Cold-start Active Learning with Robust Ordinal Matrix Factorization

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

Abstract

We present a new matrix factorization model for rating data and a corresponding active learning strategy to address the *cold-start* problem. Cold-start is one of the most challenging tasks for recommender systems: what to recommend with new users or items for which one has little or no data. An approach is to use *active learning* to collect the most useful initial ratings. However, the performance of active learning depends strongly upon having accurate estimates of i) the uncertainty in model parameters and ii) the intrinsic noisiness of the data. To achieve these estimates we propose a heteroskedastic Bayesian model for ordinal matrix factorization. We also present a computationally efficient framework for Bayesian active learning with this type of complex probabilistic model. This algorithm successfully distinguishes between informative and noisy data points. Our model yields state-of-the-art predictive performance and, coupled with our active learning strategy, enables us to gain useful information in the cold-start setting from the very first active sample.

1. Introduction

Collaborative filtering (CF) based recommender systems exploit shared regularities in people's behavior to learn about entities such as users and items. The patterns learned can then be used to make predictions and decisions such as recommending new items to a user. However, CF methods can perform poorly when new users or items are introduced and the amount of data available for such entities is minimal. This scenario is referred to as the *cold-start* problem (Maltz & Ehrlich, 1995; Schein et al., 2002). One solution to the cold-start problem is to use features (e.g. demographic information) to improve predictive performance (Claypool et al., 1999; Park et al., 2006; Ahn, 2008; Park & Chu, 2009). However, such features may not be available, e.g. for privacy reasons. A complementary strategy is then to collect additional ratings so that the system learns as much as possible about the new entities from a minimal number of user interactions. This is an instance of *active learning* (Settles, 2010).

We address the cold-start problem with a Bayesian approach to active learning. Bayesian methods are becoming increasingly popular for CF for several reasons: i) they exhibit strong predictive performance, ii) they can deal formally with missing data and iii) they provide uncertainty estimates for predictions and parameter values (Salakhutdinov & Mnih, 2008; Stern et al., 2009; Paquet et al., 2012; Marlin & Zemel, 2009). This last property is particularly important for the success of active learning. Obtaining correct estimates of uncertainty in both the model parameters and the noise levels is essential for identifying the most informative data to collect. This is especially relevant in cold-start learning as parameters relating to the new entity are highly uncertain. To achieve good models of rating uncertainty we propose a new probabilistic model for rating data. This model allows us to encode uncertainty both through a posterior distribution over the parameters and a likelihood function with different noise levels across users and items (heteroskedasticity). We demonstrate superior performance of this model on several rating datasets relative to current state-of-the-art alternatives.

A drawback of Bayesian approaches to active learning is that they can be computationally expensive. They often require computing the expected change in posterior parameter uncertainty for every candidate data instance *yet to be* collected. Most approaches speed up the process either by approximating the required posterior entropies directly (MacKay, 1992; Lewi et al., 2007) or by using heuristics such as selecting the data for which the current predictions are the most uncertain (maximum entropy sampling)

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[1]Equal contributors.

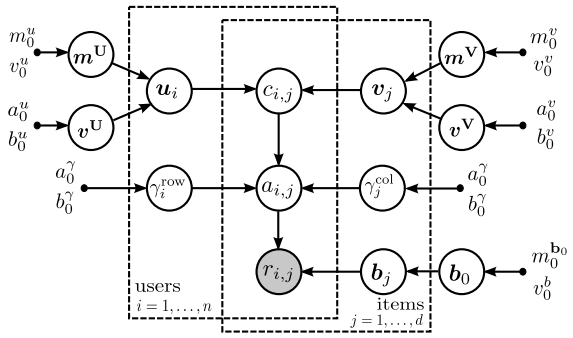

Figure 1. Graphical model for the robust method for ordinal matrix data as it is described in the main text. The observed variables are the rating values $r_{i,j}$. All the other variables are latent. Dots denote fixed hyper-parameters.

(Shewry & Wynn, 1987). As an alternative, we extend a new active learning framework (Bayesian Active Learning by Disagreement or BALD) to the cold-start CF scenario (Houlsby et al., 2012). With this framework we can compute the expected change in posterior uncertainty accurately and we only need to re-compute the posterior *after* collecting new ratings.

In cold start learning it is critical to gain maximal information from the very first sample so as not to deter a new user with multiple requests for information. We find that, after collecting a single rating with BALD, random sampling and maximum entropy sampling require 60% and 85% more data, respectively, to achieve the same predictive performance. An increase from one to two initial rating requests may be critical to whether a user stays with the system. In summary, we propose a complete Bayesian framework to address the cold-start problem in recommender systems. This includes a new heteroskedastic model for ordinal matrix factorization that accurately estimates uncertainty and the intrinsic noisiness of the data, and a computationally efficient algorithm for Bayesian active learning with this model.

2. A Robust Model for Ordinal Matrix Data

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 rating values are ordinal, $1 < \ldots < L$, for example, 1 to L 'stars' assigned to a product. $\mathcal{O}$ denotes the set of pairs of users and items for which a rating is available (observed). We assume that the dataset $\mathcal{D}$ is a sample from a full $n \times d$ rating matrix $\mathbf{R}$, where the entry $r_{i,j}$ in the i-th row and j-th column of $\mathbf{R}$ contains the i-th user's rating for the j-th item. In practice, $\mathcal{D}$ contains only a small fraction of the entries in $\mathbf{R}$.

We propose a new probabilistic model for $\mathbf{R}$ that allows the noise levels to vary across rows and columns of $\mathbf{R}$, providing robustness. This is particularly important for active learning, where collecting data from users or items that are too noisy is wasteful. To capture the discrete nature and natural ordering of rating data, our model takes an ordinal regression approach (Chu & Ghahramani, 2005; Stern et al., 2009). This is an advantage over the more common Gaussian likelihood that inappropriately assumes continuous entries in $\mathbf{R}$. To obtain better predictions, we learn different threshold parameters in the ordinal likelihood for each column (item) of $\mathbf{R}$. The model also has a low rank matrix factorization with a hierarchical prior on the latent factors. The hierarchical prior allows us to avoid specifying hyper-parameter values and increases robustness to overfitting. The graphical model for this new probabilistic method is shown in Figure 1.

