A BAYESIAN MODEL FOR COLLABORATIVE FILTERING

Yung-Hsin Chen
Department of MSIS, CBA 5.202
University of Texas at Austin
Austin, TX 78712
yhchien@mail.utexas.edu

Edward I. George
Department of MSIS, CBA 5.202
University of Texas at Austin
Austin, TX 78712
egeorge@mail.utexas.edu

Abstract

Consider the general setup where a set of *items* have been partially *rated* by a set of *judges*, in the sense that not every item has been rated by every judge. For this setup, we propose a Bayesian approach for the problem of predicting the missing ratings from the observed ratings. This approach incorporates similarity by assuming the set of judges can be partitioned into groups which share the same ratings probability distribution. This leads to a predictive distribution of missing ratings based on the posterior distribution of the groupings and associated ratings probabilities. Markov chain Monte Carlo methods and a hybrid search algorithm are then used to obtain predictions of the missing ratings.

1 INTRODUCTION

Consider the general setup where a set of *items* have been partially *rated* by a set of *judges*, in the sense that not every item has been rated by every judge. For this setup, we consider the problem of predicting the missing ratings from the observed ratings, the so-called collaborative filtering problem.

As a simple illustration of the problem, suppose a web site maintains a set of 6 news articles (the items), some which have been read by 5 Internet users (the judges) and rated as 1 (boring) through 4 (interesting). Suppose the ratings in Table 1 occurred, ("-" denotes a missing rating). Prediction of the missing ratings from this information would help the users select unread articles. For example, user 1 might like to know which of articles 4 or 5 would be more interesting. Because the preferences of user 1 seem more similar to those of user 2, it seems likely that user 1 would also prefer article 4 to article 5.

Table 1: Sample Web Site Ratings

	Articles					
Users	1	2	3	4	5	6
1	2	1	2	-	-	4
2	1	-	1	4	3	-
3	3	-	3	4	2	1
4	-	4	-	3	-	1
5	4	4	3	-	1	2

Shardanand and Maes (1995) (hereafter SM) proposed a collaborative filtering method which predicts missing ratings using linear combinations of observed ratings. Their key idea is to use formal similarity measures to put more weight on judges with similar preferences. The SM predictor is of the form

$$\hat{R}_{ij} = \frac{\sum_{j' \neq j} w_{jj'} R_{ij'}}{\sum_{j' \neq j} w_{jj'}} \tag{1}$$

where R_{ij} is the rating of item i by judge j, and $w_{jj'}$ is a measure of preference similarity between judge j and judge j'. SM propose various similarity measures to determine the weights $w_{jj'}$, and recommend the mean the constrained correlation

$$w_{jj'} = \frac{\sum_i (R_{ij} - Neutral)(R_{ij'} - Neutral)}{\sqrt{\sum_i (R_{ij} - Neutral)^2 \sum_i (R_{ij'} - Neutral)^2}} \tag{2}$$

where *Neutral* stands for a neutral rating, usually the midpoint of the rating scale. All summations above are over items which have been simultaneously rated by judges j and j'. The constrained correlation is motivated by observing that ordinary correlation may not capture similarity, since shifting ratings by a constant leaves correlation unchanged. SM also recommend setting $w_{jj'} = 0$ whenever the measure is below a preset threshold, effectively discarding dissimilar judges from (1). Paul, R., Neophytos, I., Mitesh, S., Peter, B. and John, R. (1994) proposed a related collaborative filtering algorithm based on ordinary correlation which incorporates both similar and dissimilar judge preferences for making predictions.

A limitation of the SM approach occurs when pairs of judges have simultaneously rated very few items, as might occur with sparse data. This can lead to high similarity scores based on little information. For instance, in the web example, suppose the ratings from 3 users in Table 2 were observed. To predict the user 2 rating of article 4, the SM measures would weight user 3 more heavily than user 1. This seems inappropriate since users 1 and 2 are very similar, and there is hardly any evidence of similarity between users 2 and 3. We consider the effect of patterned missingness again in Section 5.

