Mixed Membership Matrix Factorization

Lester Mackey LMACKEY@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720, USA

David Weiss DJWEISS@CIS.UPENN.EDU
Computer and Information Science, University of Pennsylvania, Philadephia, PA 19104, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA

Abstract

Discrete mixed membership modeling and continuous latent factor modeling (also known as matrix factorization) are two popular, complementary approaches to dyadic data analysis. In this work, we develop a fully Bayesian framework for integrating the two approaches into unified Mixed Membership Matrix Factorization (M^{3}F) models. We introduce two M^{3}F models, derive Gibbs sampling inference procedures, and validate our methods on the EachMovie, MovieLens, and Netflix Prize collaborative filtering datasets. We find that, even when fitting fewer parameters, the M^{3}F models outperform state-of-the-art latent factor approaches on all benchmarks, yielding the greatest gains in accuracy on sparsely-rated, high-variance items.

1. Introduction

This work is concerned with unifying discrete mixed membership modeling and continuous latent factor modeling for probabilistic dyadic data prediction. In the dyadic data prediction (DDP) problem (Hofmann et al., 1999), we observe labeled *dyads*, i.e., ordered pairs of objects, and form predictions for the labels of unseen dyads. For example, in the collaborative filtering setting, we observe U users, M items, and a training set $\mathcal{T} = \{(u_n, j_n, r_n)\}_{n=1}^N$ with real-valued ratings r_n representing the preferences of certain users u_n for certain items j_n. The goal is then to predict unobserved ratings based on users' past preferences. Other concrete examples of DDP include link prediction in social network analysis, binding affinity prediction in bioinformatics, and click prediction in web search.

Matrix factorization methods (Rennie & Srebro, 2005; DeCoste, 2006; Salakhutdinov & Mnih, 2007; 2008; Takács et al., 2009; Lawrence & Urtasun, 2009) represent the state of the art for dyadic data prediction tasks. These methods view a dyadic dataset as a sparsely observed ratings matrix, $R \in \mathbb{R}^{U \times M}$, and learn a constrained decomposition of that matrix as a product of two latent factor matrices: $R \approx A^t B$ for $A \in \mathbb{R}^{D \times U}$, $B \in \mathbb{R}^{D \times M}$, and D small. While latent factor methods perform remarkably well on the DDP task, they fail to capture the heterogeneous nature of objects and their interactions. Such models, for instance, do not account for the fact that a user's ratings are influenced by instantaneous mood, that protein interactions are affected by transient functional contexts, or even that users with distinct behaviors may be sharing a single account or web browser.

The fundamental limitation of continuous latent factor methods is a result of the static way in which ratings are assumed to be produced: a user generates all of his item ratings using the same factor vector, without regard for context. Discrete mixed membership models, like Latent Dirichlet Allocation (Blei et al., 2003), were developed to address a similar limitation of mixture models. Whereas mixture models assume that each generated object is underlyingly a member of a single latent topic, mixed membership models represent objects as distributions over topics. Mixed membership dyadic data models such as the Mixed Membership Stochastic Blockmodel (Airoldi et al., 2008) for relational prediction and Bi-LDA (Porteous et al., 2008) for rating prediction introduce context dependence by allowing each object to select a new topic for each new interaction. However, the relatively poor

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

predictive performance of Bi-LDA suggests that the blockmodel assumption—that objects only interact via their topics—is too restrictive.

In this paper we develop a fully Bayesian framework for wedding the strong performance and expressiveness of continuous latent factor models with the context dependence and topic clustering of discrete mixed membership models. In Section 2, we provide additional background on matrix factorization and mixed membership modeling. We introduce our Mixed Membership Matrix Factorization (M^3F) framework in Section 3, and discuss inference and prediction under two M^3F models in Section 4. Section 5 describes experimental evaluation and analysis of our models on a variety of real-world collaborative filtering datasets. The results demonstrate that Mixed-Membership Matrix Factorization methods outperform their context-blind counterparts and simultaneously reveal interesting clustering structure in the data. Finally, we conclude in Section 6.

2. Background

2.1. Latent Factor Models

We begin by considering a prototypical latent factor model, Bayesian Probabilistic Matrix Factorization of Salakhutdinov & Mnih (2008) (see Figure 1). Like most factor models, BPMF associates with each user u an unknown factor vector $\mathbf{a}_u \in \mathbb{R}^D$ and with each item j an unknown factor vector $\mathbf{b}_j \in \mathbb{R}^D$. A user generates a rating for an item by adding Gaussian noise to the inner product, $r_{uj} = \mathbf{a}_u \cdot \mathbf{b}_j$. We refer to this inner product as the *static rating* for a user-item pair, for, as discussed in the introduction, the latent factor rating mechanism does not model the context in which a rating is given and does not allow a user to don different moods or "hats" in different dyadic interactions. Such contextual flexibility is desirable for capturing the context-sensitive nature of dyadic interactions, and, as such, we turn our attention to mixed membership models.

