Probabilistic Matrix Factorization with Non-random Missing Data

José Miguel Hernández-Lobato JMH233@CAM.AC.UK
Neil Houlsby NMTH2@CAM.AC.UK
Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
University of Cambridge, Department of Engineering, Cambridge CB2 1PZ, UK

Abstract

We propose a probabilistic matrix factorization model for collaborative filtering that learns from data that is *missing not at random* (MNAR). Matrix factorization models exhibit state-of-the-art predictive performance in collaborative filtering. However, these models usually assume that the data is *missing at random* (MAR), and this is rarely the case. For example, the data is not MAR if users rate items they like more than ones they dislike. When the MAR assumption is incorrect, inferences are biased and predictive performance can suffer. Therefore, we model both the generative process for the data and the missing data mechanism. By learning these two models jointly we obtain improved performance over state-of-the-art methods when predicting the ratings and when modeling the data observation process. We present the first viable MF model for MNAR data. Our results are promising and we expect that further research on NMAR models will yield large gains in collaborative filtering.

1. Introduction

Collaborative filtering (CF) data consists of ratings by users on a set of items. CF systems learn patterns in this data to make accurate predictions, for example, in order to recommend new items to users. Typically, most users rate only a small fraction of the available items, so most of the CF data is *missing*. Probabilistic matrix factorization (MF) models have become popular for CF because i) they can be robust to overfitting and come with automatic estimates of uncertainty (Mnih & Salakhutdinov, 2007); ii) they can be adapted to different data types, such as continuous, binary or ordinal data (Stern et al., 2009); and iii) they can theoretically handle missing data in a formal manner. A common assumption that MF models make is that CF data is *missing at random* (MAR) (Little & Rubin, 1987). That is, the process that selects the observed data (the observation process) is independent of the value of this unobserved data.

When the data is MAR, the observation process can be ignored and standard inference methods can be used without introducing bias. However, there is evidence that CF data is missing not at random (MNAR) (Marlin & Zemel, 2007). Consider the following scenarios: i) users only watch movies that they like, and only rate movies that they watch; ii) a user only provides extreme ratings, that is, they only provide feedback particularly bad or good items; iii) certain basic items (e.g. stationary) are simply expected to function correctly and these items are only rated when they are defective. The first scenario is an example of global censoring where low-valued ratings are usually missing. The second two are examples of local censoring where missing ratings for specific users or items take higher or lower values on average than the observed ones. In these scenarios dependencies between the missing data values and the observation process exist, and consequently the data is *not* MAR. When the MAR assumption is incorrect, inferences can be biased and prediction accuracy suffers.

The lack of robustness of the MAR assumption has been widely addressed in the statistics literature. However, almost all probabilistic models for CF ignore this issue. We fill this gap by extending state-of-the-art MF models for CF to the MNAR scenario. The general approach for dealing with MNAR data is to learn jointly a complete data model (CDM), that explains how the data is generated, and a missing data model (MDM), that explains the observation process for the data (Little & Rubin, 1987). Our CDM is a new MF model for ordinal rating data with state-of-the-art predictive performance. This model uses hierarchical priors to increase robustness to the selection of hyper-parameters and is heteroskedastic in the sense that the ratings of different users and items can exhibit different levels of noise (Lakshminarayanan et al., 2011). Our MDM is also a MF model that can capture complex dependencies between the missing data and the observation process, such as those de-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

scribed in the previous paragraph. We perform efficient inference in the resulting MF model for MNAR data (MF-MNAR) using expectation propagation (Minka, 2001) and stochastic variational inference (Hoffman et al., 2013).

Experimentally, the combination of the MDM and the CDM in MF-MNAR produces gains in both the modeling of the ratings and the modeling of the data observation process. We also find that MF-MNAR outperforms the MAR version of this method, other state-of-the-art MF MAR alternatives (Paquet et al., 2012) as well as an alternative model for MNAR rating data based on mixtures of multinomials (Marlin & Zemel, 2009). In summary, we present the first attempt to model MNAR data using probabilistic matrix factorization. Our results are promising and we expect that additional research in this domain will yield significant further improvements in CF systems.

2. Probabilistic Treatment of Missing Data

The theory of missing data has been widely studied. We review here the main principles developed in (Little & Rubin, 1987) in the context of rating data. Our data is formed by ratings $r_{i,j}$ given by user i on item j, where $i = 1, \ldots,$ n and $j = 1, \ldots, d$. We collect these into an $n \times d$ rating matrix $\mathbf{R} = \{\mathbf{R}^{\mathcal{O}}, \mathbf{R}^{\neg\mathcal{O}}\}$, where $\mathbf{R}^{\mathcal{O}}$ and $\mathbf{R}^{\neg\mathcal{O}}$ denote the the sets of *observed* and *missing* entries in $\mathbf{R}$, respectively. For each $r_{i,j}$, we define a Bernoulli random variable $x_{i,j}$ that indicates whether $r_{i,j}$ is observed ($x_{i,j} = 1$) or not ($x_{i,j} = 0$), and collect all the $x_{i,j}$ in the $n \times d$ binary matrix $\mathbf{X}$. We assume that $\mathbf{R}$ is generated by a complete data model (CDM) with parameters $\mathbf{\Theta}$, and $\mathbf{X}$ is generated by a missing data model (MDM) with parameters $\mathbf{\Omega}$. Both models may also share a set of latent variables $\mathbf{Z}$. The joint distribution for $\mathbf{R}$, $\mathbf{X}$, and $\mathbf{Z}$ given $\mathbf{\Theta}$ and $\mathbf{\Omega}$ is

$$p(\mathbf{X}, \mathbf{R}, \mathbf{Z}|\mathbf{\Theta}, \mathbf{\Omega}) = p(\mathbf{X}|\mathbf{R}, \mathbf{\Omega}, \mathbf{Z}) p(\mathbf{R}, \mathbf{Z}|\mathbf{\Theta}) . \quad (1)$$

Most machine learning focuses on the estimation of the CDM given by $p(\mathbf{R}, \mathbf{Z}|\mathbf{\Theta})$. In (1), $p(\mathbf{X}|\mathbf{R}, \mathbf{\Theta}, \mathbf{Z})$ is the MDM, which is normally ignored in CF systems.