2.1. Model Description

We now describe our probabilistic model, additional details are in the supplementary material. We model the generation of $\mathbf{R}$ 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 $h \ll \min(n, d)$. $r_{i,j}$ 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 $\mathbf{v}_j$ is the j-th row of $\mathbf{V}$, and ii) a partition of the real line into $L-1$ contiguous intervals with thresholds, or boundaries, $b_{j,0} < \ldots < b_{j,L}$, with $b_{j,0} = -\infty$ and $b_{j,L} = \infty$. The value of $r_{i,j}$ is a function of the interval in which $\mathbf{u}_i^{\mathrm{T}} \mathbf{v}_j$ lies. Note that the interval boundaries are different for each column of $\mathbf{R}$. A simple model would be $r_{i,j} = l$ if $\mathbf{u}_i^{\mathrm{T}} \mathbf{v}_j \in (b_{j,l-1}, b_{j,l}]$. However, in practice, due to noise there may be no $b_{j,0}, \ldots, b_{j,L}$, $\mathbf{U}$ and $\mathbf{V}$ that guarantee $\mathbf{u}_i^{\mathrm{T}} \mathbf{v}_j \in (b_{j,r_{i,j}-1}, b_{j,r_{i,j}}]$ for all ratings in $\mathcal{D}$. Therefore, we add zero-mean Gaussian noise $e_{i,j}$ to $\mathbf{u}_i^{\mathrm{T}} \mathbf{v}_j$ before generating $r_{i,j}$ and introducing the latent variable $a_{i,j} = \mathbf{u}_i^{\mathrm{T}} \mathbf{v}_j + e_{i,j}$. The probability of $r_{i,j}$ given $a_{i,j}$ and $\mathbf{b}_j = (b_{j,1}, \ldots, b_{j,L-1})$ is

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

Θ denotes the step function, $\Theta[x] = 1$ for $x \geq 0$ and 0 otherwise. Thus, the likelihood (1) takes value 1 when $a_{i,j} \in (b_{r_{i,j}-1}, b_{r_{i,j}}]$ and 0 otherwise. Note that the dependence of (1) on all the entries in $\mathbf{b}_j$, not just the neighboring boundaries, allows us to learn the value of $\mathbf{b}_j$ whilst guaranteeing that $b_{j,0} < \ldots < b_{j,L}$. We put a hierarchical Gaussian prior on the boundary variables $\mathbf{b}_j$, $p(\mathbf{b}_j | \mathbf{b}_0) = \prod_{k=1}^{L-1} \mathcal{N}(b_{j,k} | b_{0,k}, v_0)$, where $\mathbf{b}_0$ are base interval boundaries, with prior $p(\mathbf{b}_0) = \prod_{k=1}^{L-1} \mathcal{N}(b_{0,k} | m_k^{\mathbf{b}_0}, v_0^b)$. $m_1^{\mathbf{b}_0}, \ldots, m_{L-1}^{\mathbf{b}_0}$ and v_0^b are hyper-parameters.

We include heteroskedasticity in the additive noise $e_{i,j}$, that

is, the noise level varies across users and items. To do this we put a a zero-mean Gaussian distribution on $e_{i,j}$ with variance $\gamma_i^{\text{row}}\gamma_j^{\text{col}}$. Thus γ_i^{row} and γ_j^{col} are factors that specify the noise level in the i-th row and j-th column of $\mathbf{R}$, respectively. We define $c_{i,j} = \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$ and assume that the conditional distribution of $a_{i,j}$ 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 user and item specific noise levels we put Inverse Gamma priors on γ_i^{row} and γ_j^{col}.

For robustness to fixing hyper-parameter values, we use a hierarchical Gaussian prior for $\mathbf{U}$ and $\mathbf{V}$, $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}}) = \prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_k^{\mathbf{U}},v_k^{\mathbf{U}})$ 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 $\mathbf{m}^{\mathbf{U}},\mathbf{m}^{\mathbf{V}}$ are mean parameters for the rows of $\mathbf{U},\mathbf{V}$, respectively, and are given factorized standard Gaussian priors. $\mathbf{v}^{\mathbf{U}},\mathbf{v}^{\mathbf{V}}$ are variance parameters for the rows of $\mathbf{U},\mathbf{V}$ and are given factorized Inverse Gamma priors.

Finally, let $\mathbf{C}$ denote the set of variables $c_{i,j}$ for which $r_{i,j}$ is observed, then $p(\mathbf{C}|\mathbf{U},\mathbf{V}) = \prod_{(i,j)\in\mathcal{O}}\delta(c_{i,j} - \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j)$. Similarly we collect the variables $a_{i,j}$ into $\mathbf{A}$, and the boundary variables $\mathbf{b}_j$ into a $d\times(L-1)$ matrix $\mathbf{B}$. $\mathbf{R}^{\mathcal{O}}$ denotes the set of entries in $\mathbf{R}$ that are observed. The likelihood factorizes as $p(\mathbf{R}^{\mathcal{O}}|\mathbf{A},\mathbf{B}) = \prod_{(i,j)\in\mathcal{O}}p(r_{i,j}|a_{i,j},\mathbf{b}_j)$. Given $\mathbf{R}^{\mathcal{O}}$, the posterior distribution over all of the variables $\mathbf{\Xi} = \{\mathbf{U},\mathbf{V},\mathbf{B},\mathbf{A},\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}},\mathbf{b}_0,\mathbf{m}^{\mathbf{U}},\mathbf{m}^{\mathbf{V}},\mathbf{v}^{\mathbf{U}},\mathbf{v}^{\mathbf{V}}\}$ is

$$
\begin{aligned}
&p(\mathbf{\Xi}|\mathbf{R}^{\mathcal{O}}) = \\
&p(\mathbf{R}^{\mathcal{O}}|\mathbf{A},\mathbf{B})p(\mathbf{A}|\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}})p(\mathbf{C}|\mathbf{U},\mathbf{V})p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}}) \\
&p(\mathbf{V}|\mathbf{m}^{\mathbf{V}},\mathbf{v}^{\mathbf{V}})p(\mathbf{B}|\mathbf{b}_0)p(\mathbf{b}_0)p(\boldsymbol{\gamma}^{\text{row}})p(\boldsymbol{\gamma}^{\text{col}}) \\
&p(\mathbf{m}^{\mathbf{U}})p(\mathbf{m}^{\mathbf{V}})p(\mathbf{v}^{\mathbf{U}})p(\mathbf{v}^{\mathbf{V}})[p(\mathbf{R}^{\mathcal{O}})]^{-1}\,, \qquad (2)
\end{aligned}
$$

where $p(\mathbf{R}^{\mathcal{O}})$ is the normalization constant (conditioning on hyper-parameters has been omitted for clarity). The hyper-parameters values are in the supplementary material.

