Unsupervised Learning with Non-Ignorable Missing Data

Benjamin M. Marlin, Sam T. Roweis, Richard S. Zemel
Department of Computer Science
University of Toronto
Toronto, Ontario

Abstract

In this paper we explore the topic of unsupervised learning in the presence of non-ignorable missing data with an unknown missing data mechanism. We discuss several classes of missing data mechanisms for categorical data and develop learning and inference methods for two specific models. We present empirical results using synthetic data which show that these algorithms can recover both the unknown selection model parameters and the underlying data model parameters to a high degree of accuracy. We also apply the algorithms to real data from the domain of collaborative filtering, and report initial results.

1 Introduction

In large, real world data sets (such as those commonly used for machine learning research), the presence of a certain amount of missing data is inevitable. Probabilistic methods offer a natural framework for dealing with missing data, and there is a large body of work devoted to statistical analysis in this setting.

There are two important classes of missing data: missing data that is ignorable, and missing data that is non-ignorable. Ignorable missing data includes data that is missing completely at random (MCAR), and data that is missing at random (MAR). Intuitively, missing data is ignorable if the probability of observing a data item is independent of the value of that data item. Conversely, missing data is non-ignorable if the probability of observing a data item is dependent on the value of that data item.

The majority of statistical literature deals with the case where missing data is missing at random. However, there are several important cases where the missing at random assumption fails to hold. Well studied examples from statistics include non-response in surveys, panel data studies, and longitudinal studies. In surveys non-ignorable missing data often results from asking questions about income where the probability of non-response has been found to vary according to the income of the respondent. In such cases computing statistics like average income without taking the missing data mechanism into account will result in a biased estimator.

A much more complex domain where missing data may be non-ignorable is rating-based collaborative filtering [5]. The data in this domain often comes from recommender systems where users rate different items and receive recommendations about new items they might like. When a user is free to chose which items they rate, we hypothesize that many users will exhibit a bias toward rating items they like (and perhaps a few they strongly dislike). Thus the probability of observing a rating for a given item will depend on the user's rating for that item, and the missing ratings will not be missing at random. The best known methods for predicting user ratings are based on using unsupervised learning techniques to estimate the parameters of a probabilistic model over rating profiles. Just as in the simple mean income estimation problem, the model parameter estimates will be biased in the presence of non-ignorable missing data.

In this paper we consider the general problem of learning latent variable models in the presence of non-ignorable missing data with an unknown missing data mechanism. We present learning methods based on the Expectation Maximization (EM) algorithm for several different models. We present empirical results on several synthetic data sets showing that the learning procedure recovers the data and selection model parameters to a high degree of accuracy. We also present interesting results on real data from the collaborative filtering domain.

$$\mathcal{L}(\theta|Y^{obs}) = \log f(Y^{obs}|\theta) = \log \int f(Y^{obs}, Y^{mis}|\theta) dY^{mis} \tag{1}$$

$$\mathcal{L}(\theta,\mu|Y^{obs},R) = \log f(Y^{obs},R|\theta,\mu) = \log \int f(R|Y^{obs},Y^{mis},\mu) f(Y^{obs},Y^{mis}|\theta) dY^{mis} \tag{2}$$

$$\mathcal{L}_{\mathrm{MAR}}(\theta,\mu|Y^{obs},R) = \log f(R|Y^{obs},\mu) + \log \int f(Y^{obs},Y^{mis}|\theta) dY^{mis} = \mathcal{L}(\theta|Y^{obs}) + \mathcal{L}(\mu|Y^{obs},R) \tag{3}$$

2 Non-Ignorable Missing Data Theory

We begin with technical definitions of ignorable and non-ignorable missing data due to Little and Rubin, and review the theory of maximum likelihood estimation with non-ignorable data.

Let Y denote a complete data set and let Y^{obs} and Y^{mis} be the observed and missing elements of Y. Let R be a matrix of response indicators where $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. We define $f(Y,R|\theta,\mu) = f(Y|\theta)f(R|Y,\mu)$ to be the joint distribution over the data and response indicators. We refer to $f(Y|\theta)$ as the *data model* and $f(R|Y,\mu)$ as the *selection model*.