2.2. Mixed Membership Models

Two recent examples of dyadic mixed membership (DMM) models are the Mixed Membership Stochastic Blockmodel (MMSB) (Airoldi et al., 2008) and Bi-LDA (Porteous et al., 2008) (see Figure 1). In DMM models, each user u and item j has its own discrete distribution over topics, represented by topic parameters θ_u^U and θ_j^M. When a user desires to rate an item, both the user and the item select interaction-specific topics according to their distributions; the selected topics then determine the distribution over ratings.

One drawback of DMM models is the reliance on purely groupwise interactions: one learns how a user group interacts with an item group but not how a user group interacts directly with a particular item. M^3F models address this limitation in two ways—first, by modeling interactions between groups and specific users or items and second, by incorporating the user-item specific static rating of latent factor models.

3. Mixed Membership Matrix Factorization

In this section, we present a general Mixed Membership Matrix Factorization framework and two specific models that leverage the predictive power and static specificity of continuous latent factor models while allowing for the clustered context-sensitivity of mixed membership models. In each M^3F model, users and items are endowed both with latent factor vectors ($\mathbf{a}_u$ and $\mathbf{b}_j$) and with topic distribution parameters (θ_u^U and θ_j^M). To rate an item, a user first draws a topic z_{uj}^U from his distribution, representing, for example, his mood at the time of rating (in the mood for romance vs. comedy), and the item draws a topic z_{uj}^M from its distribution, representing, for example, the context under which it is being rated (in a theater on opening night vs. in a high-school classroom). The user and item topics, i and k, together with the identity of the user and item, u and j, jointly specify a rating bias, β_{uj}^{ik}, tailored to the user-item pair. Different M^3F models will differ principally in the precise form of this *contextual bias*. To generate a complete rating, the user-item-specific static rating $\mathbf{a}_u \cdot \mathbf{b}_j$ is added to the contextual bias β_{uj}^{ik}, along with some noise.

Rather than learn point estimates under our M^3F models, we adopt a fully Bayesian methodology and place priors on all parameters of interest. Topic distribution parameters θ_u^U and θ_j^M are given independent exchangeable Dirichlet priors, and the latent factor vectors $\mathbf{a}_u$ and $\mathbf{b}_j$ are drawn independently from $\mathcal{N}\left(\mu^U, (\Lambda^U)^{-1}\right)$ and $\mathcal{N}\left(\mu^M, (\Lambda^M)^{-1}\right)$, respectively. As in Salakhutdinov & Mnih (2008), we place normal-Wishart priors on the hyper-parameters (μ^U, Λ^U) and (μ^M, Λ^M). Suppose K^U is the number of user topics and K^M is the number of item topics. Then, given the contextual biases β_{uj}^{ik}, ratings are generated according to the following M^3F generative process:

$$\Lambda^U \sim \text{Wishart}(\mathbf{W}_0, \nu_0),\ \Lambda^M \sim \text{Wishart}(\mathbf{W}_0, \nu_0)$$

$$\mu^U \sim \mathcal{N}\left(\mu_0, (\lambda_0 \Lambda^U)^{-1}\right),\ \mu^M \sim \mathcal{N}\left(\mu_0, (\lambda_0 \Lambda^M)^{-1}\right)$$

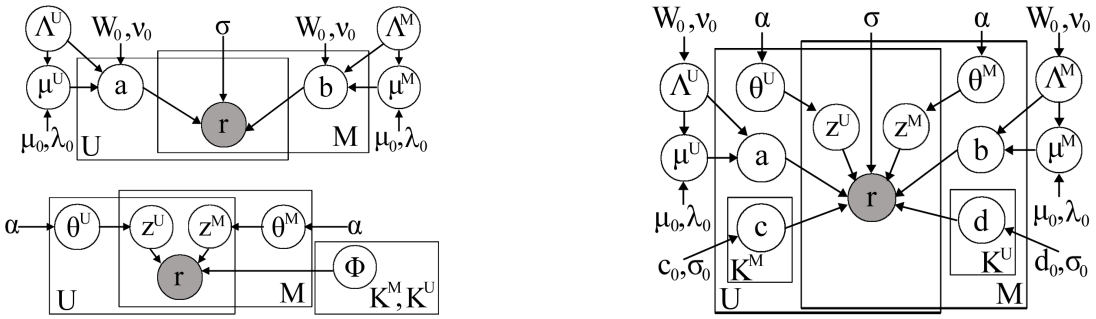

Figure 1. Graphical model representations of BPMF (top left), Bi-LDA (bottom left), and M^{3}F-TIB (right).

