

Probabilistic Latent Factor Model for Collaborative Filtering with Bayesian Inference

Jiansheng Fang^{1,2,3}, Xiaoqing Zhang², Yan Hu², Yanwu Xu⁴, Ming Yang³, Jiang Liu^{*2,4}

¹School of computer science and technology, Harbin Institute of Technology, Harbin 150001, China

Email: 11949039@mail.sustech.edu.cn

²Department of computer science and engineering, Southern University of Science and Technology, Shenzhen 518055, China

Email: liuj@sustech.edu.cn

³CVTE Research, Guangzhou 510530, China

Email: yangming@cvte.com

⁴Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Ningbo 315201, China

Abstract—Latent Factor Model (LFM) is one of the most successful methods for Collaborative filtering (CF) in the recommendation system, in which both users and items are projected into a joint latent factor space. Base on matrix factorization applied usually in pattern recognition, LFM models user-item interactions as inner products of factor vectors of user and item in that space and can be efficiently solved by least square methods with optimal estimation. However, such optimal estimation methods are prone to overfitting due to the extreme sparsity of user-item interactions. In this paper, we propose a Bayesian treatment for LFM, named Bayesian Latent Factor Model (BLFM). Based on observed user-item interactions, we build a probabilistic factor model in which the regularization is introduced via placing prior constraint on latent factors, and the likelihood function is established over observations and parameters. Then we draw samples of latent factors from the posterior distribution with Variational Inference (VI) to predict expected value. We further make an extension to BLFM, called BLFMBias, incorporating user-dependent and item-dependent biases into the model for enhancing performance. Extensive experiments on the movie rating dataset show the effectiveness of our proposed models by compared with several strong baselines.

I. INTRODUCTION

One of the popular approaches to recommender systems [1], [2] which meet users' personalized needs for items is Collaborative filtering (CF) [3], [4], [5], [6], [7], [8] which heavily relies on the past user behavior (e.g., ratings and clicks). A major appeal of CF is that CF techniques require no domain knowledge and avoid the need for extensive data collection [9], [10]. Many online services have widely adopted personalized recommendation to enhance user satisfaction. One of the most successful realizations to CF is Latent Factor Model (LFM) in which both user and item are projected into a joint latent factor space. The main idea behind LFM is to model user-item interactions as inner products of factor vectors of user and item in the space.

LFM is capable of learning a joint latent factor space from user-item interactions which include high-quality explicit rating [11] and abundant implicit feedback [12], [13], [14]. Due to scalability and flexibility of LFM [15], much research

effort have been devoted to enhancing its recommendation effect [16], [17], [18]. For example, LFM incorporates side information into the user-item interaction function to improve performance [19], [20], [21]. Furthermore, probabilistic factor models have been proposed to overcome the sparseness and solve the non-convex optimization problem [22], [23], [24], [25].

Conventional probabilistic factor models [26], [27], [28], [29] construct the likelihood function over the rating matrix. However, in our model, the likelihood function is given over only observed user-item interactions, and the benefit is that we can incorporate additional information into the model, such as biases, attributes of a user. Both of additional information and latent factors are placed prior constraint for each user and item. Because the posterior distribution is hard to exact inference, several probabilistic factor models use Markov chain Monte Carlo (MCMC) [30], [31], [32] to conduct approximate posterior inference. MCMC provides guarantee of producing exact samples from the target distribution. An alternative method of approximate posterior inference is Variational Inference (VI) which tends to be faster and easier to scale to large data [33], [34], [35]. Considering run efficiency, we adopt VI to draw samples from the posterior distribution.

Training LFM amounts to fit the inner products of factor vectors to observed user-item interactions over the given loss function. Optimal estimation methods are widely used to estimate model parameters, such as Maximum Likelihood (ML) [36], Maximum-A-Posteriori (MAP) [37]. Although optimal estimation methods can be worked efficiently on large datasets, they are easy to cause overfitting due to extreme sparsity of user-item interactions. Hence we explore the Bayesian treatment of LFM based on observed user-item interactions to alleviate overfitting. The main contributions of this work are as follows:

- We propose a Bayesian treatment for LFM, named Bayesian Latent Factor Model (BLFM), in which a probabilistic factor model is given over latent factors and observed data, and the values of model parameters are drawn from posterior distribution.