2.2. Inference

As with most non-trivial models, the posterior (2) is intractable. Therefore, we approximate this distribution using expectation propagation (EP) (Minka, 2001) and variational Bayes (VB) (Ghahramani & Beal, 2001). We use the following parametric approximation to the exact posterior:

$$
\begin{aligned}
&\mathcal{Q}(\mathbf{\Xi}) = \\
&\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1}\mathcal{N}(b_{i,k}|m_{i,k}^{b},v_{i,k}^{b})\right]\left[\prod_{(i,j)\in\mathcal{O}}\mathcal{N}(a_{i,j}|m_{i,j}^{a},v_{i,j}^{a})\right] \\
&\left[\prod_{(i,j)\in\mathcal{O}}\mathcal{N}(c_{i,j}|m_{i,j}^{c},v_{i,j}^{c})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u},v_{i,k}^{u})\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v},v_{j,k}^{v})\right]\left[\prod_{k=1}^{L-1}\mathcal{N}(b_{0,k}|m_k^{b_0},v_k^{b_0})\right] \\
&\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}}},v_k^{m^{\mathbf{U}}})\right]\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}}},v_k^{m^{\mathbf{V}}})\right] \\
&\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}}},a_k^{v^{\mathbf{U}}})\right]\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}}},b_k^{v^{\mathbf{V}}})\right] \\
&\left[\prod_{i=1}^{n}\mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}},b_i^{\gamma^{\text{row}}})\right]\left[\prod_{j=1}^{d}\mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}}},b_j^{\gamma^{\text{col}}})\right]. \qquad (3)
\end{aligned}
$$

The parameters on the right hand side of Equation (3) are learned by running a combination of EP and VB. We choose EP as our main workhorse for inference because it has shown good empirical performance in the related problem of binary classification (ordinal regression with only 2 rating values) (Nickisch & Rasmussen, 2008). However, it is well known that for factors corresponding to the matrix factorizations, EP provides poor approximations (Stern et al., 2009), so for these we use VB. Implementational details are in the supplementary material.

2.3. Predictive Distribution

Given the approximation to the posterior in (3) we estimate the predictive probability of a new entry $r_{i,j}^{\star}$ in $\mathbf{R}$ that is not contained in the observed ratings $\mathbf{R}^{\mathcal{O}}$ using

$$
\begin{aligned}
\mathcal{P}(r_{i,j}^{\star}|\mathbf{R}^{\mathcal{O}}) &\approx \int p(r_{i,j}^{\star}|a_{i,j}^{\star},\mathbf{b}_j)p(a_{i,j}^{\star}|c_{i,j}^{\star},\gamma_i^{\text{row}},\gamma_j^{\text{col}}) \\
&\quad p(c_{i,j}^{\star}|\mathbf{u}_i,\mathbf{v}_j)\mathcal{Q}(\mathbf{\Xi})\,d\mathbf{\Xi}\,da_{i,j}^{\star}\,dc_{i,j}^{\star} \\
&\approx \Phi\left\{\zeta(r_{i,j}^{\star})\right\} - \Phi\left\{\zeta(r_{i,j}^{\star}-1)\right\}\,, \qquad (4)
\end{aligned}
$$

where $\zeta(r_{i,j}^{\star}) = (m_{i,r_{i,j}^{\star}}^{b} - m_{i,j}^{c,\star})(v_{i,j}^{c,\star} + v_{j,r_{i,j}^{\star}}^{b} + v_{i,j}^{\gamma})^{-0.5}$, $m_{i,j}^{c,\star} = \sum_{k=1}^{h}m_{i,k}^{u}m_{j,k}^{v}$, $v_{i,j}^{c,\star} = \sum_{k=1}^{h}[m_{i,k}^{u}]^2v_{j,k}^{v} + v_{i,k}^{u}[m_{j,k}^{v}]^2 + v_{i,k}^{u}v_{j,k}^{v}$, $v_{i,j}^{\gamma} = [b^{\gamma^{\text{row}}}b^{\gamma^{\text{col}}}][(a^{\gamma^{\text{row}}}+1)(a^{\gamma^{\text{col}}}+1)]^{-1}$ and Φ is the standard Gaussian cdf (details in the supplementary material).

Intuitively, the above predictive distribution incorporates *two sources* of uncertainty. The first comes from the unknown values of the variables in $\mathbf{\Xi}$. This uncertainty is captured by the width (variance) of the different factors that form $\mathcal{Q}$ and it is summarized in $\zeta(r_{i,j}^{\star})$ by the variance terms $v_{i,j}^{c,\star}$ and $v_{j,r_{i,j}^{\star}}^{b}$. The second comes from the heteroskedastic additive noise in $a_{i,j}^{\star}$. This uncertainty is encoded in $\zeta(r_{i,j}^{\star})$ by the variance term $v_{i,j}^{\gamma}$. Therefore, (4) allows us to take into account the uncertainty in model parameters $\mathbf{\Xi}$ and the intrinsic noisiness of the data when making predictions. Equipped with this model we can take a Bayesian approach to active learning. We first outline our active learning strategy in its general form.

3. Bayesian Active Learning

In active learning, the system *selects* which data points it wants to be labelled, rather than passively receiving labelled data. A central objective of Bayesian active learning

is to select points to minimize uncertainty over the parameters of interest, which we denote by $\mathbf{\Theta}$. Uncertainty in a random variable is most naturally captured by the Shannon entropy of its distribution. Hence, a popular utility function for Bayesian active learning is the reduction in entropy of the posterior on $\mathbf{\Theta}$ resulting from the selected point (MacKay, 1992). However, besides $\mathbf{\Theta}$, we may have a set of additional parameters $\mathbf{\Phi}$ that are of lesser interest. We want to focus on actively learning about $\mathbf{\Theta}$ and not waste data gaining information about $\mathbf{\Phi}$. Most Bayesian active learning algorithms do not make this distinction between parameters of interest and *nuisance* parameters. We make our interest on $\mathbf{\Theta}$ explicit by integrating out $\mathbf{\Phi}$ from the posterior distribution. The utility (information gain about $\mathbf{\Theta}$) of collecting an additional rating $r^\star_{i,j}$ from $\mathbf{R}$ is then