For each $u \in \{1, \ldots, U\}$:

$$\mathbf{a}_u \sim \mathcal{N}\left(\mu^U, (\Lambda^U)^{-1}\right)$$
$$\theta_u^U \sim \text{Dir}(\alpha / K^U)$$

For each $j \in \{1, \ldots, M\}$:

$$\mathbf{b}_j \sim \mathcal{N}\left(\mu^M, (\Lambda^M)^{-1}\right)$$
$$\theta_j^M \sim \text{Dir}(\alpha / K^M)$$

For each rating r_{uj}:

$$z_{uj}^U \sim \text{Multi}(1, \theta_u^U),\ z_{uj}^M \sim \text{Multi}(1, \theta_j^M)$$
$$r_{uj} \sim \mathcal{N}\left(\beta_{uj}^{ik} + \mathbf{a}_u \cdot \mathbf{b}_j, \sigma^2\right).$$

For each model discussed below, we let Θ^U denote the collection of all user parameters (e.g., $\mathbf{a}, \theta^U, \Lambda^U, \mu^U$), Θ^M denote all item parameters, and Θ_0 denote all global parameters (e.g., $\mathbf{W}_0, \nu_0, \mu_0, \lambda_0, \alpha, \sigma_0^2, \sigma^2$). We now describe in more detail the specific forms of two M^{3}F models and their contextual biases.

3.1. The M^{3}F Topic-Indexed Bias Model

The M^{3}F Topic-Indexed Bias (TIB) model assumes that the contextual bias decomposes into a latent user bias and a latent item bias. The user bias is influenced by the interaction-specific topic selected by the item. Similarly, the item bias is influenced by the user's selected topic. We denote the latent rating bias of user u under item topic k as c_u^k and denote the bias for item j under user topic i as d_j^i. The contextual bias for a given user-item interaction is then found by summing the two latent biases and a fixed global bias, χ_0[1]:

$$\beta_{uj}^{ik} = \chi_0 + c_u^k + d_j^i.$$

Topic-indexed biases c_u^k and d_j^i are drawn independently from Gaussian priors with variance σ_0^2 and means c_0 and d_0 respectively. Figure 1 compares the graphical model representations of M^{3}F-TIB, BPMF, and Bi-LDA. Note that M^{3}F-TIB reduces to BPMF when K^U and K^M are both zero.

Intuitively, the topic-indexed bias model captures the "*Napoleon Dynamite* effect," (Thompson, 2008) whereby certain movies provoke strongly differing reactions from otherwise similar users. Each user-topic-indexed bias d_j^i represents one of K^U possible predispositions towards liking or disliking each item in the database, irrespective of the static latent factor parameterization. Thus, in the movie-recommendation problem, we expect the variance in user reactions to movies such as *Napoleon Dynamite* to be captured in part by a corresponding variance in the bias parameters d_j^i (see Section 5). Moreover, because the model is symmetric, each rating is also influenced by the item-topic-indexed bias c_u^k. This can be interpreted as the predisposition of each perceived item class towards being liked or disliked by each user in the database. Finally, because M^{3}F-TIB is a mixed-membership model, each user and item can choose a different topic and hence a different bias for each rating (e.g., when multiple users share a single account).

3.2. The M^{3}F Topic-Indexed Factor Model

The M^{3}F Topic-Indexed Factor (TIF) model assumes that the joint contextual bias is an inner product of topic-indexed factor vectors, rather than the sum of topic-indexed biases as in the TIB model. Each item topic k maintains a latent factor vector $\mathbf{c}_u^k \in \mathbb{R}^{\tilde{D}}$ for each user, and each user topic i maintains a latent factor vector $\mathbf{d}_j^i \in \mathbb{R}^{\tilde{D}}$ for each item. Each user and each item additionally maintains a single static rating bias, ξ_u and χ_j respectively. The joint contextual bias is formed by summing the user bias, the item bias, and the inner product between the topic-indexed factor vectors:

$$\beta_{uj}^{ik} = \xi_u + \chi_j + \mathbf{c}_u^k \cdot \mathbf{d}_j^i.$$

[1]The global bias, χ_0, is suppressed in the remainder of the paper for clarity.

Algorithm 1 Gibbs Sampling for M^{3}F-TIB.