The mechanisms for missing data are usually divided into three classes (Little & Rubin, 1987): completely missing at random (CMAR), missing at random (MAR) and missing not at random (MNAR). CMAR is the most restrictive assumption, where the probability of observing a rating is independent of the value of any rating or latent variable generated by the CDM, that is, $p(\mathbf{X}|\mathbf{R}, \mathbf{Z}, \mathbf{\Omega}) = p(\mathbf{X}|\mathbf{\Omega})$. With MAR data, the observation probability depends only upon the value of the *observed* data and the MDM parameters, that is $p(\mathbf{X}|\mathbf{R}, \mathbf{\Omega}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{R}^{\mathcal{O}}, \mathbf{\Omega})$. This assumption is popular in machine learning because it means that the MDM can be ignored without introducing any biases during inference. Under the MAR assumption, the likelihood for $\mathbf{\Theta}$ given $\mathbf{R}^{\mathcal{O}}$ is

$$\mathcal{L}(\mathbf{\Theta}|\mathbf{R}^{\mathcal{O}}) = \sum_{\mathbf{R}^{\neg\mathcal{O}}} \int p(\mathbf{X}|\mathbf{R}, \mathbf{Z}, \mathbf{\Omega}) p(\mathbf{R}, \mathbf{Z}|\mathbf{\Theta}) d\mathbf{Z} \quad (2)$$

$$\overset{\text{MAR}}{=} p(\mathbf{X}|\mathbf{R}^{\mathcal{O}}, \mathbf{\Omega}) \sum_{\mathbf{R}^{\neg\mathcal{O}}} \int p(\mathbf{R}, \mathbf{Z}|\mathbf{\Theta}) d\mathbf{Z} \quad (3)$$

$$= p(\mathbf{X}|\mathbf{R}^{\mathcal{O}}, \mathbf{\Omega}) p(\mathbf{R}^{\mathcal{O}}|\mathbf{\Theta}) \propto p(\mathbf{R}^{\mathcal{O}}|\mathbf{\Theta}) .$$

Since $p(\mathbf{X}|\mathbf{R}^{\mathcal{O}}, \mathbf{\Omega})$ is constant with respect to $\mathbf{\Theta}$, we can ignore the MDM when learning $\mathbf{\Theta}$.

However, in the general case of MNAR data, $\mathbf{X}$ will not be independent of $\mathbf{R}$ or $\mathbf{Z}$. In this case, the step from (2) to (3) does not hold and, when integrating over $\mathbf{Z}$ and summing over $\mathbf{R}^{\neg\mathcal{O}}$, one must weight the CDM likelihood $p(\mathbf{R},$ $\mathbf{Z}|\mathbf{\Theta})$ by the observation probabilities given by the MDM $p(\mathbf{X}|\mathbf{R}, \mathbf{Z}, \mathbf{\Omega})$. This is because the binary matrix $\mathbf{X}$ has *information* about the possible values that $\mathbf{R}^{\neg\mathcal{O}}$ or $\mathbf{Z}$ may have taken. Maximum likelihood estimates and Bayesian inference will be biased if the MAR assumption is used but the data is MNAR. For example, consider users who mainly rate items that they like and seldom rate items that they dislike. Under the MAR assumption, the estimated CDM will over-estimate the value of the missing ratings that have not yet been provided by such users.

We can correct this observational bias by jointly learning the CDM and the MDM. We now describe how to do this with an ordinal matrix factorization model for MNAR data.

3. An Ordinal Matrix Factorization Model with Data not Missing at Random

We are given a dataset $\mathcal{D} = \{r_{i,j} : 1 \leq i \leq n, 1 \leq j \leq d,$ $r_{i,j} \in \{1, \ldots L\}, (i,j) \in \mathcal{O}\}$ of discrete ratings by n users on d items, where the possible ratings are $1 < \ldots < L$ and $\mathcal{O}$ is the set of pairs of users and items for which a rating is available. $\mathcal{D}$ is a subset of the entries of a *complete* $n \times d$ rating matrix $\mathbf{R}$. In practice, $\mathcal{D}$ contains only a small fraction of the entries in $\mathbf{R}$. We model the location of the entries included in $\mathcal{D}$ using an $n \times d$ binary matrix $\mathbf{X}$, where $x_{i,j} = 1$ if $r_{i,j} \in \mathcal{D}$ (observed) and $x_{i,j} = 0$ otherwise (missing). We first describe a probabilistic model for the generation of $\mathbf{R}$, then we describe another model that generates $\mathbf{X}$ given $\mathbf{R}$. The first model is the *complete data model* (CDM) and the second one the *missing data model* (MDM). Our new method for ordinal Matrix Factorization with data Missing Not At Random (MF-MNAR) is formed by combining these two models into a single model.

The factor graph for the distribution implied by MF-MNAR is shown in Figure 2. In this graph the square nodes correspond to factors in the distribution and the circular nodes represent random variables. The edges show dependencies of factors on variables (Kschischang et al., 2001). A full description of the factors in Figure 2 is given in the supple-

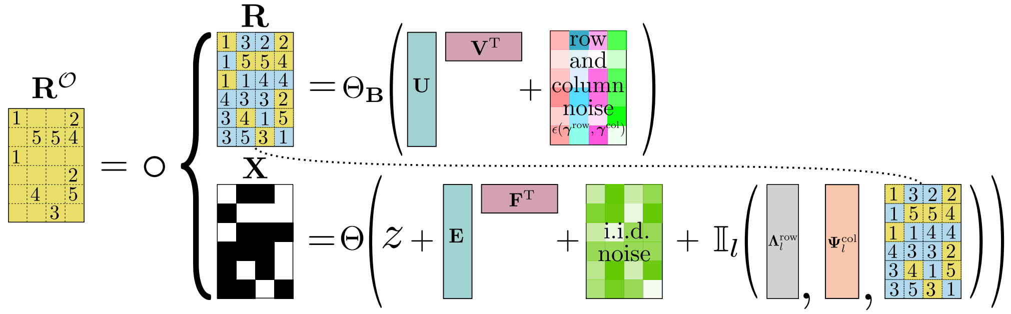

Figure 1. High-level visualization of the components of MF-MNAR. The observed data $\mathbf{R}^{\mathcal{O}}$ is obtained by 'masking' (denoted by the Hadamard product $\circ$) the complete data $\mathbf{R}$ with the binary matrix $\mathbf{X}$. The CDM (top) generates $\mathbf{R}$ by filtering a low rank matrix and heteroskedastic noise through the ordinal likelihood $\Theta_{\mathbf{B}}(\cdot)$ with specific interval boundaries $\mathbf{B}$. The MDM (bottom) generates $\mathbf{X}$ using as input a global bias z, a low rank matrix, i.i.d. noise and the result of combining $\mathbf{R}$ with $\mathbf{\Lambda}^{\text{row}}$ and $\mathbf{\Psi}^{\text{col}}$ to produce an $n \times d$ matrix (indicated by the function $\mathbb{I}_l$). The resulting matrix is filtered through the Heaviside step function Θ.

mentary material. Figure 1 visualizes the generative model of MF-MNAR, which is described in detail below.

3.1. The Complete Data Model

We describe now the CDM in the left half of Figure 2. Full details are in the supplementary material. We propose a matrix factorization model for the rating matrix $\mathbf{R}$. This model has three key features: i) an appropriate ordinal likelihood for rating data (rather than the usual Gaussian likelihood); ii) heteroskedastic noise, that is, variable noise for each user and item; and iii) hierarchical priors to increase robustness to selection of hyper-parameter values.