In the presence of missing data the correct maximum likelihood inference procedure is to maximize the full data likelihood $\mathcal{L}(\theta,\mu|Y^{obs},R) = \log f(Y^{obs},R|\theta,\mu)$ shown in equation 2 as opposed to the observed data log likelihood shown in equation 1. The only case where it is acceptable to rely on the observed data likelihood is when the missing at random (MAR) condition holds. The MAR condition is satisfied when $f(R|Y^{obs},Y^{mis},\mu) = f(R|Y^{obs},\mu)$ for all μ, and μ and θ are distinct parameters. If we suppose that the MAR condition holds we find that the full data log likelihood and the observed data log likelihood will give identical inferences for θ, as shown in equation 3.

When missing data is missing not at random (MNAR) this convenient result does not hold, and maximum likelihood estimation of the data model parameters θ based only on the observed likelihood will be biased. To obtain correct maximum likelihood estimates of the data model parameters a selection model is needed along with the data model [3, p. 218]. In most cases the parameters of the selection model will also be unknown. Fortunately the parameters of the combined data and selection model can be estimated simultaneously by maximizing the full data log likelihood using the standard EM algorithm.

3 Non-Ignorable Missing Data Models

Suppose we are given a data set containing N data vectors $\mathbf{y}_n$, each of length M. The value of each element y_{mn} is categorical, and is drawn from the set of possible values $\{1, ..., V\}$. We will assume that the $\mathbf{y}_n$ are independent, that any element y_{mn} may be unobserved, and that the selection process is unknown.

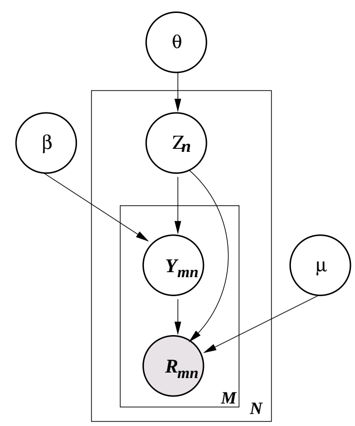

Figure 1: Combined data and selection model.

To specify a model for non-ignorable missing data we must choose a data model and a selection model. We choose a multinomial mixture model for the data model. This is a simple model for categorical data that has proved to be quite robust in the collaborative filtering domain [5]. Several other models for categorical data could also be used including the aspect model [2] and the URP model [4].

The more interesting choice is the choice of a selection model. In general, the probability that a particular variable is observed can depend on any other variable in the data model. Learning such an unstructured model would be intractable, so we must assume additional independence structure. To begin with, we assume that the probability of being observed is independent for each component of the data vector; this is formalized in equation 4. (Analogous simplifying assumptions have been used previously in similar discrete, non-ignorable contexts [6].)

$$P(\mathbf{r}|\mathbf{y},z) = \prod_{m=1}^{M} \mu_{y_m m z}{}^{(r_m)} (1 - \mu_{y_m m z})^{(1-r_m)} \tag{4}$$

If we represent the Bernoulli observation parameters $P(r_m = 1|y_m = v, Z = z) = \mu_{vmz}$ using conditional probability tables and combine this selection model with the multinomial mixture data model, we obtain a simple, tractable model for non-ignorable missing data. We call this the *CPT-vmz* model since the μ_{vmz} probabilities are represented by conditional probability tables (CPTs), and the selection probability depends on the settings of the variables v, m, and z. This combined data and selection model is depicted graphically in figure 1.

The variables in the model are the latent variable Z_n, the data variables Y_{mn}, and the corresponding response indicators R_{mn}. Recall that m indexes dimensions of the data vectors, and n indexes data cases. We suppress the data case index for simplicity when it is clear that we are referring to a single, generic data case. The parameters are the prior mixing proportions θ, the component distributions β, and the selection model parameters μ. To generate data from the combined multinomial mixture data model and *CPT-vmz* selection model we begin by sampling a state z of the latent variable Z according to $P(Z = z|\theta) = \theta_z$. We then sample a value v for each element y_m according to $P(y_m = v|Z = z, \beta) = \beta_{vmz}$. This is simply the standard generative process for the multinomial mixture model. Next, for each element y_m, we sample a response indicator variable according to the selection probability $P(r_m = r|y_m = v, M = m, Z = z, \mu) = \mu_{vmz}$. We then discard the values of y_m for which $r_m = 0$.