Input: $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}, \mathbf{c}^{(0)}, \mathbf{d}^{(0)}, \theta^{U(0)}, \theta^{M(0)}, \mathbf{z}^{M(0)})$
for $t = 1$ **to** T **do**
  // Sample Hyperparameters
  for $(u, j) \in \mathcal{T}$ **do**
    $(\mu^U, \Lambda^U)^t \sim \mu^U, \Lambda^U \mid \mathbf{a}^{t-1}, \Theta_0$
    $(\mu^M, \Lambda^M)^t \sim \mu^M, \Lambda^M \mid \mathbf{b}^{t-1}, \Theta_0$
  end for
  // Sample Topics
  for $(u, j) \in \mathcal{T}$ **do**
    $z_{uj}^{U(t)} \sim z_{uj}^U | (z_{uj}^M, \theta_u^U, \mathbf{a}_u, \mathbf{b}_j, \mathbf{c}_u, \mathbf{d}_j)^{t-1}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
    $z_{uj}^{M(t)} \sim z_{uj}^M | (\theta_j^M, \mathbf{a}_u, \mathbf{b}_j, \mathbf{c}_u, \mathbf{d}_j)^{t-1}, z_{uj}^{U(t)}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
  end for
  // Sample User Parameters
  for $u = 1$ **to** U **do**
    $\theta_u^{U(t)} \sim \theta_u^U \mid \mathbf{z}^{U(t)}, \Theta_0$
    $\mathbf{a}_u^t \sim \mathbf{a}_u \mid (\Lambda^U, \mu^U, \mathbf{z}_u^U, \mathbf{z}^M)^t, (\mathbf{b}, \mathbf{c}_u, \mathbf{d})^{t-1}, \Theta_0$
    for $i = 1$ **to** K^M **do**
      $c_u^{i(t)} \sim c_u^i \mid (\mathbf{z}^U, \mathbf{z}^M, \mathbf{a}_u)^t, (\mathbf{b}, \mathbf{d})^{t-1}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
    end for
  end for
  // Sample Item Parameters
  for $j = 1$ **to** M **do**
    $\theta_j^{M(t)} \sim \theta_j^M \mid \mathbf{z}^{M(t)}, \Theta_0$
    $\mathbf{b}_j^t \sim \mathbf{b}_j \mid (\Lambda^U, \mu^U, \mathbf{z}_u^U, \mathbf{z}^M, \mathbf{a}, \mathbf{c}_u)^t, \mathbf{d}^{t-1}, \Theta_0$
    for $k = 1$ **to** K^U **do**
      $d_j^{k(t)} \sim d_j^k \mid (\mathbf{z}^U, \mathbf{z}^M, \mathbf{a}, \mathbf{b}_j, \mathbf{c})^t, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
    end for
  end for
end for

The topic-indexed factors $\mathbf{c}_u^k$ and $\mathbf{d}_j^i$ are drawn independently from $\mathcal{N}\left(\tilde{\mu}^U, (\tilde{\Lambda}^U)^{-1}\right)$ and $\mathcal{N}\left(\tilde{\mu}^M, (\tilde{\Lambda}^M)^{-1}\right)$ priors, and conjugate normal-Wishart priors are placed on the hyper-parameters $(\tilde{\mu}^U, \tilde{\Lambda}^U)$ and $(\tilde{\mu}^M, \tilde{\Lambda}^M)$. The static user and item biases, ξ_u and χ_j, are drawn independently from Gaussian priors with variance σ_0^2 and means ξ_0 and χ_0 respectively.[2]

Intuitively, the topic-indexed factor model can be interpreted as an extended matrix factorization with both *global* and *local* low-dimensional representations. Each user u has a single global factor $\mathbf{a}_u$ but K^U local factors $\mathbf{c}_u^k$; similarly, each item j has both a global factor $\mathbf{b}_j$ and multiple local factors $\mathbf{d}_j^i$. A strength of latent factor methods is their ability to discover globally predictive intrinsic properties of users and items. The topic-indexed factor model extends this representation to allow for intrinsic properties that are predictive in some but perhaps not all contexts. For example, in the movie-recommendation setting, is *Lost In Translation* a dark comedy or a romance film? The answer may vary from user to user and thus may be captured by different vectors $\mathbf{d}_j^i$ for each user-indexed topic.

[2]Static biases ξ and χ are suppressed in the remainder of the paper for clarity.

4. Inference and Prediction

The goal in dyadic data prediction is to predict unobserved ratings $\mathbf{r}^{(\mathrm{h})}$ given observed ratings $\mathbf{r}^{(\mathrm{v})}$. As in Salakhutdinov & Mnih (2007; 2008) and Takács et al. (2009), we adopt root mean squared error (RMSE)[3] as our primary error metric and note that the Bayes optimal prediction under RMSE loss is the posterior mean of the predictive distribution $p(\mathbf{r}^{(\mathrm{h})} | \mathbf{r}^{(\mathrm{v})}, \Theta_0)$.

In our M^{3}F models, the predictive distribution over unobserved ratings is found by integrating out all topics and parameters. The posterior distribution $p(\mathbf{z}^U, \mathbf{z}^M, \Theta^U, \Theta^M | \mathbf{r}^{(\mathrm{v})}, \Theta_0)$ is thus our main inferential quantity of interest. Unfortunately, as in both LDA and BPMF, analytical computation of this posterior is intractable, due to complex coupling in the marginal distribution $p(\mathbf{r}^{(\mathrm{v})} | \Theta_0)$ (Blei et al., 2003; Salakhutdinov & Mnih, 2008).

4.1. Inference via Gibbs Sampling

In this work, we use a Gibbs sampling MCMC procedure (Geman & Geman, 1984) to draw samples of topic and parameter variables $\{(\mathbf{z}^{U(t)}, \mathbf{z}^{M(t)}, \Theta^{U(t)}, \Theta^{M(t)})\}_{t=1}^T$ from their joint posterior. Our use of conjugate priors ensures that each Gibbs conditional has a simple closed form.[4]

Alg. 1 displays the Gibbs sampling algorithm for the M^{3}F-TIB model; the M^{3}F-TIF Gibbs sampler is similar. Note that we choose to sample the topic parameters θ^U and θ^M rather than integrate them out as in a collapsed Gibbs sampler (see, e.g., Porteous et al. 2008). This decision allows us to sample the interaction-specific topic variables in parallel. Indeed, each loop in Alg. 1 corresponds to a block of parameters that can be sampled in parallel. In practice, such parallel computation yields substantial savings in sampling time for large-scale dyadic datasets.