We assume that $\mathbf{R}$ is generated as a function of two low rank latent matrices $\mathbf{U} \in \mathbb{R}^{n\times h}$ and $\mathbf{V} \in \mathbb{R}^{d\times h}$, where $h \ll \min(n,d)$. Each discrete rating $r_{i,j}$ in $\mathbf{R}$ is determined by i) the scalar $\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$, where $\mathbf{u}_i$ is the vector in the i-th row of $\mathbf{U}$ and $\mathbf{v}_j$ corresponds to the j-th row of $\mathbf{V}$, and ii) a partition of $\mathbb{R}$ into $L-1$ contiguous intervals with boundaries $b_{j,0} < \ldots < b_{j,L}$, where $b_{j,0} = -\infty$ and $b_{j,L} = \infty$. The value of $r_{i,j}$ is obtained from the interval in which $\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$ falls. Note that the interval boundaries are different for each column (item) of $\mathbf{R}$. In practice data is noisy. We model this by adding zero-mean Gaussian noise $\epsilon_{i,j}$ to $\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$ before generating $r_{i,j}$ and introducing the latent variable $a_{i,j} = \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j + \epsilon_{i,j}$. The probability of $r_{i,j}$ given $a_{i,j}$ and $\mathbf{b}_j = (b_{j,1},\ldots,b_{j,L-1})$ is

$$p(r_{i,j}|a_{i,j},\mathbf{b}_j) = \prod_{k=1}^{r_{i,j}-1}\Theta[a_{i,j}-b_{j,k}]\prod_{k=r_{i,j}}^{L-1}\Theta[b_{j,k}-a_{i,j}] = \prod_{k=1}^{L-1}\Theta\left[\text{sign}[r_{i,j}-k-0.5](a_{i,j}-b_{j,k})\right]\,, \tag{4}$$

where Θ is the Heaviside step function. This likelihood is 1 when $a_{i,j} \in (b_{r_{i,j}-1}, b_{r_{i,j}}]$ and 0 otherwise. The dependence of (4) on all the entries in $\mathbf{b}_j$ and not only on $b_{r_{i,j}-1}$ and $b_{r_{i,j}}$ allows us to learn $\mathbf{b}_j$. We put a prior on each $\mathbf{b}_j$ to learn these item specific boundaries, $p(\mathbf{b}_j|\mathbf{b}_0) = \prod_{k=1}^{L-1}\mathcal{N}(b_{j,k}|b_{0,k},v_0)$, where $\mathbf{b}_0$ is a vector of base interval boundaries. To avoid specifying these base boundaries, we use a hyper-prior $p(\mathbf{b}_0) = \prod_{k=1}^{L-1}\mathcal{N}(b_{0,k}|m_k^{\mathbf{b}_0}, v_0)$, where $m_1^{\mathbf{b}_0},\ldots,m_{L-1}^{\mathbf{b}_0}$ and v_0 are hyper-parameters.

Real world rating matrices exhibit variable levels of noise across rows and columns (Lakshminarayanan et al., 2011). To model this heteroskedasticity, we allow the additive noise $\epsilon_{i,j}$ to be row and column dependent: $\epsilon_{i,j}$ has a Gaussian prior with zero-mean and variance $\gamma_i^{\text{row}} \times \gamma_j^{\text{col}}$, where γ_i^{row} and γ_j^{col} govern the noise level in the i-th row and j-th column of $\mathbf{R}$, respectively. Define $c_{i,j} = \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$, then the conditional distribution for $a_{i,j}$ given $c_{i,j}$, γ_i^{row} and γ_j^{col} is $p(a_{i,j}|c_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}|c_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})$. To learn the noise levels we put Inverse Gamma priors on γ_i^{row} and γ_j^{col}.

We use a hierarchical Gaussian prior for the two low-rank matrices $\mathbf{U}$ and $\mathbf{V}$. We select $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}}) = \prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_k^{\mathbf{U}},v_k^{\mathbf{U}})$ and $p(\mathbf{V}|\mathbf{m}^{\mathbf{V}},\mathbf{v}^{\mathbf{V}}) = \prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_k^{\mathbf{V}},v_k^{\mathbf{V}})$, where $\mathbf{m}^{\mathbf{U}}$ and $\mathbf{m}^{\mathbf{V}}$ are mean parameters for the rows of $\mathbf{U}$ and $\mathbf{V}$, respectively. These are given factorized Gaussian hyper-priors. Similarly, $\mathbf{v}^{\mathbf{U}}$ and $\mathbf{v}^{\mathbf{V}}$ are variance parameters for the rows of $\mathbf{U}$ and $\mathbf{V}$ and are given factorized Inverse Gamma hyper-priors. The parameters of these Gaussian hyper-priors are given standard values and the parameters of the Inverse Gamma hyper-priors are in the supplementary material.

We collect the boundary vectors $\mathbf{b}_j$ into the $d \times (L-1)$ matrix $\mathbf{B}$, the $a_{i,j}$ and $c_{i,j}$ variables into the $n \times d$ matrices $\mathbf{A}$ and $\mathbf{C}$, and the row and column noise levels into the n and d dimensional vectors $\boldsymbol{\gamma}^{\text{row}}$ and $\boldsymbol{\gamma}^{\text{col}}$, respectively. The joint distribution for $\mathbf{R}$ and the parameters $\mathbf{\Theta} = \{\mathbf{U},\mathbf{V},\mathbf{B},\mathbf{A},\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}},\mathbf{b}_0,\mathbf{m}^{\mathbf{U}},\mathbf{m}^{\mathbf{V}},\mathbf{v}^{\mathbf{U}},\mathbf{v}^{\mathbf{V}}\}$ is