Depending on the type of selection effect that is present in the data, it may be desirable to further constrain the selection model. Imposing independence and factorization conditions on the Bernoulli probabilities μ_{vmz} results in a range of selection models with different properties. In this paper we concentrate on two models in particular: a model we call *CPT-v* that makes the additional independence assumption $P(r_m|y_m, m, z) = P(r_m|y_m)$, and a model we call *LOGIT-v,mz* which proposes a functional decomposition of $P(r_m|y_m, m, z)$ based on the logistic function.

3.1 The *CPT-v* Model

While the *CPT-vmz* model is highly flexible and can model a range of very different selection effects, this flexibility may result in over fitting on many data sets. By contrast the *CPT-v* model asserts that only the value of a random variable affects its chance of being observed or missing. The *CPT-v* model is highly constrained, but this makes it appealing when we have limited observations and cannot robustly fit the full *CPT-vmz* model. Examples of the type of effects this model captures are "mostly high values are observed", or "only extreme values are observed". However, it can not efficiently represent effect of the type "data item m is observed almost always." As we will see, the strict assumptions of the *CPT-v* model can cause problems during model fitting if the data contains strong item-based effects.

Algorithm 1 Expectation Maximization Algorithm for the *CPT-v* model

E-Step:

$\lambda_{vmzn} \leftarrow (\delta(y_{mn}, v)\mu_v\beta_{vmz})^{r_{mn}}((1-\mu_v)\beta_{vmz})^{1-r_{mn}}$

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

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

M-Step:

$\theta_z \leftarrow \frac{\sum_{n=1}^{N}\phi_{zn}}{\sum_{n=1}^{N}\sum_{z=1}^{K}\phi_{zn}}$

$\beta_{vmz} \leftarrow \frac{\sum_{n=1}^{N}\phi_{zn}\lambda_{vmzn}/\gamma_{mzn}}{\sum_{n=1}^{N}\phi_{zn}}$

$\mu_v \leftarrow \frac{\sum_{n=1}^{N}\sum_{z=1}^{K}\phi_{zn}\sum_{m=1}^{M}r_{mn}\lambda_{vmzn}/\gamma_{mzn}}{\sum_{n=1}^{N}\sum_{z=1}^{K}\phi_{zn}\sum_{m=1}^{M}\lambda_{vmzn}/\gamma_{mzn}}$

Equation 5 gives the full data likelihood for the *CPT-v* model. We define the intermediate variables λ_{vmzn} and γ_{mzn} in equations 6 and 7.

$$\mathcal{L}(\theta, \beta, \mu|[\mathbf{y}]^{obs}, [\mathbf{r}]) = \sum_{n=1}^{N}\log\sum_{z=1}^{K}\theta_z\prod_{m=1}^{M}\gamma_{mzn} \quad (5)$$

$$\lambda_{vmzn} = (\delta(y_{mn}, v)\mu_v)^{r_{mn}}(1-\mu_v)^{1-r_{mn}}\beta_{vmz} \quad (6)$$

$$\gamma_{mzn} = \sum_{v=1}^{V}\lambda_{vmzn} \quad (7)$$

The posterior distribution over settings of the latent variable Z is given in equation 8 using the intermediate variables. Of course, the intractable, full posterior over both $\mathbf{y}_n^{mis}$, and Z_n is never needed during learning.

$$P(z|\mathbf{y}_n^{obs}, \mathbf{r}_n) = \phi_{zn} = \frac{\theta_z\prod_{m=1}^{M}\gamma_{mzn}}{\sum_{z=1}^{K}\theta_{z'}\prod_{m=1}^{M}\gamma_{mzn}} \quad (8)$$

Expectation maximization updates can now be found by maximizing the expected complete log likelihood with respect to the data model and selection model parameters. We show the EM algorithm for the *CPT-v* model in algorithm 1.

3.2 The *LOGIT-v,mz* Model

The *CPT-v* model makes the very strong assumption that a single value-based selection effect is responsible for generating all missing data. We would like to allow different effects for different data items m, as well as allowing the setting of the latent variable z to influence the missing data process. As a modeling choice of intermediate complexity, we propose the *LOGIT* family of selection models. The main feature of *LOGIT* models is the assumption that the selection model parameters μ_{vmz} result from the interaction of multiple lower dimensional factors. In particular, these models allow all of v, m, z to influence the probability of a data element being missing, but constrain the effects to a particular functional family.