$$\mathrm{H}\left[\int p(\mathbf{\Theta},\mathbf{\Phi}|\mathbf{R}^{\mathcal{O}})\,\mathrm{d}\mathbf{\Phi}\right] - \mathbb{E}_{p(r^\star_{i,j}|\mathbf{R}^{\mathcal{O}})}\left\{\mathrm{H}\left[\int p(\mathbf{\Theta},\mathbf{\Phi}|r^\star_{i,j},\mathbf{R}^{\mathcal{O}})\,\mathrm{d}\mathbf{\Phi}\right]\right\}, \quad (5)$$

where $\mathrm{H}[\cdot]$ denotes the entropy of a distribution. The expectation with respect to $p(r^\star_{i,j}|\mathbf{R}^{\mathcal{O}})$ is taken because $r^\star_{i,j}$ is unknown prior to requesting the rating. A Bayesian approach to active learning selects the (user, item) pair (i,j) that maximizes (5). However, this can be computationally prohibitive since one must calculate the new parameter posterior $p(\mathbf{\Theta},\mathbf{\Phi}|r^\star_{i,j},\mathbf{R})$ for *every possible* entry under consideration and each possible value of that entry. Existing methods avoid this problem by using simple models whose approximate posterior can be updated quickly, e.g. aspect and flexible mixture models (Jin & Si, 2004; Harpale & Yang, 2008). However, our model is more complex and running the EP-VB routine to update the posterior approximation $\mathcal{Q}$ is relatively expensive.

To avoid having to simplify or approximate our model, we describe a more efficient approach to evaluating the utility function in (5). The previous objective can be recognized as the mutual information between $\mathbf{\Theta}$ and $r^\star_{i,j}$ given $\mathbf{R}^{\mathcal{O}}$, that is, $\mathrm{I}[\mathbf{\Theta}, r^\star_{i,j}|\mathbf{R}^{\mathcal{O}}]$. This means that we can exploit the symmetry properties of the mutual information between two random variables to re-arrange (5) into

$$\mathrm{H}[p(r^\star_{i,j}|\mathbf{R}^{\mathcal{O}})] - \mathbb{E}_{p(\mathbf{\Theta}|\mathbf{R}^{\mathcal{O}})}\left\{\mathrm{H}[\mathbb{E}_{p(\mathbf{\Phi}|\mathbf{\Theta},\mathbf{R}^{\mathcal{O}})}p(r^\star_{i,j}|\mathbf{\Theta},\mathbf{\Phi})]\right\}. \quad (6)$$

The rearrangement is highly advantageous because we no longer have to compute $p(\mathbf{\Theta},\mathbf{\Phi}|r^\star_{i,j},\mathbf{R}^{\mathcal{O}})$ multiple times (for every possible $r^\star_{i,j}$), we only require the current posterior $p(\mathbf{\Theta},\mathbf{\Phi}|\mathbf{R}^{\mathcal{O}})$. Therefore we only need to update the posterior once per sample *after* collecting the rating, as in online passive learning. Direct exploitation of this rearrangement is uncommon in the Bayesian active learning literature. Previously, it has been used for preference elicitation and is called Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2012).

BALD provides intuition about the most informative entries in $\mathbf{R}$. The first term in (6) seeks entries for which the predictive distribution has highest entropy, that is, the entries we are most uncertain about. This is maximum entropy sampling (MES) (Shewry & Wynn, 1987). However, the second term penalizes entries with high intrinsic noise. That is, if we know $\mathbf{\Theta}$ exactly, and the conditional predictive distribution for $r^\star_{i,j}$ still has high entropy, then $r^\star_{i,j}$ is not informative about $\mathbf{\Theta}$. For example, this second term will penalize selecting matrix entries corresponding to users who provide noisy, unreliable ratings. The formulation in (6) is particularly convenient as it allows the information gain to be computed accurately whilst requiring only a single update to the posterior per active sample, as in MES.

Equation (6) indicates that for effective Bayesian active learning we must capture both the uncertainty in the model parameters (to compute the first term), and the implicit noisiness in the data (to compute the second term).

3.1. Active Learning for the Cold-start Problem

Let i be the index of a new user. We want to make good predictions for this user using minimal interactions. For this, we must gain maximal information about the user's latent vector $\mathbf{u}_i$. In this active learning scenario $\mathbf{u}_i$ forms the parameters of interest $\mathbf{\Theta}$ and all the other model parameters $\mathbf{\Xi} \setminus \{\mathbf{u}_i\}$ are collected into the set of nuisance parameters $\mathbf{\Phi}$. The BALD objective (6) involves the computation of two terms: the first one can be approximated easily by the entropy of the approximate predictive distribution (4). The second term requires the computation of

$$\mathbb{E}_{Q(\mathbf{u}_i)}\mathrm{H}[\mathbb{E}_{Q(\mathbf{\Phi})}p(r^\star_{i,j}|\mathbf{u}_i,\mathbf{\Phi})] = \mathbb{E}_{Q(\mathbf{u}_i)}\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)], \quad (7)$$

where $p(r^\star_{i,j}|\mathbf{u}_i) = \Phi(\zeta(r^\star_{i,j})) - \Phi(\zeta(r^\star_{i,j}-1))$. Now $\zeta(\cdot)$ is given by $m^{a,\star}_{i,j} = \sum_{k=1}^{h} m^{v}_{j,k}u_{i,k}$ and $v^{a,\star}_{i,j} = \sum_{k=1}^{h} v^{v}_{j,k}u^2_{i,k}$ because we have conditioned on a particular $\mathbf{u}_i = (u_{i,1},\ldots,u_{i,h})$. Equation (7) includes an intractable h-dimensional Gaussian integral over $\mathbf{u}_i$. We approximate this integral by Monte Carlo sampling. In particular, we compute the expectation of $p(r^\star_{i,j}|\mathbf{u}_i)$ using a random sample from $Q(\mathbf{u}_i)$. Experimentally, this estimate converged quickly; fewer than 100 samples were required for accurate computation of (6). When computational time is critical we use the unscented approximation which uses only $2h+1$ samples placed at fixed locations (Julier & Uhlmann, 1997). This method is fast, but can generate biased estimates. Empirically, we found that the unscented approximation is sufficiently accurate to identify the most informative item in most cases, see Section 5.

We use the same method to learn about new items. In this case we draw samples from $\mathcal{Q}(\mathbf{v}_j)$ for a new item with index j, where $\mathbf{v}_j$ is the item's latent vector.