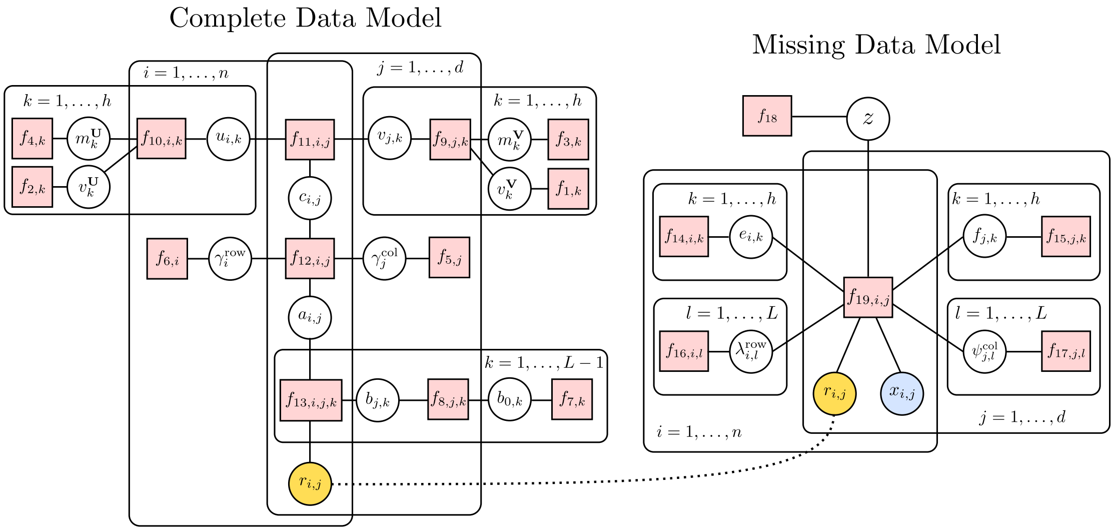

Figure 2. Factor graph for MF-MNAR. The graph includes one component for the CDM (left) and another one for the MDM (right). They are connected through the rating variables $r_{i,j}$, The nodes for $r_{i,j}$ are connected with a dotted line because they encode the same variables. The node for $x_{i,j}$ is shaded in blue because $x_{i,j}$ is always known. This variable indicates whether $r_{i,j}$ is observed. Only a few of the $r_{i,j}$ are observed, to indicate this we have shaded their nodes in orange.

$$
\begin{aligned}
p(\boldsymbol{\Theta}, \mathbf{R}) = p(\mathbf{R}|\mathbf{A}, \mathbf{B}) & p(\mathbf{A}|\mathbf{C}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}}) p(\mathbf{C}|\mathbf{U}, \mathbf{V}) \\
& p(\mathbf{U}|\mathbf{m}^{\mathbf{U}}, \mathbf{v}^{\mathbf{U}}) p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}}) p(\mathbf{B}|\mathbf{b}_0) p(\mathbf{b}_0) \\
& p(\boldsymbol{\gamma}^{\text{row}}) p(\boldsymbol{\gamma}^{\text{col}}) p(\mathbf{m}^{\mathbf{U}}) p(\mathbf{m}^{\mathbf{V}}) p(\mathbf{v}^{\mathbf{U}}) p(\mathbf{v}^{\mathbf{V}}) \,. \qquad (5)
\end{aligned}
$$

This specifies the CDM. We now describe the MDM that explains which entries of $\mathbf{R}$ are contained in the set of observed ratings $\mathcal{D}$, as specified by the binary matrix $\mathbf{X}$.

3.2. The Missing Data Model

The MDM generates the binary matrix $\mathbf{X}$ as a function of $\mathbf{R}$. We also assume a matrix factorization model for $\mathbf{X}$. However, our MDM has two main differences to the CDM. Firstly, we use a likelihood for binary data and secondly, we use additional variables to model the effect of each rating's value $r_{i,j}$ on its observation status $x_{i,j}$.

We define two additional low rank matrices $\mathbf{E} \in \mathbb{R}^{n \times h}$ and $\mathbf{F} \in \mathbb{R}^{d \times h}$. $\mathbf{X}$ is obtained as a function of $\mathbf{E}$, $\mathbf{F}$, $\mathbf{R}$ and some additive noise. In particular, we assume $x_{i,j} = \Theta\{\mathbf{e}_i \mathbf{f}_j^{\text{T}} + z + g_{i,j} + \sum_{l=1}^{L} (\lambda_{i,l}^{\text{row}} + \psi_{j,l}^{\text{col}}) \mathbf{I}[r_{i,j} = l]\}$, where $\Theta\{\cdot\}$ is the Heaviside step function, $\mathbf{I}[\cdot]$ is the binary indicator function, $z \in \mathbb{R}$ is a bias that governs the overall sparsity of $\mathbf{X}$ and $g_{i,j} \in \mathbb{R}$ is an i.i.d. noise variable with a logistic c.d.f. $\sigma(x) = 1/(1 + \exp(-x))$. This generative process results in a logistic likelihood, which is commonly used in binary classification tasks.

The parameters $\lambda_{i,l}^{\text{row}}, \psi_{j,l}^{\text{col}} \in \mathbb{R}$ determine the influence of the value of $r_{i,j}$ on whether $r_{i,j}$ is contained in $\mathcal{D}$ or not. Importantly, this influence can vary across the rows and columns of $\mathbf{R}$. A large positive value of $(\lambda_{i,l}^{\text{row}} + \psi_{j,l}^{\text{col}})$ increases the probability that $r_{i,j}$ is included in $\mathcal{D}$ when $r_{i,j} = l$ and a low value of $(\lambda_{i,l}^{\text{row}} + \psi_{j,l}^{\text{col}})$ reduces this probability. Thus $\lambda_{i,l}^{\text{row}}$ captures effects such as some users i being more likely to rate movies they like, and others rating movies they dislike, while $\psi_{j,l}^{\text{col}}$ captures the analogous effects for movies j. We collect the $\lambda_{i,l}^{\text{row}}$ and the $\psi_{j,l}^{\text{col}}$ in two $n \times L$ and $d \times L$ matrices $\boldsymbol{\Lambda}^{\text{row}}$ and $\boldsymbol{\Psi}^{\text{col}}$. The likelihood for the missing data model is

$$
\begin{aligned}
& p(\mathbf{X}|\mathbf{E}, \mathbf{F}, z, \boldsymbol{\Lambda}^{\text{row}}, \boldsymbol{\Psi}^{\text{col}}, \mathbf{R}) = \\
& \left[\prod_{\substack{(i,j): \\ x_{i,j}=1}} \sigma\{\mathbf{e}_i \mathbf{f}_j^{\text{T}} + z + \sum_{l=1}^{L} (\lambda_{i,l}^{\text{row}} + \psi_{j,l}^{\text{col}}) \mathbf{I}[r_{i,j} = l]\} \right] \\
& \left[\prod_{\substack{(i,j): \\ x_{i,j}=0}} \sigma\{-\mathbf{e}_i \mathbf{f}_j^{\text{T}} - z - \sum_{l=1}^{L} (\lambda_{i,l}^{\text{row}} + \psi_{j,l}^{\text{col}}) \mathbf{I}[r_{i,j} = l]\} \right] . \qquad (6)
\end{aligned}
$$

We use fully factorized standard Gaussian priors for all the parameters in (6). Finally we introduce row and column specific offset biases in $\mathbf{E}$ and $\mathbf{F}$ by setting to one all the entries in one of the columns in $\mathbf{E}$ and in another of the columns of $\mathbf{F}$ (details in the supplementary material).