Table 2: A Pattern of Missingness

	Articles				
Users	1	2	3	4	5
1	1	1	3	5	5
2	1	1	2	-	5
3	1	-	-	2	-

In this paper, we propose a Bayesian collaborative filtering solution. Rather than use a similarity measure such as mean square error or constrained correlation, our approach incorporates similarity by assuming the set of judges can be partitioned into groups which share the same ratings probability distribution. This leads to a predictive distribution of missing ratings based on the posterior distribution of the groupings and associated ratings probabilities. Markov chain Monte Carlo methods and a hybrid search algorithm are then used to obtain predictions of the missing ratings. A related Bayesian approach has been recently proposed by Breese, J. S. , Heckerman, D. and Kadie, C. (1998). Also related to our approach is Consonni and Veronese (1995), which proposes a Bayesian approach for partitioning observations from binomial experiments into homogeneous groups.

2 A BAYESIAN MODEL

In this section, we describe our proposed Bayesian collaborative filtering model. The underlying premise of our model is that judges who tend to give similar ratings can be modeled as having identical rating probabilities. Thus, our model entails a partition of the judges into subgroups which are identified by common probability structure for the ratings. Although such a model is not universally applicable, it may often be a reasonable approximation for popular preference groupings of movies, television programs, restaurants, politics, etc.

Our model is precisely formulated as follows. For any particular set of items and judges, we denote *items* by $i = 1, \ldots, I$, *ratings* by $r = 1, \ldots, R$, *judges* by $j = 1, \ldots, J$, and *groups* of judges by $g = 1, \ldots, G$. Clearly, I, R, J are known, and the assignment of judges to groups is unknown. For the moment, we will treat G as known. The two key building blocks of our model are:

- Each judge is assigned to group g with the same probability q_g, and judges are assigned independently. Let $\mathbf{q} = (\mathbf{q_1}, \ldots, \mathbf{q_G})$ be all the grouping probabilities.

- The rating of item i by any judge in group g is simple multinomial with parameter $\mathbf{p_{ig}} = (\mathbf{p_{ig1}}, \ldots, \mathbf{p_{igR}})$, (i.e. rating r is assigned with probability p_{igr}). Ratings are assigned independently across all items and judges. Let $\mathbf{p_g} = (\mathbf{p_{1g}}, \ldots, \mathbf{p_{Ig}})$ be the group g rating probabilities, and let $\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_G})$ be all the rating probabilities.

To represent the data, it will be convenient to code the rating assignments by letting $x_{ijr} = 0$ or 1 according to whether item i was rated r by judge j. If item i was not rated by judge j, we set $x_{ijr} = 0,\ r = 1, \ldots, R$. The rating of item i by judge j is thus given by $\mathbf{x_{ij}} = (\mathbf{x_{ij1}}, \ldots, \mathbf{x_{ijR}})$. For example, suppose item 1 was rated 2 by judge 1, then $\mathbf{x_{11}} = (\mathbf{0}, \mathbf{1})$. The ratings of all items by judge j is summarized by $\mathbf{x_j} = (\mathbf{x_{1j}}, \ldots, \mathbf{x_{Ij}})$, and the ratings across all items and judges is summarized by $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_J})$.

Finally, to simplify likelihood and posterior calculations, it will be convenient to introduce latent variables for group membership. More precisely, let $z_{jg} = 1$ if judge j belongs to group g, and 0 otherwise. The group membership of judge j is then identified by $\mathbf{z_j} = (\mathbf{z_{j1}}, \ldots, \mathbf{z_{jG}})$. For example, if there are 5 groups, and judge 2 belongs to group 3, then $\mathbf{z_2} = (\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0})$. Note that the probability that $z_{jg} = 1$ is given by q_g above.