4. Related Work

Bayesian ordinal matrix factorization is addressed in Paquet et al. (2012), but their model does not include heteroskedasticity nor does it learn the boundary variables $\mathbf{B}$. Both components yield substantial improvements to predictive performance (see Section 5). Paquet's method uses Gibbs sampling for inference, whereas our EP-VB method produces accurate and compact approximations that can be easily stored and manipulated. Heteroskedasticity has been included in a MF model with a Gaussian likelihood (Lakshminarayanan et al., 2011). However, our experiments confirm that the Gaussian likelihood yields poor predictions on rating data. Another alternative for discrete ratings has been proposed by Marlin & Zemel (2009). This work assumes that each row in the rating matrix $\mathbf{R}$ is sampled i.i.d. from a Bayesian Mixture of Multinomials (BMM) model. This model is not as expressive or accurate as MF models.

Other probabilistic approaches have been proposed for cold-start active learning (Boutilier et al., 2002; Jin & Si, 2004; Harpale & Yang, 2008). These methods either maximize the expected value of information or compute posterior entropies directly, that is, they use the more expensive utility function in (5). To reduce computational cost they approximate (5) or perform approximate updates with simple models where updates are fast, such as multiple-cause vector quantizations (Ross & Zemel, 2002), naive Bayes (Boutilier et al., 2002), the aspect model (Hofmann, 2003) and flexible mixture models (Si & Jin, 2003). We perform exact computations of the utility function by using the rearrangement in (6). Furthermore, we only need to update our posterior distribution only *after* collecting the new data and not for each possible data entry that can be collected. Alternative active selection criterion are investigated in MF with a Gaussian likelihood (Sutherland et al., 2013), but learning specific parameters of interest is not addressed. Model-free strategies have been proposed for active data collection (Rashid et al., 2002; 2008), where empirical statistics of the data such as item popularity or rating entropy are used to select items. These heuristics are computationally cheap, but perform poorly relative to our model-based approach.

Outside of collaborative filtering, methods for Bayesian active learning based on posterior entropy have been widely studied (Lindley, 1956; MacKay, 1992; Settles, 2010). However, the entropy computation is often intractable or expensive and so requires approximations. Recently the BALD formulation presented in Section 3 has been used for preference learning (Houlsby et al., 2012). However, this work makes no distinction between parameters of primary interest and nuisance parameters, that is, they have $\mathbf{\Phi} = \emptyset$. This distinction is particularly important in the cold-start setting. For example, when a new user arrives, we would like to learn quickly about their corresponding latent vector but we already have ample information about the items in the system and so do not want to waste actively selected data learning about these items further.

Like BALD, methods such as maximum entropy sampling or margin sampling (Campbell et al., 2000) are cheaper to compute than Equation (5). However, unlike BALD, these methods fail to discriminate between predictive uncertainty and inherent noise in the data.

5. Experiments

We evaluate our new model and active learning strategy on a diverse collection of rating datasets: i) *MovieLens100K* and *MovieLens1M*: two collections of ratings of movies; ii) *MovieTweets*: movie ratings obtained from Twitter; iii) *Webscope*: ratings of songs; iv) *Jester*: ratings of jokes; v) *Book*: ratings of books; vi) *Dating*: ratings from an online dating website and vii) *IPIP*: ordinal responses to a psychometrics questionnaire. All the matrix entries in IPIP are observed, the other datasets have many missing entries. Descriptions, links to the data, and our pre-processing steps are in the supplementary material. We first evaluate the predictive accuracy of our model against a number of state-of-the-art alternatives. We then investigate the performance of our method for cold-start active learning.

5.1. Comparison to Other Models for Rating Data

We compare our model for heteroskedastic ordinal matrix factorization (HOMF) against the following methods: i) the homoskedastic model with an ordinal likelihood in Paquet et al. (2012) (Paquet); ii) a method for robust Bayesian matrix factorization (RBMF) based on a Gaussian likelihood which includes heteroskedastic noise (Lakshminarayanan et al., 2011); iii) the Bayesian mixture of multinomials model (BMM) (Marlin & Zemel, 2009); and iv) a matrix factorization model like RBMF but with homoskedastic noise (BMF). We directly evaluate the improvements in predictive performance produced in HOMF from both considering heteroskedasticity and learning the boundary variables $\mathbf{B}$. For these evaluations we first compare to OMF, a homoskedastic version of HOMF, where the variance parameters γ_i^{row} are all constrained to be equal to each other (similarly for the γ_j^{col}). Secondly, HOMF-NoB uses fixed boundary parameters $\mathbf{b}_j$ rather than learning them for each item. Finally, OMF-NoB is a homoskedastic version of HOMF that does not learn $\mathbf{B}$. For all models we fix the latent dimension to $h = 10$.

We split the available ratings for each dataset randomly into a training and a test set with 80% and 20% of the ratings respectively. Each method was adjusted using the entries in the training set and then evaluated using predictive log-likelihood on the corresponding test set. The entire

Table 1. Average test log likelihood. Bold typeface denotes the best method and those statistically indistinguishable.

Method	HOMF	OMF	HOMF -NoB	OMF -NoB	Paquet	RBMF	BMF	BMM
Books	**-1.415**	-1.436	-1.507	-1.439	-1.427	-1.545	-1.544	-1.622
Dating	**-0.867**	-0.906	-0.890	-1.028	-1.009	-1.045	-1.140	-0.948
IPIP	**-1.096**	-1.140	-1.131	-1.189	-1.188	-1.194	-1.225	-1.270
Jest	**-1.238**	-1.306	**-1.240**	-1.320	-1.320	-1.312	-1.368	-1.290
ML1M	**-1.136**	-1.165	-1.141	-1.177	-1.170	-1.173	-1.210	-1.324
ML100K	**-1.203**	-1.234	-1.208	-1.243	-1.232	-1.238	-1.277	-1.493
MTweet	**-0.956**	-0.991	-0.984	-1.025	-1.012	-1.014	-1.077	-1.115
WebSc.	**-1.207**	-1.253	-1.209	-1.257	-1.236	-1.529	-1.532	-1.298

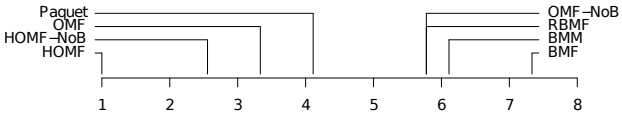

Figure 2. Average rank of each method across all the datasets.

procedure, including dataset partitioning, was repeated 20 times.