Let $\boldsymbol{\Omega}$ be the set of variables $\boldsymbol{\Omega} = \{\mathbf{E}, \mathbf{F}, z, \boldsymbol{\Lambda}^{\text{row}}, \boldsymbol{\Psi}^{\text{col}}\}$. The joint distribution for $\mathbf{X}$ and $\boldsymbol{\Omega}$ given $\mathbf{R}$ is

$$
\begin{aligned}
p(\mathbf{X}, \boldsymbol{\Omega}|\mathbf{R}) = & \, p(\mathbf{X}|\mathbf{E}, \mathbf{F}, z, \boldsymbol{\Lambda}^{\text{row}}, \boldsymbol{\Psi}^{\text{col}}, \mathbf{R}) \\
& p(\mathbf{E}) p(\mathbf{F}) p(z) p(\boldsymbol{\Lambda}^{\text{row}}) p(\boldsymbol{\Psi}^{\text{col}}) \,. \qquad (7)
\end{aligned}
$$

3.3. The Joint Model

We obtain MF-MNAR by combining the MDM with the CDM. Recall from Section 2 that $\mathbf{R}^{\mathcal{O}}$ is the set of observed entries in $\mathbf{R}$ and $\mathbf{R}^{\neg\mathcal{O}}$ is the set of non-observed entries. The posterior distribution over the parameters $\mathbf{\Theta}$ of the CDM, the parameters $\mathbf{\Omega}$ of the MDM and the missing data itself $\mathbf{R}^{\neg\mathcal{O}}$ given $\mathbf{X}$ (which ratings are observed) and $\mathbf{R}^{\mathcal{O}}$ (the values of the observed ratings) is

$$p(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}},\mathbf{X}) = \frac{p(\mathbf{X},\mathbf{\Omega}|\mathbf{R})p(\mathbf{\Theta},\mathbf{R})}{p(\mathbf{R}^{\mathcal{O}},\mathbf{X})}\,, \tag{8}$$

where $p(\mathbf{R}^{\mathcal{O}},\mathbf{X})$ is a normalization constant. The factor graph for the resulting model is shown in Figure 2. The graph includes 19 factors, described in the supplementary material. To predict the value of an entry $r_{i,j}$ in $\mathbf{R}^{\neg\mathcal{O}}$, we have to marginalize (8) with respect to $\mathbf{\Omega}$, $\mathbf{\Theta}$ and all of the entries in $\mathbf{R}^{\neg\mathcal{O}}$. This is intractable and we have to use computational approximations. We describe next how to approximate the posterior with a tractable distribution.

4. Approximate Inference

As in most non-trivial Bayesian models, the posterior (8) is intractable. Therefore we approximate this distribution using expectation propagation (EP) (Minka, 2001) and variational Bayes (VB) (Ghahramani & Beal, 2001). We approximate $p(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}})$ with $\mathcal{Q}(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}}) = \mathcal{Q}_1(\mathbf{\Theta})\mathcal{Q}_2(\mathbf{\Omega})\mathcal{Q}_3(\mathbf{R}^{\neg\mathcal{O}})$, where $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$ are given by

$$\begin{aligned}
\mathcal{Q}_1(\mathbf{\Theta}) = &\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}}},a_k^{v^{\mathbf{U}}})\right]\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}}},b_k^{v^{\mathbf{V}}})\right] \\
&\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1}\mathcal{N}(b_{i,k}|m_{i,k}^{b},v_{i,k}^{b})\right]\left[\prod_{i=1}^{n}\prod_{j=1}^{d}\mathcal{N}(a_{i,j}|m_{i,j}^{a},v_{i,j}^{a})\right] \\
&\left[\prod_{i=1}^{n}\prod_{j=1}^{d}\mathcal{N}(c_{i,j}|m_{i,j}^{c},v_{i,j}^{c})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u},v_{i,k}^{u}\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v},v_{j,k}^{v}\right]\left[\prod_{k=1}^{L-1}\mathcal{N}(b_{0,k}|m_k^{b_0},v_k^{b_0})\right] \\
&\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}}},v_k^{m^{\mathbf{U}}})\right]\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}}},v_k^{m^{\mathbf{V}}})\right] \\
&\left[\prod_{i=1}^{n}\mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}},b_i^{\gamma^{\text{row}}})\right]\left[\prod_{j=1}^{d}\mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}}},b_j^{\gamma^{\text{col}}})\right]\,,
\end{aligned} \tag{9}$$

$$\begin{aligned}
\mathcal{Q}_2(\mathbf{\Omega}) = &\,\mathcal{N}(z|m^z,v^z)\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(e_{i,k}|m_{i,k}^{e},v_{i,k}^{e})\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(f_{j,k}|m_{j,k}^{f},v_{j,k}^{f})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\mathcal{N}(\lambda_{i,l}^{\text{row}}|m_{i,l}^{\lambda^{\text{row}}},v_{i,l}^{\lambda^{\text{row}}})\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{L}\mathcal{N}(\psi_{j,l}^{\text{col}}|m_{j,l}^{\psi^{\text{col}}},v_{j,l}^{\psi^{\text{col}}})\right]\,,
\end{aligned} \tag{10}$$

$$\mathcal{Q}_3(\mathbf{R}^{\neg\mathcal{O}}) = \prod_{(i,j)\notin\mathcal{O}}\prod_{l=1}^{L}p_{i,j,l}^{\mathbf{I}[r_{i,j}=l]}\,. \tag{11}$$

Algorithm 1 Approximate Inference in the Joint Model

```
**Input:** Rating dataset $\mathcal{D}$.
Adjust $\mathcal{Q}$ using EP on CDM ignoring MDM.
Adjust $\mathcal{Q}$ using SVI on MDM ignoring CDM.
**for** $t = 1$ **to** $T$ **do**
  Adjust $\mathcal{Q}$ using SVI on MDM with CDM predictions.
  Adjust $\mathcal{Q}$ using EP-SVI on CDM with MDM predictions.
**end for**
**Output:** Posterior approximation $\mathcal{Q}$.
```

The parameters of $\mathcal{Q}$ are fixed by running EP and VB on the complete data model (CDM) and VB on the missing data model (MDM). We use EP for the CDM as it performs well in ordinal regression (Chu & Ghahramani, 2005). However, EP is known to perform poorly for factors corresponding to matrix factorizations, so we use VB for these. In particular, we use a stochastic version of VB called stochastic variational inference (SVI) (Hoffman et al., 2013) so that our computational cost scales with the number of observed ratings $|\mathcal{O}|$ and not with the size of $\mathbf{R}$, which can be very large. Algorithm 1 summarizes our inference procedure. We first adjust the components in $\mathcal{Q}$ for the two models independently. After that, we co-train the models. For this, we iteratively re-adjust $\mathcal{Q}$ by jointly refining the approximate posterior of each model while taking into account the predictions of the other. Full details can be found in the supplementary material. The code for our MF-MNAR method is publicly available at `http://jmhl.org`.