* Corresponding author: Jiang Liu

- We further make an extension to BLFM, called BLFMBias, incorporating user-dependent and item-dependent biases into model to enhance performance. This shows that BLFM can be extended to incorporate additional information to solve cold start problem in the recommendation system.
- Extensive experiments are performed on the movie rating dataset to demonstrate the effectiveness of our proposed method.

II. PROBLEM STATEMENT

In this section, we formalize the problem and introduce optimal estimation method for LFM.

A. Rating Prediction

Let m and n denote the number of users and items, respectively. Given observed rating r_{ui} of user u on item i , let \mathbf{D} denotes the set of tuple (u, i, r_{ui}) . The recommendation problem is formulated as a problem of predicting unobserved rating. Formally, the problem of rating prediction is to learn

$$\hat{r}_{ui} = f(r_{ui}|\Theta), \quad (1)$$

where \hat{r}_{ui} denotes the estimated rating of user u on item i , and Θ denotes model parameters. The function f establishes the link between model parameters and observed rating.

B. Optimal Estimation

In LFM, rating is modeled as inner product of factor vectors with dimension K . Let \mathbf{p}_u and \mathbf{q}_i denote the factor vector for user u and item i , respectively. Equation (1) can be recast as follow:

$$\hat{r}_{ui} = f(r_{ui}|\mathbf{p}_u, \mathbf{q}_i) = \mathbf{p}_u^T \mathbf{q}_i = \sum_{k=1}^K p_{uk} q_{ik}, \quad (2)$$

where k is the dimension of factor vector. Training LFM is to fit the inner products of factor vectors \mathbf{p}_u and \mathbf{q}_i to the observed rating r_{ui} .

The optimal estimation method learns the model parameters of LFM by minimizing the regularized squared error

$$\min_{\mathbf{p}_*, \mathbf{q}_*} \sum_{(u, i, r_{ui}) \in \mathbf{D}} (r_{ui} - \mathbf{p}_u^T \mathbf{q}_i)^2 + \lambda(\|\mathbf{p}_u\|^2 + \|\mathbf{q}_i\|^2), \quad (3)$$

where λ is the regularization parameter. Given the loss function of Equation (3), gradient descent is usually applied to optimize parameters.

Such optimal estimation method can be interpreted as MAP. MAP can be seen as a regularization of ML estimation, both of MAP and ML can be used to make single optimal estimate over observed data. However, the regularization parameters should be tuned carefully to overcome overfitting. For improving generalization performance, we apply Bayesian parameter estimation method to estimate factor vectors \mathbf{p}_u and \mathbf{q}_i .

III. BAYESIAN LATENT FACTOR MODEL (BLFM)

In this section, we present the details of BLFM and its extend BLFMBias. First, we define a probabilistic factor model, in which prior assumption is introduced for model parameters and likelihood function is build on observations and parameters. Second, based on the defined probabilistic factor model, we draw samples from posterior distribution of model parameters to estimate the expected rating.

A. Prior Assumption

In fact, we recast Equation (3) by building a probabilistic factor model with observed rating being observations, factor vectors and biases terms being parameters. In this model, we introduce regularization by placing prior constraint over model parameters.

BLFM. The prior distributions over the factor vectors are assumed to be Gaussian:

$$p(p_{uk}|\mu_{uk}, \sigma_{uk}) \sim \mathcal{N}(\mu_{uk}, \sigma_{uk}), \quad (4)$$

$$p(q_{ik}|\mu_{ik}, \sigma_{ik}) \sim \mathcal{N}(\mu_{ik}, \sigma_{ik}).$$

Here, the k^{th} dimension of factor vector have its independent hyperparameters, mean μ_{uk}, μ_{ik} and variance σ_{uk}, σ_{ik} . Such that amounts to set the weight of each latent factor. It is flexible enough to control the complexity of model and adjust the effect of regularization.