[3]For work linking improved RMSE with better top-K recommendation rankings, see Koren (2008).

[4]See the Supplementary Information at the authors' websites for the exact conditional distributions.

Table 1. 1M MovieLens and EachMovie RMSE scores for varying static factor dimensionalities and topic counts for both M^{3}F models. All scores are averaged across 3 standardized cross-validation splits. Parentheses indicate topic counts (K^U, K^M). For M^{3}F-TIF, $\tilde{D} = 2$ throughout. L&U (2009) refers to (Lawrence & Urtasun, 2009). Best results for each D are boldened. Asterisks indicate significant improvement over BPMF under a one-tailed paired t-test with level 0.05.

	1M MovieLens				EachMovie			
Method	**D=10**	**D=20**	**D=30**	**D=40**	**D=10**	**D=20**	**D=30**	**D=40**
BPMF	0.8695	0.8622	0.8621	0.8609	1.1229	1.1212	1.1203	1.1163
M^{3}F-TIB (1,1)	0.8671	0.8614	0.8616	0.8605	1.1205	1.1188	1.1183	1.1168
M^{3}F-TIF (1,2)	0.8664	0.8629	0.8622	0.8616	1.1351	1.1179	1.1095	1.1072
M^{3}F-TIF (2,1)	0.8674	0.8605	0.8605	0.8595	1.1366	1.1161	1.1088	1.1058
M^{3}F-TIF (2,2)	**0.8642**	**0.8584***	0.8584	0.8592	1.1211	1.1043	1.1035	1.1020
M^{3}F-TIB (1,2)	0.8669	0.8611	0.8604	0.8603	1.1217	1.1081	1.1016	1.0978
M^{3}F-TIB (2,1)	0.8649	0.8593	**0.8581***	**0.8577***	1.1186	1.1004	1.0952	1.0936
M^{3}F-TIB (2,2)	0.8658	0.8609	0.8605	0.8599	**1.1101***	**1.0961***	**1.0918***	**1.0905***
L&U (2009)	0.8801 (RBF)		0.8791 (Linear)		1.1111 (RBF)		1.0981 (Linear)	

4.2. Prediction

Given posterior samples of parameters, we can approximate the true predictive distribution by the Monte Carlo expectation

$$\hat{p}(\mathbf{r}^{(\mathrm{h})}|\mathbf{r}^{(\mathrm{v})}, \Theta_0) = \frac{1}{T}\sum_{t=1}^{T}\sum_{\mathbf{z}^U,\mathbf{z}^M} p(\mathbf{z}^U, \mathbf{z}^M|\Theta^{U(t)}, \Theta^{M(t)}) \\ p(\mathbf{r}^{(\mathrm{h})}|\mathbf{z}^U, \mathbf{z}^M, \Theta^{U(t)}, \Theta^{M(t)}, \Theta_0), \quad (1)$$

where we have integrated over the unknown topic variables. Eq. 1 yields the following posterior mean prediction for each user-item pair under the M^{3}F-TIB model:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)} + \sum_{k=1}^{K^M} c_u^{k(t)}\theta_{jk}^{M(t)} + \sum_{i=1}^{K^U} d_j^{i(t)}\theta_{ui}^{U(t)}\right).$$

Under the M^{3}F-TIF model, posterior mean prediction takes the form

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)} + \sum_{i=1}^{K^U}\sum_{k=1}^{K^M}\theta_{ui}^{U(t)}\theta_{jk}^{M(t)}\mathbf{c}_u^{k(t)}\cdot\mathbf{d}_j^{i(t)}\right).$$

5. Experimental Evaluation

We evaluate our models on several movie rating collaborative filtering datasets including the Netflix Prize dataset[5], the EachMovie dataset, and the 1M and 10M MovieLens Datasets[6]. The Netflix Prize dataset contains 100 million ratings in $\{1, \ldots, 5\}$ distributed across 17,770 movies and 480,189 users. The EachMovie dataset contains 2.8 million ratings in $\{1, \ldots, 6\}$ distributed across 1,648 movies and 74,424 users. The 1M MovieLens dataset has 6,040 users, 3,952 movies, and 1 million ratings in $\{1, \ldots, 5\}$. The 10M MovieLens dataset has 10,681 movies, 71,567 users, and 10 million ratings on a .5 to 5 scale with half-star increments. In all experiments, we set W_0 equal to the identity matrix, ν_0 equal to the number of static matrix factors, μ_0 equal to the all-zeros vector, χ_0 equal to the mean rating in the data set, and $(\lambda_0, \sigma^2, \sigma_0^2) = (10, .5, .1)$. For M^{3}F-TIB experiments, we set $(c_0, d_0, \alpha) = (0, 0, 10000)$, and for M^{3}F-TIF, we set $\tilde{W}_0$ equal to the identity matrix, $\tilde{\nu}_0$ equal to the number of topic-indexed factors, $\tilde{\mu}_0$ equal to the all-zeros vector, and $(\tilde{D}, \xi_0, \alpha, \tilde{\lambda}_0) = (2, 0, 10, 10000)$. Free parameters were selected by grid search on an EachMovie hold-out set, disjoint from the test sets used for evaluation. Throughout, reported error intervals are of plus or minus one standard error from the mean.