In the case of *LOGIT-v,mz* two factors are proposed. One factor σ_v models a value-based effect, while the other factor ω_{mz} models a joint element index/latent variable effect. This latter effect can include factors that are item-specific (a given data item m can have its own probability of being missing), and latent variable-specific (each mixture component z generates its own pattern of missing data). The values of these factors can be arbitrary real numbers and they are combined to obtain the selection probabilities through the logistic function as seen in equation 9. This parameterization was selected because it is more flexible than simple factorizations, such as a product of Bernoulli probabilities. Suppose a data set contains strong value-based selection effects for most data items, but the values for one data items are always observed regardless of their values. *LOGIT-v,mz* can account for this by setting ω_{mz} to a large value. The logistic combination rule then allows ω_{mz} to override σ_v and produce a selection probability of 1 for just this data item. In a product of distributions decomposition this simply is not possible. As we will later see, this flexibility is needed for modeling "blockbuster" effects in the collaborative filtering domain.

$$\mu_{vmz} = \frac{1}{1 + \exp^{-(\sigma_v + \omega_{mz})}} \tag{9}$$

Given values for the selection model parameters σ_v and ω_{mz}, we can compute the complete set of selection probability parameters μ_{vmz} according to equation 9. If we then redefine the intermediate variable λ_{vmzn} according to equation 10, the full likelihood and the posterior over the latent variable have exactly the same form for the *LOGIT-v,mz* model as they do for the *CPT-v* model.

$$\lambda_{vmzn} = (\delta(y_{mn}, v)\mu_{vmz})^{r_{mn}}(1 - \mu_{vmz})^{1-r_{mn}}\beta_{vmz} \tag{10}$$

Unlike the *CPT-v* case, closed form selection model parameter updates cannot be found for *LOGIT-v,mz*. Instead, numerical optimization methods must be used to adapt these parameters. We sketch a suitable EM algorithm for the *LOGIT-v,mz* model in algorithm 2. Note that to ensure the full likelihood is non-decreasing, line search must be used to determine acceptable step sizes α at each iteration.

Algorithm 2 Expectation Maximization Algorithm for the *LOGIT-v,mz* model

E-Step:

$\mu_{vmz} \leftarrow \frac{1}{1+\exp^{-(\sigma_v+\omega_{mz})}}$

$\lambda_{vmzn} \leftarrow (\delta(y_{mn}, v)\mu_{vmz})^{r_{mn}}(1 - \mu_{vmz})^{1-r_{mn}}\beta_{vmz}$

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

$\phi_{zn} \leftarrow \dfrac{\theta_{z_n} \prod_{m=1}^{M} \gamma_{mzn}}{\sum_{z=1}^{K} \theta_{z'} \prod_{m=1}^{M} \gamma_{mzn}}$

M-Step:

$\theta_z \leftarrow \dfrac{\sum_{n=1}^{N} \phi_{zn}}{\sum_{n=1}^{N}\sum_{z=1}^{K} \phi_{zn}}$

$\beta_{vmz} \leftarrow \dfrac{\sum_{n=1}^{N} \phi_{zn}\lambda_{vmzn}/\gamma_{mzn}}{\sum_{n=1}^{N} \phi_{zn}}$

$\sigma \leftarrow \sigma - \alpha_\sigma \displaystyle\sum_{n=1}^{N}\sum_{z=1}^{K} \phi_{zn} \sum_{m=1}^{M} \delta(y_{mn}, v)(r_{mn} - \mu_{vmz})$

$\omega \leftarrow \omega - \alpha_\omega \displaystyle\sum_{n=1}^{N} \phi_{zn} \sum_{v=1}^{V} (r_{mn} - \mu_{vmz})$

4 Synthetic Data Experiments

Our first goal is to examine whether, in situations where the assumptions underlying them are satisfied, the proposed models are able to recover both the unknown selection mechanism and a correct model of the data. To this end, we generated synthetic data sets patterned after real data sets from the collaborative filtering domain.

4.1 Generating Synthetic Data

We generate complete synthetic data sets according to a hierarchical Bayesian procedure. In particular, we choose a $K = 6$ component multinomial mixture data model with M=100 data variables, and $V = 5$ values per variable. The mixture model parameters are sampled from an appropriate length Dirichlet prior (uniform, strength two). We sample n=5000 data cases from the mixture model to form a single, complete data set.