BLFMBias. The biases tend to capture the vibration of explicit rating of user on item. We incorporate the biases terms of user and item into model to denoise. Let \bar{r} denotes the overall average rating, the prior for bias b_u of user u , bias b_i of item i in BLFMBias are defined as:

$$p(b_u|\bar{r}) \sim \mathcal{N}(0, \bar{r}), \quad (5)$$

$$p(b_i|\bar{r}) \sim \mathcal{N}(0, \bar{r}).$$

The parameters b_u and b_i indicate the observed deviations of user u and item i . Thus, the biases terms is defined as:

$$b_{ui} = \bar{r} + b_u + b_i. \quad (6)$$

B. Likelihood Function

The likelihood function establishes the relation between observations and parameters.

BLFM. In BLFM, we fit the inner product of factor vectors to approximate the observed rating. The likelihood function of BLFM is defined as:

$$p(r_{ui}|\mathbf{p}_u, \mathbf{q}_i) \sim \mathcal{N}\left(\sum_{k=1}^K p_{uk} q_{ik}, \bar{r}\right). \quad (7)$$

A graphical model corresponding to this generative process is shown at the Figure 1.

BLFMBias. In BMFBias, we fit the sum of inner product of factor vectors and biases terms to the observed rating. The likelihood function of BMFBias is defined as:

$$p(r_{ui}|\mathbf{p}_u, \mathbf{q}_i, b_u, b_i) \sim \mathcal{N}\left(\sum_{k=1}^K p_{uk} q_{ik} + b_{ui}, \bar{r}\right). \quad (8)$$

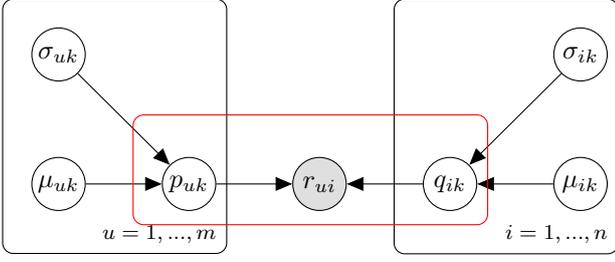


Fig. 1. The graphical model of BLFM

A graphical model corresponding to this generative process is shown at the Figure 2.

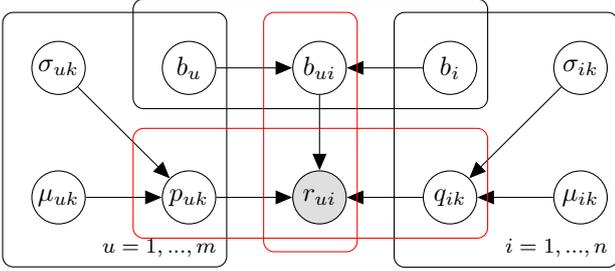


Fig. 2. The graphical model of BLFMBias.

C. Variational Inference (VI)

While exact inference of the posterior distribution over the parameters given the observed rating is intractable, we resort to use VI for posterior approximate inference. The goal of VI is to approximate a conditional density of latent variables given observed variables [38]. VI tries to find a family of approximate density and turns the inference problem into an optimization problem by minimizing the Kullback-Leibler (KL) divergence to the exact posterior.

BLFM. The KL divergence is an information-theoretical measure of proximity between two density [39]. First, Let us give the posterior density $p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})$ of BLFM. Then, we specify a family Q over the factor vectors $\mathbf{p}_u, \mathbf{q}_i$, and each $q(z) \in Q$ is a candidate approximation to the exact posterior density. Last, our purpose is to solve the following optimization:

$$q^* = \arg \min_{q(\mathbf{p}_u, \mathbf{q}_i) \in Q} KL(q(\mathbf{p}_u, \mathbf{q}_i) || p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})). \quad (9)$$

Within the family Q , the goal is to find the best approximation $q^*(\mathbf{p}_u, \mathbf{q}_i)$ of the posterior $p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})$. KL divergence is the expected value of the difference between target density and candidate density, the expectation representation of KL can be written as:

$$KL(q(\mathbf{p}_u, \mathbf{q}_i) || p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})) = \mathbb{E}[\log q(\mathbf{p}_u, \mathbf{q}_i)] - \mathbb{E}[\log p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})]. \quad (10)$$

According to the Bayesian formula, the posterior (conditional) density $p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})$ can be represented as:

$$p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D}) = \frac{p(\mathbf{p}_u, \mathbf{q}_i, \mathbf{D})}{p(\mathbf{D})}. \quad (11)$$

Here, the numerator $p(\mathbf{p}_u, \mathbf{q}_i, \mathbf{D})$ is a joint density based on the factor vectors and the observed rating, and the denominator $p(\mathbf{D})$ contains the marginal density of the observations, also called the evidence. This evidence is calculated by marginalizing out the latent variables from the joint density, and this evidence integral is intractable.