5.1. 1M MovieLens and EachMovie Datasets

We first evaluated our models on the smaller datasets, 1M MovieLens and EachMovie. We conducted the "weak generalization" ratings prediction experiment of Marlin (2004), where, for each user in the training set, a single rating is withheld for the test set. All reported results are averaged over the same 3 random train-test splits used in (Marlin, 2003; 2004; Rennie & Srebro, 2005; DeCoste, 2006; Park & Pennock, 2007; Lawrence & Urtasun, 2009). Our Gibbs samplers were

[5] http://www.netflixprize.com/

[6] http://www.grouplens.org/

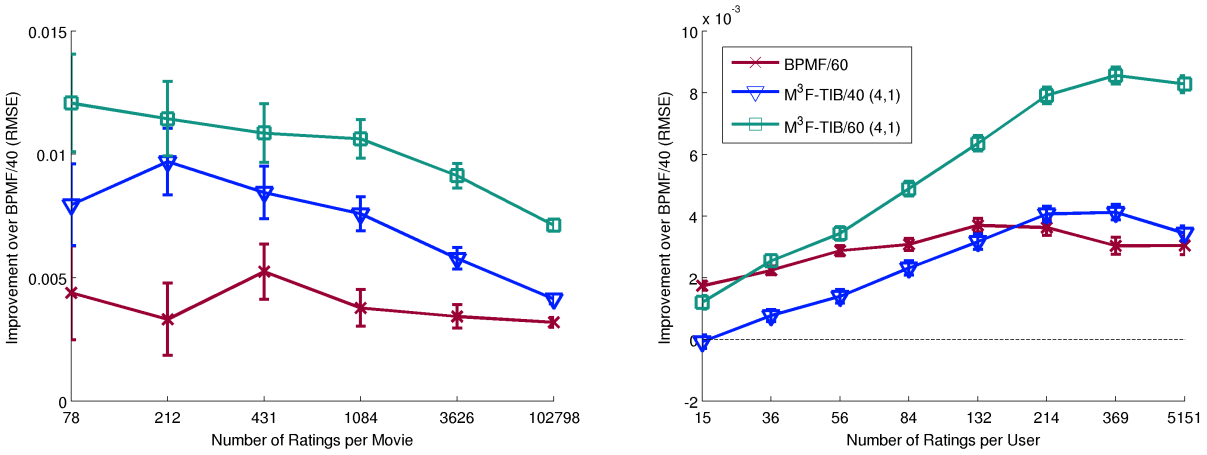

Figure 2. RMSE improvements over BPMF/40 on the Netflix Prize as a function of movie or user rating count. Left: Improvement as a function of movie rating count. Each x-axis label represents the average rating count of 1/6 of the movie base. Right: Improvement over BPMF as a function of user rating count. Each bin represents 1/8 of the user base.

initialized with draws from the prior and run for 3000 samples for M^{3}F-TIB and 512 samples for M^{3}F-TIF. No samples were discarded for "burn-in."

Table 1 reports the predictive performance of our models for a variety of static factor dimensionalities (D) and topic counts (K^U, K^M). We compared all models against BPMF as a baseline by running the M^{3}F-TIB model with K^U and K^M set to zero. For comparison with previous results that report the normalized mean average error (NMAE) of Marlin (2004), we additionally ran M^{3}F-TIB with $(D, K^U, K^M) = (300, 2, 1)$ on EachMovie and achieved a weak RMSE of ($\mathbf{1.0878} \pm 0.0025$) and a weak NMAE of ($\mathbf{0.4293} \pm 0.0013$).

On both the EachMovie and the 1M MovieLens datasets, both M^{3}F models systematically outperformed the BPMF baseline for almost every setting of latent dimensionality and topic counts. For $D = 20$, increasing K^U to 2 provided a boost in accuracy for both M^{3}F models equivalent to doubling the number of BPMF static factor parameters ($D = 40$). We also found that the M^{3}F-TIB model outperformed the more recent Gaussian process matrix factorization model of Lawrence & Urtasun (2009).

The results indicate that the mixed-membership component of M^{3}F offers greater predictive power than simply increasing the dimensionality of a pure latent factor model. While the M^{3}F-TIF model sometimes failed to outperform the BPMF baseline due to overfitting, the M^{3}F-TIB model always outperformed BPMF regardless of the setting of K^U, K^M, or D. Note that the increase in the number of parameters from the BPMF model to the M^{3}F models is independent of D (M^{3}F-TIB requires $(U + M)(K^U + K^M)$ more parameters than BPMF with equal D), and therefore the ratio of the number of parameters of BPMF and M^{3}F approaches 1 if D increases while K^U, K^M, and $\tilde{D}$ are held fixed. Nonetheless, the modeling of joint contextual bias in the M^{3}F-TIB model continues to improve predictive performance even as D increases, suggesting that the M^{3}F-TIB model is capturing aspects of the data that are not captured by a pure latent factor model.