Using this notation, the probability of ratings $\mathbf{x_j}$ by judge j, given $\mathbf{p}$ and $\mathbf{q}$, is a mixture of a product of simple multinomial probabilities, namely

$$f(\mathbf{x_j} \mid \mathbf{p}, \mathbf{q}) = \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{q_g f(x_j \mid z_{jg} = 1, p_g)} \qquad (3)$$

where

$$f(\mathbf{x_j} \mid \mathbf{z_{jg}} = \mathbf{1}, \mathbf{p_g}) = \prod_{\mathbf{i=1}}^{\mathbf{I}} \prod_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{p_{igr}^{x_{ijr}}} \qquad (4)$$

Hence the probability of ratings $\mathbf{x}$ by all judges is given by

$$f(\mathbf{x} \mid \mathbf{p}, \mathbf{q}) = \prod_{\mathbf{j=1}}^{\mathbf{J}} \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{q_g f(x_j \mid z_{jg} = 1, p_g)} \tag{5}$$

Note that by setting $x_{ijr} = 0,\ r = 1, \ldots, R$ for when item i is not rated by judge j, we effectively remove the inclusion of missing values in the likelihood of $\mathbf{p}, \mathbf{q}$.

To complete the Bayesian model setup, we consider priors for the parameters $\mathbf{p}$ and $\mathbf{q}$. A natural and computationally convenient choice are the Dirichlet priors which are conjugate for our setup. In particular, we consider priors for which the rating probabilities $\mathbf{p_{ig}}$'s and the grouping probabilities $\mathbf{q} = (\mathbf{q_1}, \ldots, \mathbf{q_G})$ are all independent with symmetric Dirichlet distributions

$$\pi(\mathbf{p_{ig}}) \propto \prod_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{p_{igr}^{\alpha-1}},\ \mathbf{-1} < \alpha \leq \mathbf{1}$$

and

$$\pi(\mathbf{q}) \propto \prod_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{q_g^{\beta-1}},\ \mathbf{-1} < \beta \leq \mathbf{1}$$

respectively. Furthermore, we consider only small hyperparameter values $-1 < \alpha, \beta \leq 1$ to keep the priors relatively noninfluential. Although the Jeffrey's priors, obtained with $\alpha = \beta = 1/2$, are a natural noninformative choice, we have obtained essentially the same results using the uniform priors, obtained with $\alpha = \beta = 1$. The symmetric and noninfluential aspects of these prior choices are motivated for situations where there is little or no prior information, as is apt to be the case. Of course, it may be preferable to use asymmetric, influential Dirichlet priors when there is prior information, such as when certain items are expected to get consistently high or consistently low ratings.

3 PREDICTION FOR FIXED G

Suppose judge ℓ did not rate item k so that $x_{k\ell r} = 0,\ r = 1, \ldots, R$. The ultimate goal of collaborative filtering is to predict the future value of this rating. Let us denote this by $\mathbf{y_{k\ell}} = (\mathbf{y_{k\ell 1}}, \ldots, \mathbf{y_{k\ell R}})$ where $y_{k\ell r} = 1$ if judge ℓ would (in the future) assign rating r to item 1, and 0 otherwise. The Bayesian predictive distribution of $\mathbf{y_{k\ell}}$ is given by

$$\begin{aligned} f(\mathbf{y_{k\ell}} \mid \mathbf{x}) \quad = \quad & \int_{\mathbf{p},\mathbf{q}} f(\mathbf{y_{k\ell}} \mid \mathbf{p}, \mathbf{q})\ \mathbf{f(x \mid p, q)} \\ & \qquad \pi(\mathbf{p})\ \pi(\mathbf{q})\ \mathbf{dp\ dq} \end{aligned} \tag{6}$$

where the conditional distribution of $\mathbf{y}$ is

$$f(\mathbf{y_{k\ell}} \mid \mathbf{p}, \mathbf{q}) = \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{q_g} \prod_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{p_{kgr}^{y_{k\ell r}}} \tag{7}$$

and the remaining distributions are provided in the previous section.

Direct evaluation of the integral in (6) is complicated by the summations which appear in (5) and (7). To circumvent this problem, we condition on $\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_J})$, the latent variables for group membership defined in the previous section. We can reexpress (6) as

$$f(\mathbf{y_{k\ell}} \mid \mathbf{x}) = \sum_{\mathbf{z}} \mathbf{f(y_{k\ell} \mid z, x) f(z \mid x)}. \tag{8}$$