Because the evidence is hard to compute, VI optimize an alternative objective function:

$$ELBO(q) = \mathbb{E}[\log p(\mathbf{p}_u, \mathbf{q}_i, \mathbf{D})] - \mathbb{E}[\log q(\mathbf{p}_u, \mathbf{q}_i)]. \quad (12)$$

This function is called the evidence lower bound (ELBO). Maximizing the ELBO is equivalent to minimizing the KL divergence.

According to the Bayesian formula, the joint density $p(\mathbf{p}_u, \mathbf{q}_i, r_{ui})$ can be represented as:

$$p(\mathbf{p}_u, \mathbf{q}_i, \mathbf{D}) = p(\mathbf{D} | \mathbf{p}_u, \mathbf{q}_i) p(\mathbf{p}_u) p(\mathbf{q}_i). \quad (13)$$

Here, $p(\mathbf{D} | \mathbf{p}_u, \mathbf{q}_i)$ denotes likelihood function, and the $p(\mathbf{p}_u), p(\mathbf{q}_i)$ denote the prior distribution of $\mathbf{p}_u, \mathbf{q}_i$, respectively.

With integrating the prior distribution (4), the likelihood function (conditional density) (7), and the ELBO function (12), we can approximate posterior over the factor vectors $\mathbf{p}_u, \mathbf{q}_i$ given the observed rating \mathbf{D} . Then we draw samples of model parameters from the posterior $p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D})$. Given a pair (u, i) , the expected value of rating \hat{y}_{ui} is predicted as:

$$\hat{r}_{ui} = \int_{\mathbf{p}_u, \mathbf{q}_i} p(r_{ui} | \mathbf{p}_u, \mathbf{q}_i) p(\mathbf{p}_u, \mathbf{q}_i | \mathbf{D}) d\{\mathbf{p}_u, \mathbf{q}_i\}. \quad (14)$$

BLFMBias. In BLFMBias, we incorporate the parameters b_u and b_i into Bayesian model. We use Θ unified denotes the parameters of posterior density in BLFMBias, including $\mathbf{p}_u, \mathbf{q}_i, b_u, b_i$. The objective function of VI is:

$$ELBO(q) = \mathbb{E}[\log p(\Theta, \mathbf{D})] - \mathbb{E}[\log q(\Theta)]. \quad (15)$$

Here, our aim is to find the best $q(\Theta)$ to approximate the posterior density $p(\Theta | \mathbf{D})$ by maximizing the ELBO. And the joint density $p(\Theta, \mathbf{D})$ of BLFMBias can be represented as:

$$p(\Theta, \mathbf{D}) = p(\mathbf{D} | \Theta) p(\Theta). \quad (16)$$

Here, $p(\Theta)$ is the prior distribution of parameters in BMFBias, and $p(\mathbf{D} | \Theta)$ is the likelihood function given the observed rating. As same with BLFM, we can predict expectation rating for user u and item i :

$$\hat{r}_{ui} = \int_{\Theta} p(r_{ui} | \Theta) p(\Theta | \mathbf{D}) d\Theta. \quad (17)$$

The posterior probability is a conditional probability conditioned on observed rating. For BLFM, the parameter of posterior probability is latent factors. For BLFMBias, the parameter of posterior probability is latent factors and biases terms. By learning probabilistic factor model, we draw samples of parameters from posterior probability to predict expected value of rating.

IV. EXPERIMENTS

In this section, the following research questions will be answered by conducting corresponding experiments.

- RQ1** Do our proposed BLFM outperforms the state-of-the-art methods for rating prediction?
- RQ2** By incorporating biases into BLFM, how is our proposed BLFMBias performance?
- RQ3** Is Bayesian treatment more robust in alleviating overfitting?
- RQ4** Do the samples of parameter from posterior probability converge by applying VI?
- RQ5** Which hyperparameter most affect performance for training BLFM model?