Finally, because the M^{3}F-TIB model offered superior performance to the M^{3}F-TIF model in most experiments, we focus on the M^{3}F-TIB model in the remainder of this section.

5.2. 10M MovieLens Dataset

For the larger datasets, we initialized the Gibbs samplers with MAP estimates of **a** and **b** under simple Gaussian priors, which we trained with stochastic gradient descent. This is similar to the PMF initialization scheme of Salakhutdinov & Mnih (2008). All other parameters were initialized to their model means.

For the 10M MovieLens dataset, we averaged our results across the r_a and r_b train-test splits provided with the dataset after removing those test set ratings with no corresponding item in the training set. For comparison with the Gaussian process matrix factorization model of Lawrence & Urtasun (2009), we adopted a static factor dimensionality of $D = 10$. Our M^{3}F-TIB model with $(K^U, K^M) = (4, 1)$ achieved an RMSE of ($\mathbf{0.8447} \pm 0.0095$), representing a significant improvement ($p = 0.034$) over BPMF with RMSE ($\mathbf{0.8472} \pm 0.0093$) and a substantial increase in accuracy over the Gaussian process model with RMSE

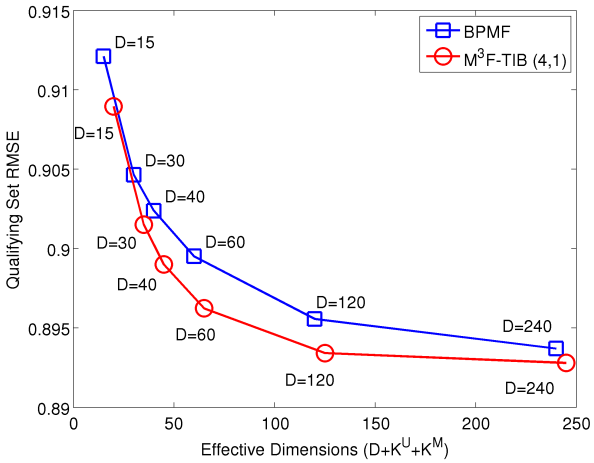

Figure 3. RMSE performance of BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$ on the Netflix Prize Qualifying set as a function of the number of parameters modeled per user or item.

($\mathbf{0.8740} \pm 0.0197$).

5.3. Netflix Prize Dataset

The unobserved ratings for the 100 million dyad Netflix Prize dataset are partitioned into two standard sets, known as the Quiz Set and the Test Set. Prior to September of 2009, public evaluation was only available on the Quiz Set, and, as a result, most prior published "test set" results were evaluated on the Quiz Set. In Table 2, we compare the performance of BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$ on the Quiz Set, the Test Set, and on their union (the Qualifying Set), across a wide range of static dimensionalities. We also report running times of our Matlab/MEX implementation on dual quad-core 2.67GHz Intel Xeon CPUs. We used the initialization scheme described in Section 5.2 and ran the Gibbs samplers for 500 iterations.

In addition to outperforming the BPMF baselines of comparable dimensionality, the M^{3}F-TIB models routinely proved to be more accurate than higher dimensional BPMF models with longer running times and many more learned parameters. This major advantage of M^{3}F modeling is highlighted in Figure 3, which plots error as a function of the number of parameters modeled per user or item $(D + K^U + K^M)$.

To determine where our models were providing the most improvement over BPMF, we divided the Qualifying Set into bins based on the number of ratings associated with each user and movie in the database. Figure 2 displays the improvements of BPMF/60, M^{3}F-TIB/40, and M^{3}F-TIB/60 over BPMF/40 as a function of the number of user or movie ratings. Consistent with our expectations, we found that adopting an M^{3}F model yielded improved accuracy for movies of small rating counts, with the greatest improvement over BPMF occurring for those high-variance movies with relatively few ratings. Moreover, the improvements realized by either M^{3}F-TIB model uniformly dominated the improvements realized by BPMF/60 across movie rating counts. At the same time, we found that the improvements of the M^{3}F-TIB models were skewed toward users with larger rating counts.

Table 2. Netflix Prize results for BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$. Hidden ratings are partitioned into Quiz and Test sets; the Qualifying set is their union. Best results in each block are boldened. Reported times are average running times per sample.

