* model captures the intuition that a user's preference for an item affects whether they choose to rate that item or not. The model assumes that the choice to rate each item is independent, and that the probability of rating a single item, given that the user's rating for that item is v, is Bernoulli distributed with parameter μ_v. We extend the basic *CPT-v* model slightly by introducing a Beta prior on the parameters μ_v. The probabilistic model is summarized below.

Algorithm 2 MAP EM Algorithm for the Bayesian multinomial mixture/*CPT-v* model.

E-Step:

$\lambda_{vmzn} \leftarrow ([x_{im} = v]\mu_v\beta_{vmz})^{r_{im}}((1-\mu_v)\beta_{vmz})^{1-r_{im}}$

$\gamma_{mzn} \leftarrow \sum_{v=1}^{V}\lambda_{vmzn}$

$q_{zi} \leftarrow \frac{\theta_{z_n}\prod_{m=1}^{M}\gamma_{mzn}}{\sum_{z=1}^{K}\theta_{z'}\prod_{m=1}^{M}\gamma_{mzn}}$

M-Step:

$\theta_z \leftarrow \frac{\alpha_z - 1 + \sum_{i=1}^{N} q_{zi}}{\sum_{z=1}^{K}(\alpha_z + \sum_{i=1}^{N} q_{zi}) - K}$

$\beta_{vmz} \leftarrow \frac{\phi_{vmk} - 1 + \sum_{i=1}^{N}\phi_{zi}\lambda_{vmzn}/\gamma_{mzn}}{\sum_{v=1}^{V}\phi_{vmk} - V + \sum_{n=1}^{N} q_{zi}}$

$\mu_v \leftarrow \frac{\xi_{1v} - 1 + \sum_{i=1}^{N}\sum_{z=1}^{K} q_{zi}\sum_{m=1}^{M} r_{mn}\lambda_{vmzn}/\gamma_{mzn}}{\xi_{0v} + \xi_{1v} - 2 + \sum_{n=1}^{N}\sum_{z=1}^{K} q_{zi}\sum_{m=1}^{M}\lambda_{vmzn}/\gamma_{mzn}}$

$$P(\mu|\xi) = \prod_{v} Beta(\mu_v|\xi_v) \tag{4.16}$$

$$P(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{X} = \boldsymbol{x}) = \prod_{m=1}^{M}\prod_{v=1}^{V}\mu_v^{r_{im}[x_{im}=v]}(1-\mu_v)^{(1-r_{im})[x_{im}=v]} \tag{4.17}$$

The Beta prior we select is the conjugate prior for the Bernoulli parameters μ_v. We give the form of the prior distribution in equation 4.18.

$$Beta(\mu_v|\xi_v) = \frac{\Gamma(\xi_{0v}+\xi_{1v})}{\Gamma(\xi_{0v})\Gamma(\xi_{1v})}\mu_v^{\xi_{1v}-1}(1-\mu_v)^{\xi_{0v}-1} \tag{4.18}$$

The factorized structure of the model is quite restrictive. However, it allows the missing data to be summed out of the posterior distribution leaving local factors that only depend on one missing data value at a time. The log posterior distribution on the model parameters is given in Equation 4.19.

$$\mathcal{L}_{CPTv} = \sum_{n=1}^{N}\log\left(\sum_{z=1}^{K}\theta_z\prod_{m=1}^{M}\gamma_{mzn}\right) + \sum_{v=1}^{V}\log Beta(\mu_v|\xi_v) \tag{4.19}$$

$$\gamma_{mzn} = \begin{cases} \prod_v(\mu_v\beta_{vmz})^{[x_{im}=v]} & \ldots\ r_{im}=1 \\ \sum_v(1-\mu_v)\beta_{vmz} & \ldots\ r_{im}=0 \end{cases}$$

As in the Bayesian mixture model case, the log posterior distribution of the combined Bayesian

mixture/*CPT-v* model can be maximized using an expectation maximization algorithm. We give the details in Algorithm 2. Again, inference for the latent mixture indicator Z_i is the main operation in the expectation step. As we can see in equation 4.20, the form of the inference equation is very similar to the standard mixture case.

$$P(Z_i = z|\boldsymbol{x}_i, \boldsymbol{r_i}, \theta, \beta) = \frac{\theta_z \prod_{m=1}^{M} \gamma_{mzn}}{\sum_{z=1}^{K} \theta_z \prod_{m=1}^{M} \gamma_{mzn}} \quad (4.20)$$

4.3 Rating Prediction

To make a prediction for user i and item m we first need to perform inference in the model to compute the posterior distribution $P(Z_i = z|\boldsymbol{x}_i, \boldsymbol{r_i}, \theta, \beta)$ over the mixture indicator variable Z_i. For the multinomial mixture model under the MAR assumption we use equation 4.15. For the multinomial mixture model combined with the *CPT-v* model we use equation 4.20. For both models, we compute the predictive distribution over rating values for item m according to equation 4.21.

$$P(X_{im} = v) = \sum_{z=1}^{K} \beta_{vmz} P(Z_i = z|\boldsymbol{x}_i, \boldsymbol{r_i}, \theta, \beta) \quad (4.21)$$

5 Experimental Method and Results

Both the analysis of the user survey, and the analysis of the rating data collected in this study suggest that missing data in the LaunchCast database is not missing at random. The question we address in this section is whether treating the missing data as if it were not missing at random leads to an improvement in predictive performance relative to treating the missing data as if it were missing at random. We discuss the data set used for rating prediction experiments, the methods tested, the experimental protocol, and the results.