4.1. Predictive Distribution in the Joint Model

Given $\mathcal{Q}$, we can approximate the joint model's posterior probability $p_{i,j,l}^{\text{JM}}(x_{i,j})$ that the entry in the i-th row and j-th column of $\mathbf{R}$ takes value l, conditioned on a specific value of $x_{i,j}$. When $x_{i,j} = 0$, we assume that the entry was not selected by the MDM and it is missing. When $x_{i,j} = 1$, the entry was selected by the MDM and should have been observed, but its value is unknown (for example, if it was held out in a test set). The value of $p_{i,j,l}^{\text{JM}}(x_{i,j})$ is approximated using

$$\tilde{p}_{i,j,l}^{\text{JM}}(x_{i,j}) \propto \tilde{p}_{i,j,l}^{CDM}\tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})\,, \tag{12}$$

where $\tilde{p}_{i,j,l}^{CDM}$ and $\tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})$ are the individual predictions of the CDM and the MDM generated using the approximate posterior $\mathcal{Q}$, respectively, and their exact values can be found in the supplementary material. In this formula, $\tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})$ improves the prediction of the CDM, $\tilde{p}_{i,j,l}^{CDM}$, by using the information available in $x_{i,j}$.

5. Related Work

Marlin & Zemel (2009) have proposed a method for modeling MNAR rating data in collaborative filtering. They use a Mixture of Multinomials (MM) as their CDM. In the MM model, there are K clusters of users and each item has K multinomial rating distributions associated to it. To rate item j, all the users in the k-th cluster sample from the k-th multinomial distribution associated with the j-th item. The EM algorithm is used to adjust the MM model with MAR data. This MM CDM is simple enough to be extended to MNAR data at a low computational cost. However, in practice, the MM model lacks flexibility and is often outperformed by more powerful MF approaches such as the MF method described in Section 3.1.

Marlin & Zemel (2009) propose two different missing data models for the MM CDM. The first, CPTv, assumes that $p(x_{i,j}|r_{i,j}) = \mu_{r_{i,j}}$, where $\boldsymbol{\mu}$ is an L-dimensional probability vector such that μ_l is the probability that $r_{i,j}$ is observed if its value is l. The second MDM, Logitvd, assumes that $p(x_{i,j}|r_{i,j}) = \sigma(\kappa_{r_{i,j}} + \omega_j)$, where σ is the logistic function, $\omega_j \in \mathbb{R}$ models a per-item bias and the parameters $\kappa_1, \ldots, \kappa_L \in \mathbb{R}$ model the dependence of $x_{i,j}$ on the value of $r_{i,j}$. These MDMs can be jointly estimated with the MM CDM using EM. The resulting method is computationally efficient and has approximately the same cost as learning the MM under the MAR assumption.

Our MDM (see Section 3.2) is more flexible than CPTv or Logitvd. The variables Λ^{row} and Ψ^{col} allow us to encode dependencies between $r_{i,j}$ and $x_{i,j}$ that can change across users, items and values of $r_{i,j}$. For example, we can capture effects such as groups of users with different rating behaviors: users that rate what they like and others that rate what they dislike; or certain items that are only rated by users who dislike them. Furthermore, we use a MF component to capture global effects that are independent of $\mathbf{R}$. For example, a set of items that are strongly promoted and hence observed with high frequency. We also do full Bayesian inference instead of MAP estimation as in Logitvd and CPTv. This makes our MDM robust to overfitting problems. Finally, we obtain the same scalability as CPTv and Logitvd by using stochastic inference methods (see Section 4 and the supplementary material).

MNAR rating data is also considered by Steck (2010). This method is similar to the BPR algorithm (Rendle et al., 2009) and works by optimizing the parameters of a non-probabilistic MF model with respect to a ranking-based loss function. Steck (2010) does not learn a generative model for the data and optimizes a metric that is robust to MNAR data. The resulting method can be applied to recommendation, but not to rating prediction tasks.

The previous pioneering works have addressed the problem of modeling MNAR rating data. However, these early approaches are limited to relatively simple models. Bayesian MF models are highly flexible and often yield state-of-the-art predictive performance on rating data. We present the first approach to extend these methods to the MNAR scenario.

6. Experiments

We analyze the performance of our Matrix Factorization model with data Missing Not At Random (MF-MNAR) in a series of experiments with synthetic and real-world rating data. We compare MF-MNAR with several benchmark methods including i) a version of MF-MNAR that assumes data Missing At Random (MF-MAR); ii) a Mixture of Multinomials model for MAR rating data (MM-MAR) (Marlin & Zemel, 2009); the iii) CPTv and iv) Logitvd models for MNAR rating data proposed by Marlin & Zemel (2009); v) a state-of-the-art method for ordinal matrix factorization with MAR data (Paquet et al., 2012) (Paquet) and finally, vi) an oracle method that always predicts labels according to their empirical frequencies in the test set (Oracle). In all MF methods we use a latent rank of size 20. In the mixture of multinomials we use 20 components.

6.1. Datasets

Our first synthetic dataset is a toy example that illustrates the differences between rating data *missing at random* (MAR) and rating data *missing not at random* (MNAR) (Steck, 2010). The dataset contains items that are horror movies (items 1 to 50) or romance movies (items 51 to 100) and users who are romance-lovers (users 1 to 100 and 201 to 300) or horror-lovers (users 101 to 200 and 301 to 400). We consider discrete ratings with values from 1 to 5. Romance-lovers (horror-lovers) will rate romance movies (horror movies) by sampling from a multinomial with probability vector $\mathbf{p} = (0.05, 0.05, 0.05, 0.4, 0.45)$, where p_i is the probability of rating the movie with value i. Similarly, romance-lovers (horror-lovers) will rate horror movies (romance movies) using a multinomial with probability vector $\mathbf{p}'$, where $p'_i = p_{6-i}$. The left-hand plot in Figure 3 shows the complete rating matrix for this dataset. We have considered two missing data mechanisms. In the first (MAR), each matrix entry is observed independently with probability $p \approx 0.23$. In the second (MNAR), users 1 to 200 are 'positive', they tend to rate what they like. Each rating from these users is observed with probability p_i, where i is the value of the entry. However, users 201 to 400 are 'negative', rating what they do *not* like. Each rating with value i is now observed with probability p'_i. The plot in the middle of Figure 3 shows the observed data for the MAR setting, while the right-hand plot shows the observed data for the MNAR setting. In the latter case, there is a clear

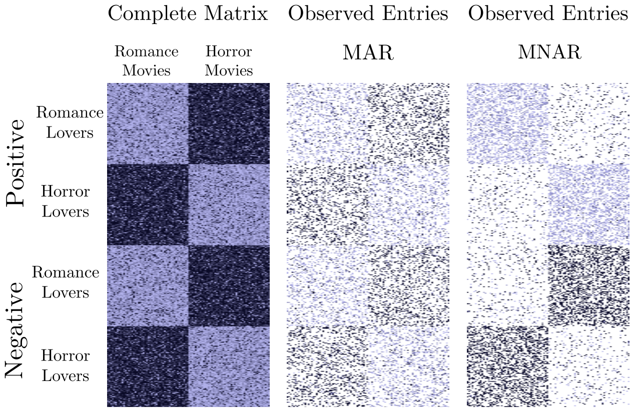

Figure 3. Toy synthetic dataset that illustrates the differences between MAR and MNAR. Left, full 400×100 rating matrix. Light blue colors correspond to high rating values, while dark blue colors correspond to low rating values. Middle, observed data under the MAR setting. Right, observed data under the MNAR setting.

pattern relating the value of an entry with its observation status. Note that the pattern changes from rows 1-200 to rows 201-400.