When $z_{kg} = 1$, the first expression on the RHS of (8) is

$$\begin{aligned} f(\mathbf{y_{k\ell}} \mid \mathbf{z}, \mathbf{x}) \quad & \propto \quad \int_{\mathbf{p}} f(\mathbf{y_{k\ell}} \mid \mathbf{p}, \mathbf{z})\ \mathbf{f(x \mid p, z)}\ \pi(\mathbf{p})\ \mathbf{dp} \\ & \propto \quad \int_{\mathbf{p}} \prod_{r=1}^{R} p_{kgr}^{\alpha - 1 + y_{k\ell r} + \sum_{j=1}^{J} z_{jg} x_{kjr}}\ d\mathbf{p} \\ & \propto \quad \prod_{r=1}^{R} \Gamma(\alpha + y_{k\ell r} + \sum_{j=1}^{J} z_{jg} x_{kjr}). \end{aligned}$$

Thus, when $z_{kg} = 1$, $f(\mathbf{y_{k\ell}} \mid \mathbf{z}, \mathbf{x})$ is given by

$$P(y_{k\ell r} = 1 \mid \mathbf{z}, \mathbf{x}) = \frac{\alpha + \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{z_{jg} x_{kjr}}}{\mathbf{R}\alpha + \sum_{\mathbf{r=1}}^{\mathbf{R}} \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{z_{jg} x_{kjr}}} \tag{9}$$

which is easy to compute. Note that $\sum_{j=1}^{J} z_{jg}$ is the number of judges in group g and $\sum_{j=1}^{J} z_{jg} x_{kjr}$ is the number of judges in group g who gave rating r to item k. Thus, this probability reflects our assumption that judges in the same group have similar rating probabilities.

The second expression on the RHS of (8) is

$$\begin{aligned} f(\mathbf{z} \mid \mathbf{x}) \quad & \propto \quad \int_{\mathbf{p},\mathbf{q}} f(\mathbf{x} \mid \mathbf{p}, \mathbf{z})\ \mathbf{f(z \mid q)}\ \pi(\mathbf{p})\ \pi(\mathbf{q})\ \mathbf{dp\ dq} \\ & \propto \quad \int_{\mathbf{p},\mathbf{q}} \prod_{g=1}^{G} q_g^{\beta - 1 + \sum_{j=1}^{J} z_{jg}} \\ & \qquad \prod_{i=1}^{I} \prod_{r=1}^{R} p_{igr}^{\alpha - 1 + \sum_{j=1}^{J} z_{jg} x_{ijr}}\ d\mathbf{p}\ \mathbf{dq} \\ & \propto \quad \frac{\prod_{g=1}^{G} \Gamma(\beta + \sum_{j=1}^{J} z_{jg})}{\Gamma(G\beta + J)} \\ & \qquad \prod_{i=1}^{I} \prod_{g=1}^{G} \frac{\prod_{r=1}^{R} \Gamma(\alpha + \sum_{j=1}^{J} z_{jg} x_{ijr})}{\Gamma(R\alpha + \sum_{r=1}^{R} \sum_{j=1}^{J} z_{jg} x_{ijr})} \end{aligned} \tag{10}$$

which is straightforward to compute.

Although $f(\mathbf{y_{k\ell}} \mid \mathbf{x})$ in (8) can in principle be computed exactly, the number of terms in the summation makes it prohibitively expensive. A much faster alternative is to use Markov chain Monte Carlo (MCMC) to sample from $f(\mathbf{z} \mid \mathbf{x})$ using (10) and then to compute $f(\mathbf{y_{k\ell}} \mid \mathbf{z}, \mathbf{x})$ using (9). The average of $f(\mathbf{y_{k\ell}} \mid \mathbf{z}, \mathbf{x})$ over the sample will then approximate $f(\mathbf{y_{k\ell}} \mid \mathbf{x})$.