Table 1 contains the test log-likelihoods and Figure 2 summarizes the performance of each algorithm. The proposed model, HOMF, outperforms all the other methods in all datasets. Significance is assessed with a paired t-test. The likelihood function for ordinal data is more appropriate for ratings than the Gaussian likelihood; HOMF and Paquet outperform RBMF and BMF. Furthermore, predictive performance is improved by modeling heteroskedasticity across rows and across columns since HOMF outperforms OMF and Paquet, and RBMF outperforms BMF. Learning the biases also results in substantial improvements to the performance of our model. Finally, the matrix factorization models (HOMF, Paquet, RBMF and BMF) usually outperform the mixture model BMM.

5.2. Cold-start Active Learning

We selected 2000 users and 1000 items (up to the maximum available) with the most ratings from each dataset. This provided the active sampling strategies with the largest possible pool of data to select from. We partitioned the data randomly into three sets: training, test and pool. For this, we sampled 75% of the users and added all of their ratings to the training set. These represented the ratings for the users that were *already in the system*. Each of the remaining 25% *test users* were initialized with a single item, adding that rating to the training set. For each test user we sampled three ratings and added these to the test set. The remaining ratings were added to the pool set. Figure 4 illustrates this set-up. We also simulated new items arriving to the system. In this case the setup was identical except that the roles of the users and items were interchanged. We denote the new-users and new-items experiments by appending -U and -I to the dataset names respectively.

HOMF was adjusted using the available ratings in the training set. Then, during each iteration of active learning, a single rating was selected from the pool set for each test user using active learning. The selected ratings were then added to the training set and HOMF was incrementally readjusted using the new training set. We evaluated the second term in (6) using Monte Carlo sampling from Q with 100 samples. As alternatives to BALD we considered random sampling (*Rand*), maximum entropy sampling (*Entropy*) and a model-free version of Entropy that selects the item whose empirical distribution of ratings in the training data has the greatest entropy (*Emp-Ent*). After each active sample was selected, we computed the predictive log-likelihood on the test set. The entire procedure was repeated 25 times.

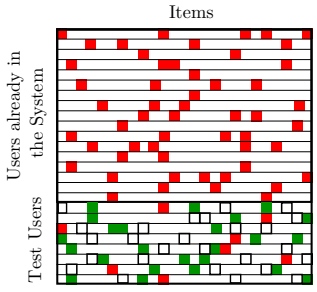

Figure 4. Experimental setup for cold-start active learning. The squares depict available ratings. Red squares form the training set. These are all the ratings for those users already in the system and one rating per test user. Green squares form the test set. The remaining hollow squares form the pool set for the test users. Note that most ratings are missing.

Active Learning Strategies Figure 3 shows the learning curves with each strategy for each new-user experiment. The curves for the new-item experiments are in the supplementary material. Table 2, left hand columns, summarizes the results with the test log-likelihood after drawing 10 samples from the pool for each test user or item. With HOMF, BALD yields the best (or joint best) predictions in all but one cases. Both the model based and empirical entropy sampling algorithms often perform poorly because they ignore the inherent noisiness of the users or items. Note that in most datasets, there are only a few ratings available for most users. This means that BALD is restricted to sampling from a limited pool set. In particular, Book, MovieTweets and Webscope are the most sparse, with only 2, 3 and 5% of ratings available, respectively. Unsurprisingly, BALD exhibits smaller performance gains on these datasets. In practice, in these domains most users would be able to provide ratings for a larger number of items; they may watch a new movie, listen to a song, read a book, etc. Consequently, in practice we would expect to see larger performance gains as in the denser datasets, IPIP and Jester. This assumption may not always hold, for example, in dating recommendation a user may not follow

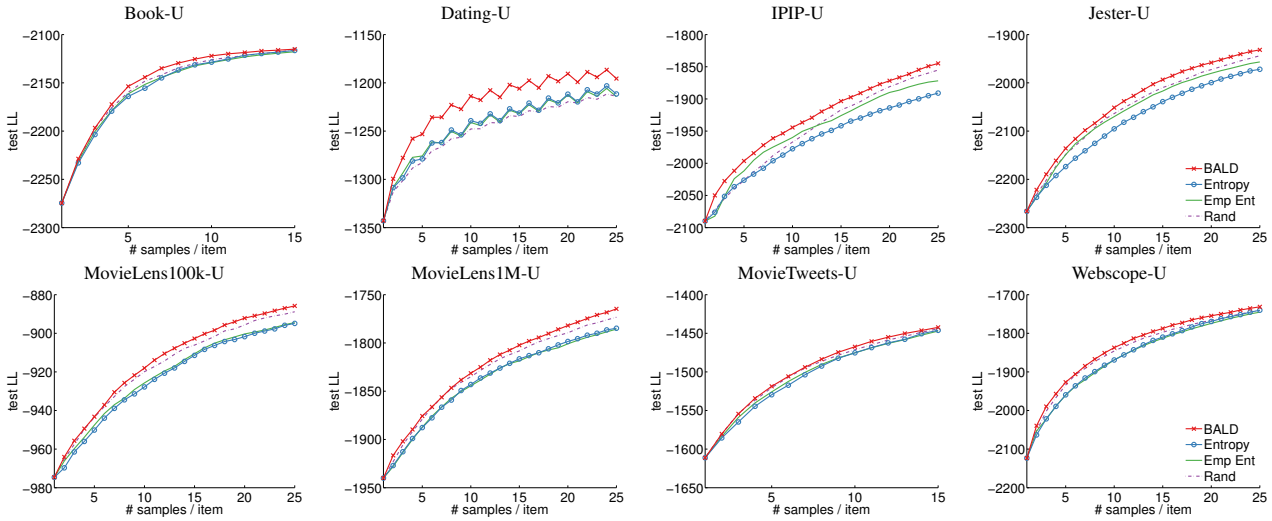

Figure 3. Predictive log likelihood after each round of active sampling for each algorithm.

Table 2. Log-likelihood after receiving 10 samples. Underlining indicates the top performing active sampling algorithms for each model and bold denotes the best overall method. The bottom row gives the number of datasets that each strategy yields the best (or joint best) performance with each model.