The second synthetic dataset is obtained by sampling from a matrix factorization model. We generate two 500×10 matrices $\mathbf{U}$ and $\mathbf{V}$ with standard Gaussian i.i.d. entries. We then generate $\mathbf{C} = \mathbf{U}\mathbf{V}^\text{T} - 3$ and partition $\mathbb{R}$ in contiguous intervals with boundaries $-\infty, -6, -2, 2, 6, \infty$. For each $c_{i,j}$, we create an $r_{i,j} \in \{\, 1, \ldots, 5\}$ according to the interval in which $c_{i,j}$ lies. We generate two extra 500×10 matrices $\mathbf{E}$ and $\mathbf{F}$ with standard Gaussian i.i.d. entries. For each $r_{i,j}$ we set the binary variable $x_{i,j}$ to 1 with probability $\sigma(\mathbf{e}_i\mathbf{f}_j^\text{T} - 4 + \sum_{l=1}^5 z_i \mathbf{I}[r_{i,j} = l])$, where σ is the logistic function and $\mathbf{e}_i$ and $\mathbf{f}_j$ are the i-th and j-th rows of $\mathbf{E}$ and $\mathbf{F}$. We consider two missing data mechanisms. In the first (MAR), all the $z_1, \ldots, z_5$ are 0 when we generate the $x_{i,j}$. In the second (MNAR), $(z_1, \ldots, z_5) = (-3, -3, -3, 3, 3)$. In the latter case we expect to see many more ratings with value 4 and 5 than in the MAR scenario.

We considered several real-world rating datasets. The MovieLens 100k and 1M datasets[1] include ratings from 1 to 5 on movies. The Yahoo! Web-scope R3 dataset[2]. contains ratings from 1 to 5 on songs. We also analyzed a dataset obtained from the reviewer bidding process for the 2013 NIPS conference. This dataset contains ratings from 1 to 4 given by reviewers on papers. Finally, the Movie Tweetings dataset[3] includes ratings on movies collected from Twitter. We pre-processed this dataset to map the original ratings from 0 to 10 to the interval $[1, 5]$ using the rule $r_\text{mapped} = \lceil r_\text{original}/2 \rceil$ except for 0, which was mapped to 1. Furthermore, because this dataset is very sparse, we only kept the users and movies that have at least 10 ratings. Table 1 shows a summary of the datasets. Note that only the Yahoo and the synthetic datasets include data entries with $x_{i,j} = 0$. These entries were collected for the Yahoo dataset by eliciting ratings on items selected randomly, not by the users themselves. In all the other real-world datasets we only have the entries from the rating matrix $\mathbf{R}$ that are selected by the missing data mechanism, that is, entries for which $x_{i,j} = 1$.

Table 1. Characteristics of the datasets.

Dataset	n	d	$\#x_{i,j} = 1$	$\#x_{i,j} = 0$	**Sparsity**
ML100K	943	1682	100000	0	0.063
ML1M	6040	3706	1000209	0	0.045
MTweet	3972	2043	106337	0	0.013
NIPS	1053	1381	52721	0	0.036
Yahoo	15400	1000	311704	54000	0.020
SMF-MNAR	500	500	12234	237766	0.049
SMF-MAR	500	500	31378	218622	0.126
SRH-MNAR	400	100	9057	30943	0.226
SRH-MAR	400	100	9258	30742	0.231

SMF: Synthetic Matrix Factorization dataset.
SRH: Synthetic Romance Horror movie dataset.

6.2. Experimental Protocol

For each dataset, we collect the available rating entries $r_{i,j}$ with $x_{i,j} = 0$ in a *special* test set; this set is only available for the Yahoo and synthetic datasets. After that, we randomly split the observed ratings (the $r_{i,j}$ with $x_{i,j} = 1$) into a training set with 99% of the ratings and a *standard* test set with the remaining 1%. We use small standard test sets to avoid interfering with the actual missing data mechanism. When an $r_{i,j}$ with $x_{i,j} = 1$ is added to the standard test set we fix that $x_{i,j}$ to zero to indicate that the entry $r_{i,j}$ is not observed. Each method is adjusted on each training set and then evaluated on the corresponding standard and special test sets by computing the average predictive log-likelihood of the ratings. The whole process is repeated 40 times. The supplementary material contains results for other performance metrics: root mean squared error, mean absolute error, and predictive accuracy.

The previous procedure focuses mainly on evaluating the performance of the complete data model. However, we are also interested in evaluating the missing data model. For this we try to find the $x_{i,j}$ that initially took value one but were flipped to zero during the generation of the standard test sets. We use recall at N to measure performance as this is a popular metric in recommendation tasks (Gunawardana & Shani, 2009). For each row i of $\mathbf{X}$, we use the missing data model to compute the probability that the variables $x_{i,j}$ with value zero in that row originally took value one. We select the top $N = 10$ variables with highest probability and the fraction of variables $x_{i,j}$ with original value one that are recovered and average the recall across rows.

[1] http://grouplens.org/datasets/movielens/

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

[3] https://github.com/sidooms/MovieTweetings

Table 2. Average Log-likelihood on the Standard Test Sets.

	MF	**MF**	**MM**	**CTPv**	**Logitvd**	**Paquet**	**Oracle**
Dataset	**MNAR**	**MAR**	**MAR**	**MNAR**	**MNAR**	**MAR**	
ML100K	**-1.181**	-1.186	-1.471	-1.463	-1.425	-1.218	-1.468
ML1M	**-1.121**	-1.125	-1.308	-1.436	-1.380	-1.162	-1.456
MTweet	**-0.941**	-0.946	-1.105	-1.245	-1.141	-0.997	-1.235
NIPS	**-0.937**	-0.956	-1.204	-1.170	-1.167	-0.995	-1.329
Yahoo	**-1.172**	-1.204	-1.278	-1.399	-1.304	-1.218	-1.551
SMF-MNAR	**-0.902**	-0.937	-1.447	-1.336	-1.326	-1.000	-1.331
SMF-MAR	-0.425	**-0.417**	-1.327	-1.238	-1.235	-0.510	-1.198
SRH-MNAR	-1.055	-1.067	-0.987	**-0.962**	-0.963	-1.143	-1.392
SRH-MAR	-1.317	-1.287	-1.272	**-1.265**	-1.266	-1.318	-1.498

6.3. Results

Table 2 shows the average test log-likelihood obtained by each method on the standard rating test sets. The best performing method is highlighted in bold and the statistically indistinguishable results, according to a paired t-test, are underlined. MF-MNAR is the best method on the real-world datasets and on the SMF-MNAR dataset. Furthermore, in accordance with intuition, MF-MNAR is better than MF-MAR on the synthetic datasets with MNAR data, while the opposite result occurs on the synthetic datasets with MAR data. The models based on mixtures of multinomials, MM, CPTv and Logitvd, obtain the best performance on the SRH-MNAR and SRH-MNAR datasets. This is because these datasets were generated by sampling from multinomials, as assumed by these models. The oracle method is generally outperformed by most methods since it predicts the same values for all entries. Finally, Paquet's method is less accurate than MF-MNAR and MF-MAR.