Although a variety of MCMC methods can be used to sample from $f(\mathbf{y_{k\ell}} \mid \mathbf{z}, \mathbf{x})$, we have successfully used the following Metropolis within Gibbs sampling scheme. More precisely, we obtain our sample $\mathbf{z^1}, \ldots, \mathbf{z^M}$ by successively sampling from $f(\mathbf{z_j} \mid \mathbf{z_{-j}}, \mathbf{x})$ which is an overall Gibbs sampling scheme. Each sample from $f(\mathbf{z_j} \mid \mathbf{z_{-j}}, \mathbf{x})$ is obtain by random proposing a new $\mathbf{z_j} \rightarrow \mathbf{z_j^*}$ for judge j. Suppose $z_{jg} = 1$ and $z_{jg^*} = 1$. From (10), the acceptance probability of this new move is then $\min(1, A_j)$ where

$$
\begin{aligned}
A_j &= \prod_{i=1}^{I} \left[\frac{\alpha + \sum_{j=1}^{J} z_{jg^*} x_{ijr}}{\alpha - 1 + \sum_{j=1}^{J} z_{jg} x_{ijr}} \right] \\
&\quad \times \prod_{i=1}^{I} \left[\frac{R\alpha - 1 + \sum_{r=1}^{R} \sum_{j=1}^{J} z_{jg} x_{ijr}}{R\alpha + \sum_{r=1}^{R} \sum_{j=1}^{J} z_{jg^*} x_{ijr}} \right] \\
&\quad \times \frac{\beta + \sum_{j=1}^{J} z_{jg^*}}{\beta - 1 + \sum_{j=1}^{J} z_{jg}}
\end{aligned}
$$

This sample is then used to approximate the predictive distribution of $\mathbf{y_{k\ell}}$ by

$$
\hat{f}(\mathbf{y_{k\ell}} \mid \mathbf{x}) = \frac{\mathbf{1}}{\mathbf{M}} \sum_{\mathbf{m=1}}^{\mathbf{M}} \mathbf{f}(\mathbf{y_{k\ell}} \mid \mathbf{z^m}, \mathbf{x}) \tag{11}
$$

An especially nice feature of this approach is that a single MCMC sample can be used to simultaneously and quickly estimate the predictive distributions of all missing ratings.

4 SEARCHING FOR AN EFFECTIVE G

Because G is in fact unknown, a fully Bayesian approach would entail putting a prior on G and averaging $\hat{f}(\mathbf{y_{k\ell}} \mid \mathbf{x})$ over $\pi(G \mid \mathbf{x})$. Although MCMC can in principle be used to calculate $\pi(G \mid \mathbf{x})$, this appears to be an overwhelming calculation for even moderately sized databases. To avoid this calculation and the specification of a prior for G, we consider the more modest goal of selecting a value of G yielding effective predictions based on $\hat{f}(\mathbf{y_{k\ell}} \mid \mathbf{x})$. To find such a G, we use a directed hybrid search algorithm which combines cross-validated split and combine moves with MCMC calculation of $\hat{f}(\mathbf{y_{k\ell}} \mid \mathbf{x})$ in (11). The split move entails randomly choosing and splitting a judge group into two groups, and the combine move entails combining two randomly chosen judge groups into one group.

The implementation of our algorithm proceeds as follows. First, a fixed percentage of the observed ratings is randomly flagged and removed for a cross-validation purposes. Starting with an initial value of G, an initial group membership allocation ($\mathbf{z}$) is obtained via $n_{initial}$ samples from (10). An additional n_{outer} sampling steps on the group membership indicators are performed and predictions are made on the flagged ratings using (11) with the n_{outer} allocations. The predictive mean square error (MSE) obtained at this stage is set to be the TARGET MSE. Now a split or combine move is randomly proposed and n_{inner} sampling steps are performed to obtain new group membership allocations. These are used to make new predictions on the flagged ratings (after discarding the first $n_{inner\ warm}$ values), and to obtain the MOVE MSE. The proposed move is accepted if (TARGET MSE - MOVE MSE) $> c$, a preset threshold value. If the move is accepted, the TARGET MSE is updated to be the MOVE MSE. Notice that the allocations between split/combine move are generated via MCMC so that the resulting process behaves like a systematically re-started MCMC. The algorithm can be summarized as follows:

```
START
Randomly flag and remove $p$ % of the observed ratings
Start with an initial $G$ judge groups
Sample $\mathbf{z}$ for $n_{initial}$ steps
Sample $\mathbf{z}$ for $n_{outer}$ steps
Compute the TARGET MSE
DO LOOP
  Sample $\mathbf{z}$ for $n_{outer}$ steps
  Propose a SPLIT or COMBINE move
  IF (split move attempt) THEN
    Randomly select and split a judge group into two
  ELSE IF (combine move attempt) THEN
    Combine two randomly selected judge groups into one
  END IF
  (According to the proposed $G$ and $\mathbf{z}$)
  Sample $\mathbf{z}$ for $n_{inner}$ steps
  Discard $n_{inner\ warm}$ steps
  Compute the MOVE MSE
  IF ((TARGET MSE - MOVE MSE) $> c$) THEN
    Accept the proposed move
    Set TARGET MSE = MOVE MSE
  END IF
END DO LOOP
STOP
```