	Heteroscedastic (HOMF)				Homoscedastic (OMF)				BMM			
Dataset	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**
Book-U	**-2122**	-2129	-2129	-2126	-2146	-2149	-2150	-2147	-2405	-2418	-2413	-2411
Dating-U	**-1214**	-1239	-1241	-1248	-1217	-1230	-1235	-1244	-1234	-1309	-1305	-1255
IPIP-U	**-1944**	-1977	-1960	-1967	**-1945**	-1978	-1964	-1973	-1964	-1988	-1983	-1987
Jester-U	-2051	-2095	-2070	-2064	-2080	-2119	-2100	-2099	**-2041**	-2075	-2054	**-2045**
MLens100k-U	**-918**	-928	-926	**-920**	-926	-927	-929	-926	-989	-1001	-997	-988
MLens1M-U	**-1831**	-1843	-1844	**-1835**	-1840	-1850	-1854	-1846	-1877	-1899	-1898	-1879
MTweets-U	**-1467**	-1475	-1475	**-1471**	-1503	-1508	-1508	-1503	-1608	-1624	-1622	-1613
Webscope-U	**-1837**	-1869	-1869	-1846	-1882	-1898	-1903	-1880	-1951	-1984	-1970	-1958
Book-I	**-2038**	-2039	**-2037**	**-2038**	-2095	-2094	-2094	-2095	-2186	-2198	-2202	-2195
Dating-I	-1630	-1720	-1655	**-1612**	-1672	-1722	-1684	-1643	**-1603**	-1691	-1631	**-1602**
IPIP-I	**-319**	-325	-339	-329	-325	-325	-339	-330	-335	-347	-346	-339
Jester-I	**-99**	**-99**	**-99**	**-100**	-102	-102	-101	-102	-104	-107	-106	-104
Mlens100k-I	**-1085**	-1103	-1095	-1099	-1110	-1112	-1111	-1113	-1160	-1186	-1171	-1170
Mlens1M-I	**-1831**	-1843	-1844	**-1835**	-1840	-1850	-1854	-1846	-1877	-1899	-1898	-1879
MTweets-I	**-1470**	-1479	-1475	-1476	-1519	-1520	-1520	-1520	-1605	-1617	-1613	-1608
Webscope-I	**-1837**	-1869	-1869	-1846	-1882	-1898	-1903	-1880	-1951	-1984	-1970	-1958
Wins / 16	15	1	2	7	15	7	5	12	16	1	2	12

any recommendation and provide a rating. In this case, the probability of receiving a rating from the user should be accounted for.

In cold start learning, it is crucial to elicit useful information from the very first sample. The average number of queries required to achieve the same predictive performance as from a single active sample chosen by BALD is 1.85 with Ent, 1.83 with Emp, and 1.59 with random sampling. This means that on average around 60% more random samples are required to gain the same performance as the first sample selected by BALD.

Heteroskedasticity vs. Homoskedasticity We ran each active learning strategy with our homoskedastic model OMF and with the homoskedastic method BMM, see Table 2. With these homoskedastic models active learning significantly outperforms random sampling on fewer datasets than with HOMF. This demonstrates that accurate estimates of the intrinsic noisiness of the data are required to unlock the full potential of Bayesian active learning. Figure 5 presents example learning curves on MovieTweets-I, where the difference in relative performance of BALD versus Rand when using HOMF and OMF is large. One can see that BALD provides much faster learning than the other strategies with HOMF, but all strategies are indistinguishable with OMF. This indicates that some users in this dataset provide highly noisy and unpredictable ratings. HOMF is able to model this and elicit ratings only from the useful, low noise users.

We also evaluated each model in terms of RMSE, see Table 2 in the supplementary material. In this case, the best model is OMF, very closely followed by HOMF. Note

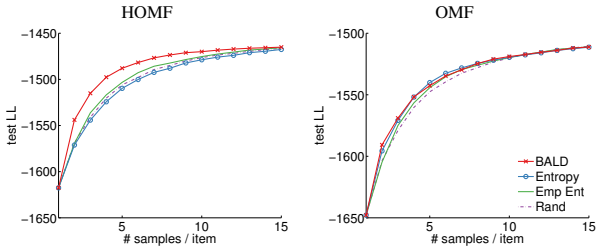

Figure 5. Comparison of active learning strategies when using HOMF and OMF with the MovieTweets-I dataset.

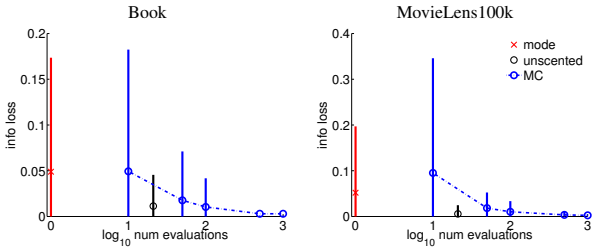

Figure 6. Information loss (8) from approximations to $\mathbb{E}_{Q(\mathbf{u}_i)}\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)]$ versus the number of samples drawn from $\mathcal{Q}$. Vertical bars indicate the 10th and 90th percentiles.

that the ranking of these two methods was the opposite in Table 1. The reason for this is that RMSE focuses only on the predictive mean and ignores the predictive variance. Modeling heteroskedasticity largely affects the predictive variance, but barely changes the predictive mean. Therefore, OMF and HOMF are expected to perform similarly under the RMSE metric. Nevertheless, in cold-start active learning HOMF with BALD performs best overall, see Table 4 in the supplementary material. This is because, although heteroskedasticity does not assist in the final evaluation under the RMSE metric, it is still important to enable BALD to find the informative ratings.

5.2.1. Speeding Up the Computation of BALD

In online settings, the time available for selecting the most informative matrix entries may be limited. We can reduce the cost of BALD by making approximations when computing the second term in the utility function in Equation (6), $\mathbb{E}_{\mathcal{Q}(\mathbf{u}_i)}\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)]$, as described in Section 3.1. We evaluate the accuracy of three approximations: Monte Carlo (MC) sampling, the unscented approximation, and evaluating the integral with a delta function located at the mode of $\mathcal{Q}$. We are interested in finding the most informative item, so we evaluate the estimation error using fraction information loss, measured as

$$\frac{\max_j \hat{I}(j) - \hat{I}(\arg\max_j I(j))}{\max_j \hat{I}(j)}, \tag{8}$$

where $I(j)$ is given by (6) evaluated on item j using the approximation and $\hat{I}(j)$ is a gold standard obtained using MC with a separate set of 1000 samples. The results are averaged over all test users. The loss across all datasets (± 1 s.d.) from MC with 100 samples, the unscented approximation and the posterior mode approximation were 0.017 ± 0.007, 0.035 ± 0.031 and 0.136 ± 0.073, respectively. Figure 6 depicts the loss as a function of the number of evaluations of $\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)]$ on Movielens100k and Webscope. Results on the other datasets are similar and are in the supplementary material. With MC the integral converges rapidly, the loss falls below 5% with fewer than 50 samples on all datasets. The unscented approximation requires only $2h + 1$ evaluations, and in most cases yields a better estimate than MC with this number of samples.