In what follows, we first present the experimental settings, followed by answering the above five research question.

A. Experimental Settings

TABLE I
STATISTICS OF THE EVALUATION DATASETS.

Dataset	Interaction	User	Item	Sparsity
MovieLens	1,000,209	6,040	3,706	95.53%

Dataset. We experimented with one public accessible dataset: MovieLens¹. The characteristic of the dataset are summarized in Table I. This movie rating dataset has been widely used to evaluate LFM algorithms. We choose the version containing about one million ratings, where each user has at least 20 ratings. The ratings are given on a scale of 0 to 5 stars with increments of 0.5 stars. The mean rating over all users and movies is 3.58 and the variance is 1.24.

Evaluation Metric. We adopted Root Mean Square Error (RMSE) to evaluate the performance of rating prediction. RMSE is widely used in literature [40], [28], [29], [24]. For each user, we held out his latest interaction as the test set and utilized the remaining data for training.

Baselines. In order to evaluate the utility of our models, four strong LFM baselines were selected as follows:

- **SVD**[9]. SVD trains model directly only on the observed ratings while avoiding overfitting through introducing regularization, and applied gradient descent to solve the optimization problem.
- **PMF**[24]. Probabilistic Matrix Factorization (PMF) adopts a probabilistic linear model with Gaussian observation noise, and performed well on very sparse and imbalanced datasets by modeling the user preference matrix as as product of two lower-rank matrices.
- **BPMF**[29]. Bayesian Probabilistic Matrix Factorization (BPMF) implements PMF by using MCMC and further placed Gaussian-Wishart priors on the users and items feature vectors.
- **SVDBias**[9]. SVDBias is a fully-regularized gradient descent SVD model with user and item biases providing a competitive baseline.

Given observed user-item interactions, both of our BLFM and BLFMBias make rating prediction by drawing samples of parameters from posterior probability.

Parameters Setting. For SVD and SVDBias, weights are updated using a learning rate of 0.001, the regularization parameter is set to 0.01, and the prediction errors on train set were observed to stabilize after approximately 30 iterations. For PMF model, we set learning rate of 0.005, a momentum of 0.9, and max iterations of 200. For BPMF model, we initialized the two lower-rank matrices of zero, the mean of Gaussian-Wishart priors of zero, the degrees of freedom (parameter of Wishart distribution) of latent factor number, and max iterations of 100. We implemented our proposed methods based on Python². We fixed the parameters of Gaussian priors, the mean of zero and the variance of the average rating. On conducting VI, we fit target density 10000 times, and 2000 samples are drawn to calculate expectation rating. For all methods, we tested in different number of latent factors (8, 16, 32, 64) on the same experimental device and reported the best performance given corresponding parameters.

B. Performance Comparison (RQ1)

Table II shows the performance of RMSE with respect to the number of latent factors. On the whole, we can see that BLFM achieves the best performance, significantly outperforming SVD and PMF, slightly outperforming BPMF.

TABLE II
PERFORMANCE OF RMSE.

Movielens				
Factors	SVD	PMF	BPMF	BLFM
RMSE@8	1.0274	1.0256	0.9871	<u>0.9781</u>
RMSE@16	1.0256	1.0076	0.9840	<u>0.9836</u>
RMSE@32	1.0278	0.9999	0.9875	<u>0.9852</u>
RMSE@64	1.0141	0.9952	0.9867	<u>0.9826</u>

SVD. SVD and BLFM are trained directly on observed rating. Let us observe the performance on factors of 32. BMF model achieves an RMSE of 0.9852, compared to an RMSE of 1.0278 of SVD model, an improvement of over 4.1%. The result demonstrates that Bayesian prior constraint is better than parameter regularization and the expectation rating prediction is better than the single optimal rating prediction.

PMF. PMF model is trained by maximizing the log-posterior over the users and items features with fixed hyperparameters (i.e. the observation noise variance and prior variance) [29]. PMF model introduced prior constraint and learned model parameters over the rating matrix (included unobserved entries of rating matrix) with MAP method. Let us observe the performance on factors of 16. BLFM achieves an RMSE of 0.9836, compared to an RMSE of 1.0076 of PMF model, an improvement of over 2.38%.