5 COMPARISON WITH OTHER APPROACHES

A natural competitor to our approach is the SM algorithm (1) based on constrained correlation (2). It may also be of interest to consider variants of our Bayesian approach which treat the missing ratings differently. Referring to our approach as Bayes 1, a variant, Bayes 2, is obtained by treating values of $\mathbf{x_{ij}} = (\mathbf{x_{ij1}}, \ldots, \mathbf{x_{ijR}})$, for which item i was not rated by judge j, as missing values in the likelihood (5). These missing values can be updated by introducing a Gibbs sampling step which samples from the full conditionals. Unfortunately, this dramatically increases the computational burden, especially if the data is sparse. Another variant, Bayes 3, is obtained by treating a missing rating as a new category, i.e. let $\mathbf{x_{ij}} = (\mathbf{x_{ij1}}, \ldots, \mathbf{x_{ij(R+1)}})$ and set $x_{ij(R+1)} = 1$ if item i was not rated by judge j. This approach, proposed by Breese et. al.(1998), requires renormalization of the posterior for prediction and appears to support larger values of G. All these approaches are described more fully in Chien (1998). To compare these various approaches, we evaluated their predictive performance on the following three different datasets.

5.1 THE SIMULATED DATA

The first data set was simulated according to the following scenario. Three groups of 40 judges (120 judges in all) are to rate each of 10 items on a scale of 1 to 5. For each item, judges are assigned preference probabilities which have been generated from one of the five template *Dirichlet* distributions: 1) $\mathcal{D}(20, 2, 1, 0.2, 0.1)$, 2) $\mathcal{D}(1.4, 20, 1.4, 0.2, 0.1)$, 3) $\mathcal{D}(0.1, 1.2, 20, 1.2, 0.1)$, 4) $\mathcal{D}(0.1, 0.2, 1.4, 20, 1.4)$ and 5) $\mathcal{D}(0.1, 0.2, 1, 2, 20)$. Note that each template corresponds to a strong preference for a particular rating. Within each group, the same templates are used across items to provide a nearly homogeneous grouping of preference probabilties. Our assigment of templates across items within groups is displayed in Table 3. From this setup, the ratings of all judges across items were then simulated from the model (5), and 20% of these ratings were then removed at random for performance evaluation.

Table 3: Simulated Data Template Arrangement

Groups	Templates									
1	4	1	1	5	5	1	5	5	1	1
2	4	3	3	4	4	3	5	5	3	3
3	2	5	5	2	2	5	2	2	5	5

5.2 THE PATTERNED DATA

The second data set was constructed with a deliberate pattern of missing ratings. Here, 160 judges rate 5 items from 1 to 5. The complete data on the LHS of Table 4 shows that there are two distinct judge groups with identical ratings, judges 1 to 80 and judges 81 to 160. However, instead of removing ratings randomly, we remove ratings from judges 41-80 and judges 121-160 in the manner shown on the RHS of Table 4. This patterned missingness was constructed to make it difficult for the SM algorithm to distinguish between distinct judge groups. In particular, the SM algorithm will here predict the missing ratings of judges 41-80 using both judges 1-40 and judges 121-160. Judges 121-160 are erroneously used because of they agree with Judges 41-80 on Item 1. A similar problem occurs for SM in predicting the missing ratings of judges 121-160. Our purpose in constructing this example is to explore the comparative performance of different approaches when ratings are not missing at random. Although the pattern of missingness is extreme in this data, nonrandom missingness is likely to occur such as when judges rate items they like but not those they dislike.