In practice, we found no statistical difference in predictive performance when running the experiments using the unscented approximation or MC with 100 samples. We therefore recommend the unscented approximation as an efficient solution for systems with computational constraints.

6. Conclusions

We have proposed new a framework for cold-start learning based on a Bayesian active learning strategy that learns as much as possible about new entities (users or items) from minimal user interactions. To achieve strong performance we have proposed a matrix factorization model that takes into account the ordinal nature of rating data and incorporates different levels of noise across the rows and columns of a rating matrix. This model uses hierarchical priors to provide additional robustness to fixing hyper-parameter values. With this model we perform efficient Bayesian active learning by extending a new framework for computing information gain (BALD) to collaborative filtering, where we only want to learn optimally about user (or item) specific model parameters. This approach removes the requirement to re-compute the parameter posterior many times per active sample, and hence permits us to use our relatively complex matrix factorization model. Our model generates state-of-the-art predictions on rating data, and when combined with BALD yields strong performance in cold-start active learning from the very first sample.

This work addresses learning about a new entity as quickly as possible. An important extension for real-world systems is to trade-off exploration and exploitation, balancing information gain with recommending a user their most favored items. Possible extensions to this setting include incorporating active search (Garnett et al., 2011) or strategies from Bandit Theory into our framework.

Acknowledgements

NMTH and JMH are grateful for support from the Google European Doctoral Fellowship scheme and Infosys Labs, Infosys Limited, respectively.

References

Ahn, Hyung Jun. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, 178(1):37–51, 2008.

Boutilier, Craig, Zemel, Richard S, and Marlin, Benjamin. Active collaborative filtering. In *UAI*, pp. 98–106, 2002.

Campbell, Colin, Cristianini, Nello, and Smola, Alex. Query learning with large margin classifiers. In *ICML*, pp. 111–118, 2000.

Chu, Wei and Ghahramani, Zoubin. Gaussian processes for ordinal regression. In *JMLR*, pp. 1019–1041, 2005.

Claypool, Mark, Gokhale, Anuja, Miranda, Tim, Murnikov, Pavel, Netes, Dmitry, and Sartin, Matthew. Combining content-based and collaborative filters in an online newspaper. In *SIGIR workshop on recommender systems*, volume 60, 1999.

Garnett, Roman, Krishnamurthy, Yamuna, Wang, Donghan, Schneider, Jeff, and Mann, Richard. Bayesian optimal active search on graphs. In *MLG*, 2011.

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

Harpale, Abhay S and Yang, Yiming. Personalized active learning for collaborative filtering. In *SIGIR*, pp. 91–98, 2008.

Hofmann, Thomas. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *SIGIR*, pp. 259–266, New York, NY, USA, 2003. ACM.

Houlsby, Neil, Hernandez-Lobato, Jose Miguel, Huszar, Ferenc, and Ghahramani, Zoubin. Collaborative Gaussian processes for preference learning. In *NIPS*, pp. 2105–2113, 2012.

Jin, Rong and Si, Luo. A Bayesian approach toward active learning for collaborative filtering. In *UAI*, pp. 278–285, 2004.

Julier, Simon J and Uhlmann, Jeffrey K. New extension of the Kalman filter to nonlinear systems. In *AeroSense*, pp. 182–193, 1997.

Lakshminarayanan, Balaji, Bouchard, Guillaume, and Archambeau, Cedric. Robust Bayesian matrix factorisation. In *AISTATS*, pp. 425–433, 2011.

Lewi, Jeremy, Butera, Robert J, and Paninski, Liam. Efficient active learning with generalized linear models. In *AISTATS*, pp. 267–274, 2007.

Lindley, Dennis V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pp. 986–1005, 1956.

MacKay, David JC. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Maltz, David and Ehrlich, Kate. Pointing the way: active collaborative filtering. In *SIGCHI*, pp. 202–209, 1995.

Marlin, Benjamin M. and Zemel, Richard S. Collaborative prediction and ranking with non-random missing data. In *RecSys*, pp. 5–12, New York, NY, USA, 2009. ACM.

Minka, Thomas P. Expectation propagation for approximate Bayesian inference. In *UAI*, pp. 362–369, 2001.

Nickisch, Hannes and Rasmussen, Carl Edward. Approximations for binary Gaussian process classification. *JMLR*, 9:2035–2078, 2008.

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

Park, Seung-Taek and Chu, Wei. Pairwise preference regression for cold-start recommendation. In *RecSys*, pp. 21–28, 2009.

Park, Seung-Taek, Pennock, David, Madani, Omid, Good, Nathan, and DeCoste, Dennis. Naïve filterbots for robust cold-start recommendations. In *SIGKDD*, pp. 699–705, 2006.

Rashid, Al Mamunur, Albert, Istvan, Cosley, Dan, Lam, Shyong K, McNee, Sean M, Konstan, Joseph A, and Riedl, John. Getting to know you: learning new user preferences in recommender systems. In *IUI*, pp. 127–134, 2002.

Rashid, Al Mamunur, Karypis, George, and Riedl, John. Learning preferences of new users in recommender systems: an information theoretic approach. *ACM SIGKDD Explorations Newsletter*, 10(2):90–100, 2008.

Ross, David A and Zemel, Richard S. Multiple cause vector quantization. In *NIPS*, pp. 1017–1024, 2002.

Salakhutdinov, Ruslan and Mnih, Andriy. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *ICML*, pp. 880–887, 2008.

Schein, Andrew I, Popescul, Alexandrin, Ungar, Lyle H, and Pennock, David M. Methods and metrics for cold-start recommendations. In *SIGIR*, pp. 253–260, 2002.

Settles, Burr. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

Shewry, Michael C and Wynn, Henry P. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.

Si, Luo and Jin, Rong. Flexible mixture model for collaborative filtering. In *ICML*, volume 3, pp. 704–711, 2003.

Stern, David H, Herbrich, Ralf, and Graepel, Thore. Matchbox: large scale online bayesian recommendations. In *WWW*, pp. 111–120, 2009.

Sutherland, Dougal J, Póczos, Barnabás, and Schneider, Jeff. Active learning and search on low-rank matrices. In *SIGKDD*, pp. 212–220, 2013.