To generate data that conforms to the *CPT-v* selection model, we created several sets of selection parameters and used these to sample the complete data. For the purpose of these experiments we use a *CPT-v* selection model with a special functional form that allows the strength of the missing data effect to be easily quantified. In particular we let $\mu_v(s) = s(v - 3) + 0.5$, where s is the parameter that controls the strength of the effect. Note that since the underlying data distribution is uniform across values, any choice of s in the range $0 < s < 0.25$ yields an overall observation rate of 0.5. We create ten sets of observation probabilities by evenly varying the parameter s from 0 to 0.225.

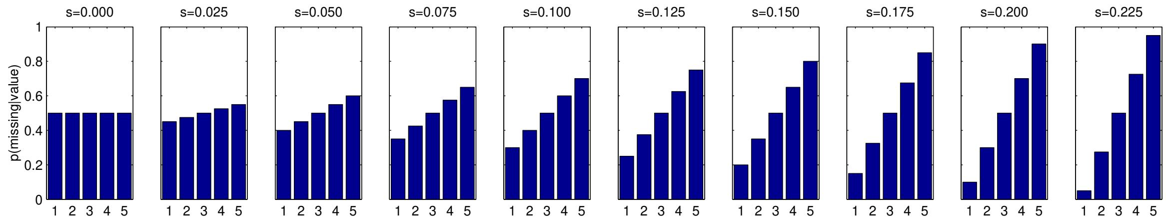

Figure 2: Selection probabilities for the effect $\mu_v(s) = s(v-3)+0.5$. The parameter s controls the strength of the missing data effect. Here we show $\mu_v(s)$ at ten equally spaced values on the interval $0 \leq s \leq 0.225$.

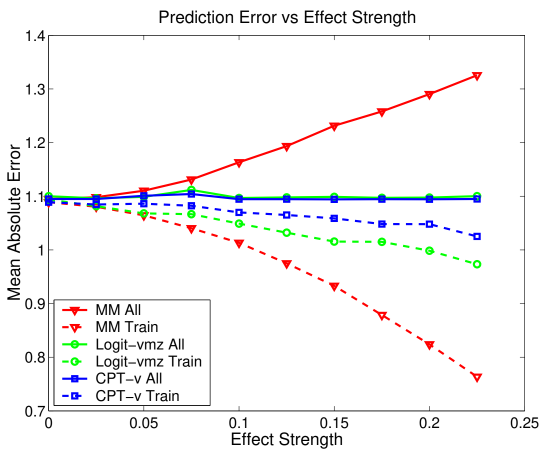

Figure 3: MAE_{All} and MAE_{Tr} versus strength of the *CPT-v* missing data effect for the multinomial mixture, *CPT-v*, and *LOGIT-v,mz* models.

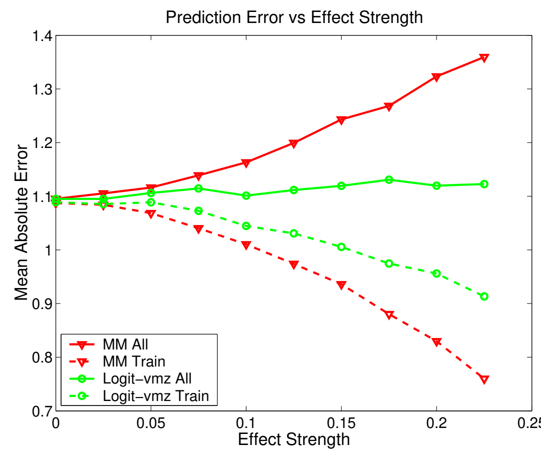

Figure 4: MAE_{All} and MAE_{Tr} versus strength of the *LOGIT-v,mz* missing data effect for the multinomial mixture, and *LOGIT-v,mz* models.

The resulting parameters are shown in figure 2. These ten sets of observation parameters were used to sample ten different training sets from the complete data set.

To generate data that conforms to the *LOGIT-v,mz* model we need to define the selection model parameters σ_v and ω_{mz}. We create 10 different selection models by setting:

$$\sigma_v(s) = \log(\mu_v(s)/(1-\mu_v(s)))$$

$$\omega_{mz} = \log(\pi_{mz}/(1-\pi_{mz}))$$

The π_{mz} values are sampled from a Beta prior. Note that we have chosen these values so that when the logistic function is applied to $\sigma_v(s)$ the result is $\mu_v(s)$, and when it is applied to ω_{mz} the result is π_{mz}. We compute the corresponding set of selection probabilities $\mu_{vmz}(s)$ and use these to sample ten different training sets.