Table 3 shows the average recall obtained by the missing data models of those methods that assume MNAR data. In this table we have also included the results obtained by the missing data model of MF-MNAR (MDM) without coupling it with the complete data model. This allows us to evaluate the gains produced in MDM by using the predictions of the complete data model. We also include the results of a baseline (Freq) that, for each row i of $\mathbf{R}$, ranks the variables $x_{i,1}, \ldots, x_{i,d}$ with value zero by their empirical frequency across rows. The missing data model of MF-MNAR obtains the best results in all cases, except again intuitively on the SMF-MAR and SHR-MAR datasets, where the MAR assumption is appropriate. Overall, all the methods outperform Freq except CPTv, which performs worst. Our sample of the SHR-MAR dataset seems to be particularly suited to Logitvd, which performs best in this case.

Finally, Table 4 shows the average test log-likelihood of each technique on the special test sets for predicting ratings when $x_{i,j} = 0$. In this case, the Oracle baseline obtains the best results on most datasets with MNAR data, assuming that the Yahoo dataset is MNAR. The Oracle method always makes the same probabilistic prediction for each test entry. This shows the difficulty of making accurate predictions on MNAR data. Despite this, MF-MNAR is better than MF-MAR in all of the datasets with MNAR data, except on the real-world Yahoo dataset, where MF-MAR performs better. In the MAR datasets, MF-MAR is better than MF-MNAR, as expected. Note that Logitvd and CPTv do not produce any improvement on the SRH-MNAR dataset with respect to MM. This is because these models make incorrect assumptions about the missing data mechanism that was used to generate this dataset.

Table 3. Average Recall on the Standard Test Sets.

	MF	**MDM**	**CTPv**	**Logitvd**	**Freq**
Dataset	**MNAR**		**MNAR**	**MNAR**	
ML100K	**0.299**	0.295	0.093	0.130	0.119
ML1M	0.204	**0.204**	0.041	0.068	0.077
MTweet	**0.203**	0.199	0.127	0.142	0.143
NIPS	**0.309**	0.309	0.011	0.009	0.013
Yahoo	**0.285**	0.283	0.145	0.198	0.182
SMF-MNAR	**0.300**	0.280	0.038	0.052	0.051
SMF-MAR	0.438	**0.445**	0.038	0.051	0.047
SRH-MNAR	**0.246**	0.245	0.209	0.157	0.121
SRH-MAR	0.113	0.115	0.134	**0.143**	0.131

Table 4. Average Log-likelihood on the Special Test Sets.

	MF	**MF**	**MM**	**CPTv**	**Logitvd**	**Paquet**	**Oracle**
Dataset	**MNAR**	**MAR**	**MAR**	**MNAR**	**MNAR**	**MAR**	
Yahoo	-1.578	-1.566	-1.558	-1.277	-1.499	-1.457	**-1.227**
SMF-MNAR	-1.215	-1.287	-1.705	-1.770	-1.724	-1.182	**-1.152**
SMF-MAR	-0.446	**-0.439**	-1.330	-2.003	-1.957	-0.535	-1.201
SRH-MNAR	-1.682	-1.718	**-1.531**	-2.335	-2.320	-1.755	-1.531
SRH-MAR	-1.298	-1.286	**-1.263**	-1.393	-1.403	-1.318	-1.515

7. Conclusions

We have presented the first practical implementation of a probabilistic matrix factorization (MF) model for ordinal matrix data with entries missing not at random (MF-MNAR). The missing data model in MF-MNAR is a MF model for binary matrices in which the observation probability of a matrix entry depends on the entry's value, with different dependence strengths across rows and across columns. The complete data model in MF-MNAR is an ordinal MF method that generates state-of-the-art predictions on rating data on its own. Approximate Bayesian inference in MF-MNAR is implemented using expectation propagation and variational Bayes. We achieve scalability to large datasets by using stochastic inference methods that randomly sub-sample missing matrix entries. The combination of the missing and complete data models in MF-MNAR produces gains in both the modeling of the missing data mechanism and the modeling of the ordinal ratings.

Acknowledgements NMTH is a recipient of the Google Europe Fellowship in Statistical Machine Learning. JMH acknowledges support from Infosys Labs, Infosys Limited and from the Spanish Dirección General de Investigación, project ALLS (TIN2010-21575-C02-02).

References

Chu, Wei and Ghahramani, Zoubin. Gaussian processes for ordinal regression. In *Journal of Machine Learning Research*, pp. 1019–1041, 2005.

Ghahramani, Z. and Beal, M. J. *Advanced Mean Field Method—Theory and Practice*, chapter Graphical models and variational methods, pp. 161–177. 2001.

Gunawardana, Asela and Shani, Guy. A survey of accuracy evaluation metrics of recommendation tasks. *The Journal of Machine Learning Research*, 10:2935–2962, 2009.

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Kschischang, Frank R., Frey, Brendan J., and Loeliger, H.-A. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

Lakshminarayanan, Balaji, Bouchard, Guillaume, and Archambeau, Cedric. Robust Bayesian matrix factorisation. In *International Conference on Artificial Intelligence and Statistics*, pp. 425–433, 2011.

Little, R.J.A. and Rubin, D.B. *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1987.

Marlin, Benjamin M. and Zemel, Richard S. Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

Marlin, Benjamin M. and Zemel, Richard S. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pp. 5–12, 2009.

Minka, Thomas P. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, pp. 362–369, 2001.

Mnih, Andriy and Salakhutdinov, Ruslan. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pp. 1257–1264, 2007.

Paquet, Ulrich, Thomson, Blaise, and Winther, Ole. A hierarchical model for ordinal matrix factorization. *Statistics and Computing*, 22(4):945–957, 2012.

Rendle, Steffen, Freudenthaler, Christoph, Gantner, Zeno, and Schmidt-Thieme, Lars. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.

Steck, Harald. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pp. 713–722, 2010.

Stern, David H, Herbrich, Ralf, and Graepel, Thore. Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, pp. 111–120. ACM, 2009.