Table 4: Ratings for the Constructed Data

	Complete					Observed				
	Items					Items				
Judges	1	2	3	4	5	1	2	3	4	5
1	3	1	4	5	5	3	1	4	5	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	3	1	4	5	5	3	1	4	5	5
41	3	1	4	5	5	3	-	4	-	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
80	3	1	4	5	5	3	-	4	-	5
81	3	5	1	1	1	3	5	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
120	3	5	1	1	1	3	5	1	1	1
121	3	5	1	1	1	3	5	-	1	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
160	3	5	1	1	1	3	5	-	1	-

5.3 THE EACHMOVIE DATA

The third data set is real ratings data obtained from the EachMovie recommendation service run by the DEC Systems Research Center. Over a period of 18 months 72,916 users entered a total of 2,811,983 numeric ratings for 1,628 different movies (films and videos). Of these, we randomly selected data for 1373 users on 41 movies. See Table 5 for a snap shot of the data. As with our simulated data set, 20% of these ratings were then removed at random for performance evaluation.

Table 5: Snap Shot of the EachMove dataset

	Movies					
Judges	1	2	3	4	5	$\cdots$
1	-	-	-	4	6	$\cdots$
2	5	6	2	-	-	$\cdots$
3	4	6	-	-	5	$\cdots$
4	3	2	2	-	-	$\cdots$
5	5	5	-	3	5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$

5.4 PERFORMANCE

We evaluated the predictive performance of the various approaches, SM, Bayes 1, Bayes 2 and Bayes 3 on the three datasets described above. Performance was measured by MSE between the actual and predicted ratings. For each of the Bayes approaches, we used the hybrid search algorithm with parameters set as follows: Total sampling steps are 10,000, warm-up steps are 8,000, $n_{outer} = 10$, $n_{inner} = 50$, $n_{inner\ warm} = 20$, 10% of the observed ratings are removed for cross-validation, and for the MOVE MSE threshold $c = 0.01$. The performance comparison is summarized in Table 6.

Table 6: Predictive MSE Performance

Data Set	SM	Bayes 1	Bayes 2	Bayes 3
Simulated	0.692	0.636	0.636	0.649
Patterned	3.563	0.091	0.087	0.064
EachMovie	1.324	1.266	1.298	1.345

Bayes 1 and Bayes 2 outperformed SM on all three data sets, and Bayes 3 was only slightly worse on the EachMovie data. On the constructed data with patterned missingness, all three Bayes approaches substantially outperformed SM. The Bayes methods appear to be robust against such missingness because they weight the extent of matches between judges rather than correlated patterns. Although Bayes 1 and Bayes 2 performed similarly, we strongly prefer Bayes 1 because it requires so much less computational effort. The comparison between Bayes 1 and Bayes 3 is interesting. By treating missingness as a separate category, Bayes 3 outperforms Bayes 1 on our patterned missingness data. However, Bayes 3 performed less well than Bayes 1 on the other two data sets, especially the EachMovie data. It should also be mentioned that Bayes 3 requires slightly more computational effort than Bayes 1. Obviously, the comparisons here are limited and more investigation is needed, especially concerning the wide variety of patterns of missingness with are likely to be encountered in practice.

Acknowledgments

This research was supported by Texas ARP grant 003658-452, NSF grant DMS 98-03756 and research funding from the Graduate School of Business at the University of Texas at Austin.

References

Breese, J. S. , Heckerman, D. and Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the 14th conference on uncertainty in artifical intelligence*, Madison, WI, Morgan Kaufmann.

Chien, Y.H. (1998). Probabilistic Preference Modeling. *Ph.D. Thesis*, University of Texas at Austin.

Consonni, G. and Veronese P. (1995). A Bayesian Method for Combining Results from Several Binomial Experiments, *Journal of the American Statistical Association*, 90, 431, 935-944.

Paul, R, Neophytos, I, Mitesh, S. Peter, B, John, R. (1994). Grouplens: An Open Architecture for Collaborative Filtering of Netnews, *Communications of the ACM.*

Shardanand, U., Maes, P. (1995) Social Information Filtering: Algorithms for Automating "Word of Mouth", *Proceedings of the CHI-95 Conference*, Denver, CO, ACM Press.