4.2 Experimental Procedure

The mixture of multinomials model, the *CPT-v* model, and the *LOGIT-v,mz* model were trained until convergence of their respective likelihood functions on all ten of the *CPT-v* training sets, and all ten of the *LOGIT-v,mz* training sets. After training, each model was used to predict the complete data vector $\hat{\mathbf{y}}_n$ for each data case n given the training set (observed) values for that data case. We repeat this training and prediction process three times with different initializations to account for local minima.

In order to judge the performance of each model under increasing missing data effect strength, we use the true and predicted ratings to measure a quantity called the mean absolute error (MAE), which is commonly used as an error measure in the collaborative filtering domain. We measure the MAE restricted to the training data as defined in equation 11, as well as the MAE computed over the complete data set as seen in equation 12. Note that on a real data set we have no choice but to compute the MAE restricted to the observed ratings, which corresponds to equation 11. If a given model accurately learns both the data model and selection model parameters for each setting of the effect strength s, the computed MAE_{All} values should be approximately constant indicating little degradation in performance.

$$MAE_{Tr} = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{r_{mn}|y_{mn} - \hat{y}_{mn}|}{\sum_{m=1}^{M} r_{mn}} \quad (11)$$

$$MAE_{All} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} |y_{mn} - \hat{y}_{mn}| \quad (12)$$

It is very important to note that under a non-ignorable selection model, the value of MAE_{Tr} as defined by equation 11 is a biased estimate of $E[|y - \hat{y}|]$, were we to sample items uniformly at random. Since the selection model is non-ignorable and also unknown, it is not possible to introduce a correction factor that will allow for an unbiased estimate of $E[|y - \hat{y}|]$ from the training data alone. On the other hand, the value of MAE_{All} as computed by equation 12 is unbiased because it is computed from the complete set of true ratings. However, such a complete set of ratings is not currently available for real data sets.

4.3 Results

In figure 3 we show the results of the synthetic data experiment with the *CPT-v* selection model effect. At $s = 0$, corresponding to the first histogram in figure 2, the selection effect is constant across values v and is thus ignorable. In this case we would expect MAE_{Tr} to be approximately equal to MAE_{All}. In addition we would expect all three models to achieve approximately the same prediction performance since all three are based on an underlying multinomial mixture data model. The experimental results we obtain at $s = 0$ are exactly in line with the theory.

As we increase s we would expect that the value of MAE_{All} would increase for the multinomial mixture model since its parameter estimates are based only on the observed training data. Both the *CPT-v* and *LOGIT-v,mz* models have the capacity to exactly model this selection effect. If these models learn accurate data and selection model parameters then the measured value of MAE_{All} should be approximately constant. A further note is that *LOGIT-v,mz* actually has excess capacity when applied to the *CPT-v* selection effect data, so over fitting may be an issue.

As we see in figure 3 the MAE_{All} curves follow exactly the trends we have predicted for each of the models. However, the MAE_{Tr} curves exhibit an interesting downward trend indicating that error on the training set actually decreases as the missing data effect strength is increased. This is not unexpected in the mixture of multinomials model since it is able to concentrate more of its capacity on fewer rating values and thus achieve lower error on the training data. *CPT-v* and *LOGIT-v,mz* exhibit a slight decrease in error on the training set as the selection strength is increased, but it is not accompanied by a corresponding increase on the complete data set.

In figure 4 we show the results of the synthetic data experiment with the *LOGIT-v,mz* selection effect. The most notable result of this experiment is the fact that the learning procedure for the *CPT-v* model, algorithm 1, converges to a "boundary solution" for the μ_v parameters for all values of the effect strength s. Specifically, at convergence the μ values have the form $\mu_1 = c$, $\mu_j \approx 1$, where c reflects the global sparsity rate and $2 \leq j \leq 5$. This appears to indicate that the *CPT-v* model lacks the capacity to model the *LOGIT-v,mz* missing data effect. This failure of the *CPT-v* model may result from the presence of strong item-based effects in the *LOGIT-v,mz* data. For example, suppose an item is always observed regardless of its value. The only way *CPT-v* can explain this is by increasing the values of μ. Of course it cannot increase all the values of μ since it must still explain the fraction of data that is missing. The most likely solution appears to be exactly the boundary solution explained above. This problem may also simply be a failure of the maximum likelihood framework. We plan to explore a Bayesian approach to prediction to determine if the problem actually lies with the model, or the estimation and prediction techniques.