5.1 Rating Prediction Data Set

The prediction data set consists of the 1000 songs in the survey set, as well as an additional 1000 songs chosen at random. Survey users were included in the prediction data set if they had at least 30 existing ratings on the additional set of items. Non-survey users were included in the prediction data set if they had at least 30 existing ratings on both sets of songs. The training data set consists of the ratings from the LaunchCast rating database for each user included the prediction data set. The test data set consists of the ratings for random songs collected during the study for each survey user in the prediction data set. The overlapping ratings in the two sets were removed from the training set. The minimum number of ratings per user, and the inclusion of non-survey users and items was necessary to create a reasonably dense, but manageably sized data set.

5.2 Rating Prediction Experiments

The baseline method for the rating prediction experiments is based on the Bayesian multinomial mixture model under the MAR assumption. We learn the model using the EM algorithm given in Algorithm 1 with the prior parameters $\phi_{vmz} = 2$ and $\alpha_z = 2$ for all v, m, z.

Initial testing of the *CPT-v* model showed that it diverged to poor boundary solutions using a uniform prior. This is not surprising since *CPT-v* has previously been observed to diverge on real data sets [9]. To remedy this problem we tried learning only the mixture model parameters with the *CPT-v* model parameters fixed to the values $\mu = [0.01, 0.01, 0.01, 0.01, 0.05]$ to express the belief that five star ratings are much more likely to be observed than the other rating values. We tried relaxing this assumption by setting the prior on μ_v to $Beta(10, 990)$ for $v = 1, ..., 4$, and $Beta(50, 950)$ for $v = 5$, and learning both μ, and the data model parameters. This prior expresses the same belief that five star ratings are more likely to be observed than the other rating values. It is worth noting that this very simple prior is the first prior we tested, and we have not yet investigated the effect of alternative priors.

Training for each model was performed until the log posterior converged to six decimal places, or 250 EM iterations were exceeded. Once each model was trained, it was used to predict the rating values of the ten randomly selected test items for each survey user in the test data set. We report error values using mean absolute error (MAE) [1].

5.3 Rating Prediction Results

Each row of table 2 gives prediction results for a different combination of mixture model, and missing data model. The missing data models are "none", corresponding to the MAR assumption; "*CPT-v* Fixed", corresponding to setting μ by hand; and "*CPT-v* Prior", corresponding to learning μ under an informative prior. The results show that the *CPT-v* selection model combined with the informative prior achieves best performance on the test set. This combination significantly out-performs prediction under the MAR assumption. It is interesting to note that the best performance is achieved at $K = 1$. The learned μ parameters for $K = 1$ essentially explain all of the missing data as having a rating value of one star.

Table 2: Mean Absolute Rating Prediction Error

Model	Train MAE	Test MAE
K=1/None	1.3354 ± 0.0000	1.2393 ± 0.0000
K=1/CPT-v Fixed	1.3554 ± 0.0000	1.0208 ± 0.0000
K=1/CPT-v Prior	1.7269 ± 0.0000	0.8468 ± 0.0000
K=2/None	0.9447 ± 0.0009	1.2344 ± 0.0007
K=2/CPT-v Fixed	1.0341 ± 0.0636	0.9715 ± 0.0073
K=2/CPT-v Prior	1.6326 ± 0.0192	0.8500 ± 0.0009
K=6/None	0.7546 ± 0.0012	1.1330 ± 0.0057
K=6/CPT-v Fixed	0.7781 ± 0.0016	0.9923 ± 0.0055
K=6/CPT-v Prior	1.2027 ± 0.0056	0.9366 ± 0.0085
K=10/None	0.7349 ± 0.0017	1.1230 ± 0.0061

6 Discussion and Conclusions

In the collaborative filtering domain, both the validity of learning algorithms, and the validity of standard testing procedures rests on the assumption that missing rating data is missing at random. In this paper we have provided compelling evidence of a violation of the missing at random assumption in real collaborative filtering data. Furthermore, we have shown that learning an explicit model of the missing data mechanism can significantly improve rating prediction on a test set.

Results of the LaunchCast user survey indicate that users are aware that their preferences impact which items they choose to rate. Ratings of randomly selected songs collected in this study show systematic differences relative to ratings of user selected songs. We introduced a new experimental protocol where models are trained on ratings of user selected songs, and tested on ratings of randomly selected songs. Using this protocol, we found that a very simple missing data model, with an informative yet not highly tuned prior, produced a surprising boost in test performance.

There remain many open questions for future work. We have yet to address the question of sensitivity to the prior, or the question of model selection. The use of approximate Bayesian inference may lead to better predictive performance than the current MAP framework. It would also allow us to consider more flexible data and missing data models including hierarchical, and non-parametric constructions. In terms of empirical development, it would be interesting to study the precision/recall of five star ratings to see if they are better predicted by more complex data models.

From a broader perspective, using rating prediction to solve the recommendation problem is much less attractive without the foundation provided by the MAR assumption. Recommendation methods that can avoid solving the rating prediction problem may be much more robust to deviations from the MAR assumption.

References

[1] J. S. Breese, D. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52, July 1998.

[2] D. DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd International Conference on Machine learning*, pages 249–256, 2006.

[3] A. Dempster, N. Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval Journal*, 4(2):133–151, July 2001.

[5] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 230–237, 1999.

[6] T. Hofmann and J. Puzicha. Latent Class Models for Collaborative Filtering. In *Proceedings of the International Joint Conference in Artificial Intelligence*, 1999.

[7] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., 1987.

[8] B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, January 2004.

[9] B. Marlin, R. S. Zemel, and S. T. Roweis. Unsupervised learning with non-ignorable missing data. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[10] D. B. Rubin. Inference and missing data. *Biometrika*, 64(3):581–592, 1976.