Method	Test	Quiz	Qual	Time
BPMF/15	0.9125	0.9117	0.9121	27.8s
TIB/15	**0.9093**	**0.9086**	**0.9090**	46.3s
BPMF/30	0.9049	0.9044	0.9047	38.6s
TIB/30	**0.9018**	**0.9012**	**0.9015**	56.9s
BPMF/40	0.9029	0.9026	0.9027	48.3s
TIB/40	**0.8992**	**0.8988**	**0.8990**	70.5s
BPMF/60	0.9004	0.9001	0.9002	94.3s
TIB/60	**0.8965**	**0.8960**	**0.8962**	97.0s
BPMF/120	0.8958	0.8953	0.8956	273.7s
TIB/120	**0.8937**	**0.8931**	**0.8934**	285.2s
BPMF/240	0.8939	0.8936	0.8938	1152.0s
TIB/240	**0.8931**	**0.8927**	**0.8929**	1158.2s

5.3.1. M^{3}F & The *Napoleon Dynamite* Effect

In our introduction to the M^{3}F-TIB model we discussed the joint contextual bias as a potential solution to the problem of making predictions for movies that have high variance. To investigate whether or not M^{3}F-TIB achieved progress towards this goal, we analyzed the correlation between the improvement in RMSE over the BPMF baseline and the variance of ratings for the 1000 most popular movies in the database. While the improvements for BPMF/60 were not significantly correlated with movie variance ($\rho = -0.016$), the improvements of the M^{3}F-TIB models were strongly correlated with $\rho = 0.117 (p < 0.001)$ and $\rho = 0.15$ ($p < 10^{-7}$) for the $(40, 4, 1)$ and $(60, 4, 1)$ models, respectively. These results indicate that a strength of the M^{3}F-TIB model lies in the ability of the topic-indexed biases to model variance in user biases toward specific items.

Table 3. Top 200 Movies from the Netflix Prize dataset with the highest and lowest cross-topic variance in $\mathbb{E}(d_j^i|\mathbf{r}^{(\mathrm{v})})$. Reported intervals are of the mean value of $\mathbb{E}(d_j^i|\mathbf{r}^{(\mathrm{v})})$ plus or minus one standard deviation.

Movie Title	$\mathbb{E}(d_j^i\|\mathbf{r}^{(\mathrm{v})})$
Napoleon Dynamite	-0.11 ± 0.93
Fahrenheit 9/11	-0.06 ± 0.90
Chicago	-0.12 ± 0.78
The Village	-0.14 ± 0.71
Lost in Translation	-0.02 ± 0.70
LotR: The Fellowship of the Ring	0.15 ± 0.00
LotR: The Two Towers	0.18 ± 0.00
LotR: The Return of the King	0.24 ± 0.00
Star Wars: Episode V	0.35 ± 0.00
Raiders of the Lost Ark	0.29 ± 0.00

To further illuminate this property of the model, we computed the posterior expectation of the movie bias parameters, $\mathbb{E}(\mathbf{d}_j|\mathbf{r}^{(\mathrm{v})})$, for the 200 most popular movies in the database. For these movies, the variance of $\mathbb{E}(d_j^i|\mathbf{r}^{(\mathrm{v})})$ across topics and the variance of the ratings of these movies were very strongly correlated ($\rho = 0.682, p < 10^{-10}$). The five movies with the highest and lowest variance in $\mathbb{E}(d_j^i|\mathbf{r}^{(\mathrm{v})})$ across topics are shown in Table 3. The results are easily interpretable, with high-variance movies such as *Napoleon Dynamite* dominating the high-variance positions and universally acclaimed blockbusters dominating the low-variance positions.

6. Conclusion

In this work, we developed a fully Bayesian dyadic data prediction framework for integrating the complementary approaches of discrete mixed membership modeling and continuous latent factor modeling. We introduced two Mixed Membership Matrix Factorization models, developed MCMC inference procedures, and evaluated our methods on the EachMovie, MovieLens, and Netflix Prize datasets. On each dataset, we found that M^{3}F-TIB significantly outperformed BPMF and other state-of-the-art baselines, even when fitting fewer parameters. We further discovered that the greatest performance improvements occurred for the high-variance, sparsely-rated items, for which accurate DDP is typically the hardest.

Acknowledgments

LM was supported by an NDSEG Fellowship. DW was supported by a NSF fellowship.

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic blockmodels. *JMLR*, 9:1981–2014, 2008.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

DeCoste, D. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *ICML*, 2006.

Geman, S. and Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Hofmann, T., Puzicha, J., and Jordan, M. I. Learning from dyadic data. In *NIPS*, 1999.

Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, 2008.

Lawrence, N.D. and Urtasun, R. Non-linear matrix factorization with Gaussian processes. In *ICML*, 2009.

Marlin, B. Modeling user rating profiles for collaborative filtering. In *NIPS*, 2003.

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

Park, S-T. and Pennock, D. M. Applying collaborative filtering techniques to movie search for better ranking and browsing. In *KDD*, 2007.

Porteous, I., Bart, E., and Welling, M. Multi-HDP: A non parametric Bayesian model for tensor factorization. In *AAAI*, 2008.

Rennie, J. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, 2005.

Salakhutdinov, R. and Mnih, A. Probabilistic matrix factorization. In *NIPS*, 2007.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Marlo. In *ICML*, 2008.

Takács, G., Pilászy, I., Németh, B., and Tikk, D. Scalable collaborative filtering approaches for large recommender systems. *JMLR*, 10:623–656, 2009.

Thompson, C. If you liked this, you're sure to love that. *New York Times Magazine*, November 2008.