The trends for the mixture of multinomials model are quite similar to the previous case with similar explanations applying. The trends for the *LOGIT-v,mz* model are also similar to the previous case. One slight difference is that the MAE_{All} curve is more noisy and increases somewhat with s. The most likely explanation is an insufficient amount of training data to properly estimate all the ω_{mz} parameters. The previous case is easier for the *LOGIT-v,mz* model in this respect since the *CPT-v* data contains no item or latent variable-based effects.

Perhaps the most important point illustrated by figures 3 and 4 is that estimating the prediction error on the observed data only can be an arbitrarily poor estimate of the error on the complete data set in the presence of a non-ignorable missing data effect. In both graphs we see that the multinomial mixture model attains the lowest error on the observed data, when in fact its true error rate is the highest among the three models.

5 Real Data Experiments

Unfortunately, the task of evaluating a method for unsupervised learning in the presence of non-ignorable

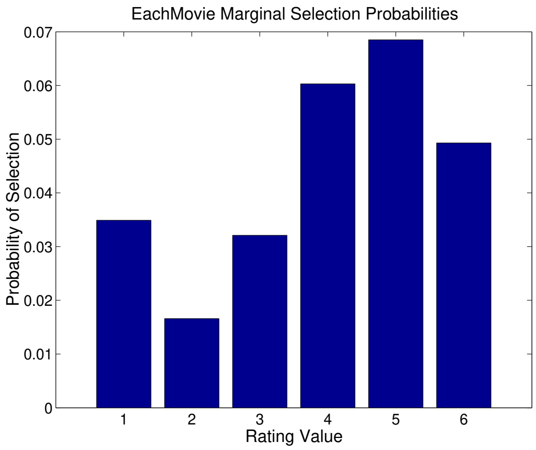

Figure 5: EachMovie marginal selection probabilities $P(r = 1|y = v)$. Computed under a learned *LOGIT-v,mz* model from parameters $\sigma, \omega, \beta, \theta$.

Figure 6: EachMovie Full Data Log Likelihood

	Full Data Log Likelihood
LOGIT-v,mz	-8.750368×10^6
Multinomial Mixture	-1.164895×10^7

missing data on real data sets is problematic. Standard evaluation procedures involve measuring prediction error on held out observed data. This hold-out method entails an inherent bias in our case, as it is only possible to evaluate predictions for items that are not missing. That is, unlike the case for synthetic data, evaluating MAE_{All} is impossible for real data. A learning method based on the MAR assumption would be expected to perform at least as well as one based on MNAR on such a biased hold-out experiment: a system can safely ignore any dependence of the item's response on missing values if it is never evaluated on missing items.

We are currently considering a novel interactive data gathering procedure designed to collect exactly the type of data needed to validate the proposed models. The main idea is to conduct an interactive, two stage survey. In the first stage participants would be free to select and rate any number of items they wished. This data would then form the training set, and would likely contain a strong non-ignorable missing data effect. In the second stage of the survey a different set of items would be randomly selected for each participant to rate. Ratings for these randomly sampled items would then form the test set. While this might be difficult for items like movies, it is quite simple for items like music where a clip can be played if the participant is not familiar with a particular randomly chosen item. Since the test data is a randomly chosen set of items, evaluating prediction error on the test set gives an unbiased estimate of the error on the whole data set.

It is still interesting to observe what the proposed models learn on currently available real data sets, even if this results can not be considered conclusive. We note that not all currently available ratings-based collaborative filtering data can be used with the proposed models. The key assumption is that users are free to rate items at will. Any source of data where the users are constrained to rate a sequence of items determined solely by a recommender system violates this assumption. In this case the proposed models would essentially be learning the selection bias caused by the recommender system attempting to predict items it thinks the user will like. While this may actually be an interesting method to test the efficacy of a recommender system, it is not what we intend here.

We chose to apply the *CPT-v* and *LOGIT-v,mz* models to a well known collaborative filtering data set: EachMovie. EachMovie is a collaborative filtering data set collected from a movie recommender system operated by the Compaq Systems Research Center. It contains about 1600 movies and 70000 users. EachMovie is well known to contain a "blockbuster" effect where several movies have been rated by almost all users, while others are rated by just a few users. EachMovie also contains a fairly strong user-based effect: the number of movies rated per user ranges from one to several thousand. EachMovie is widely believed to contain quite a substantial bias toward high rating values. The underlying recommender system used to collect the data also allowed for free user interaction, which satisfies our main requirement.

EachMovie appears too complicated for the *CPT-v* model using maximum likelihood learning. As in the synthetic *LOGIT-v,mz* experiment, *CPT-v* converges to uninformative boundary solutions on EachMovie. On the other hand, *LOGIT-v,mz* appears to converge to parameter estimates that are in very good alignment with the effects that are thought to occur in the data set.

After convergence, we can examine the learned parameters $\sigma, \omega, \beta, \theta$ and use them to compute the marginal probability $P(r = 1|y = v)$ as a summary of the learned selection effect (averaged across all items and settings of the latent variable). We show the computed marginal selection probabilities for EachMovie in figure 5. This figure exhibits a definite skew toward high ratings values, although the dip at the highest rating value is somewhat suspect.

While we cannot compute an unbiased estimator of the expected mean absolute error for the data set, we can compute another quantity that will allow us to compare the *LOGIT-v,mz* model with a MAR multinomial mixture model: the full data likelihood. The mixture of multinomials model has no explicit selection model and optimizes only an observed data likelihood as seen in equation 1. However, as we see in equation 3 we can obtain the full data likelihood from the observed data likelihood by accounting for the likelihood of the response indicators. If we suppose a simple MAR scheme with a global observability parameter μ, the log likelihood of the complete set of response indicators is given by $\log(\mu)\sum_n\sum_m r_{mn} + \log(1-\mu)\sum_n\sum_m(1-r_{mn})$. In this case the maximum likelihood estimator for μ is simply $\mu = \frac{1}{NM}\sum_n\sum_m r_{mn}$.

We show the full data likelihood values for the *LOGIT-v,mz* model and the multinomial mixture model computed for the EachMovie data set in figure 6. We see that *LOGIT-v,mz* obtains a significantly higher full data log likelihood value than the simple MAR multinomial mixture model.

6 Extensions and Future Work

In addition to the research directions mentioned in the previous sections, we are also considering extensions of the proposed framework that will allow us to model a wider range of missing data problems. In the original framework, the binary response value R_m for an item m is determined by the presence of a rating Y_m for that item. We might also decouple these two variables, and allow a response to exist for an item ($R_m = 1$) when the rating is not known. This situation often arises in Web-based systems where we may have information about many items a user has viewed, but a relatively small number of ratings. Only minor changes are required to reformulate the proposed models to handle this type of data.

7 Conclusions

In this paper, we have proposed several probabilistic selection models which treat missing data as a systematic (non-ignorable) rather than random (ignorable) effect. Coupled with a basic discrete latent variable model for user-item data, these selection mechanisms allow us to model data sets in which known (or suspected) response biases exist. We have derived efficient learning and inference algorithms to jointly estimate the data and selection model parameters in an unsupervised way, and verified that these algorithms can recover both the unknown selection model parameters and the underlying data model parameters to a high degree of accuracy under a wide variety of conditions. We have also shown that when an unbiased estimate of their performance is available, our models do substantially better than comparable models which do not account for the missing data mechanism. Finally, we have applied these models to a real world collaborative filtering data set, EachMovie, obtaining initial results that support several "folk beliefs" about the patterns of missing data in this data set.

Acknowledgments

We thank Nathan Srebro for helpful comments and discussions about missing data and collaborative filtering. His comments on an internal presentation of this work were instrumental in revising this paper. We thank Krisztina Filep for reviewing the final draft. Lastly, we thank the Compaq Computer Corporation for the use of the EachMovie data set.

References

[1] 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.

[2] T. Hofmann. Learning What People (Don't) Want. In *Proceedings of the European Conference on Machine Learning (ECML)*, 2001.

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

[4] B. Marlin. Modeling user rating profiles for collaborative filtering. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS-2003)*, 2003.

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

[6] E. A. Ramalho and R. J. Smith. Discrete choice models for nonignorable missing data. In *Proceedings of the Econometric Society Fifty-Seventy European Meeting (ESEM'2002)*, 2002.