¹<https://grouplens.org/datasets/movielens/1m/>

²<https://github.com/fjssharpword/fjsgcode2019/tree/master/BMF>

BPMF. BPMF model adopted fully Bayesian treatment of the PMF model and used MCMC to train model over the rating matrix. Compared to BPMF, BLFM applies Bayesian treatment in the same, but uses the VI to train model over the observed rating. Although BLFM only slightly outperforms BPMF, but it is faster than BPMF and is flexible to extend model. On training latent factors of 64, the run time of BPMF is over 10,000 seconds with executing sampler of 100 step, and BLFM approximates 4,000 seconds with fitting target posterior density of 10000 times. On average, the run time of training BPMF is double longer than BLFM.

C. Performance with Adding Biases (RQ2)

Table III shows the performance of RMSE with adding biases on different latent factors. Let us observe the performance on factors of 16. BLFMBias achieves an RMSE of 0.9397, compared to an RMSE of 0.9836 of BLFM, an improvement of over 4.46%. This result demonstrate that the performance can be improved by adding biases. Compared to an RMSE of 0.9570 of SVDBias model, BLFMBias also have an improvement of over 1.80%.

TABLE III
PERFORMANCE OF RMSE WITH ADDING BIASES.

Movielens			
Factors	BLFM	SVDBias	<i>BLFMBias</i>
RMSE@8	0.9781	0.9581	<u>0.9406</u>
RMSE@16	0.9836	0.9570	<u>0.9397</u>
RMSE@32	0.9852	0.9573	<u>0.9410</u>
RMSE@64	0.9826	0.9572	<u>0.9452</u>

D. Performance of Alleviating Overfitting (RQ3)

To make a comparison between the performance of the train set and testing set, we randomly select rating entries of the same size as the test set from the train set. Figure 3 shows that SVD and PMF achieves better on the validation set (green line) but worse on the test set (blue line), compared to BLFM. The length of dashed red line in the Figure 3 denotes the difference size of performance between validation set and test set. It can be concluded that the length of BLFM is shorter than PMF and the length of SVD is longer than PMF. On alleviating overfitting, this experiment clearly demonstrates that BLFM which applies prior constraint and Bayesian estimation method is best, next is PMF which applies prior constraint and MAP estimation method, and last is SVD which applies parameter regularization and optimal estimation method. This shows that BLFM with prior constraint and Bayesian estimation method is more robust against overfitting.

E. Convergence of Parameter Estimation (RQ4)

In BLFM, the value of parameters (latent factors) is drawn from approximating posterior distribution. As shown in Figure 4, we draw 500 samples (x-axis) from the approximation distribution for user 100 and item 100, respectively. We

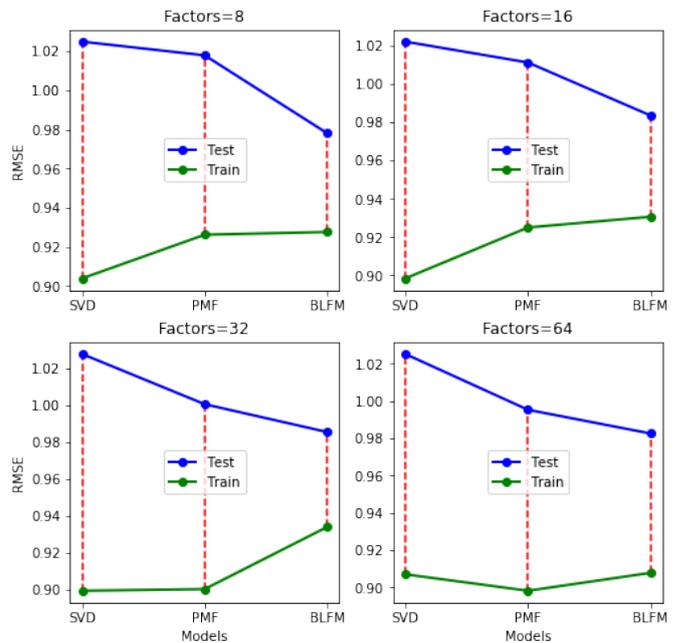


Fig. 3. Performance of RMSE on validation set and test set.

observe the value (y-axis) distribution (yellow circle point) of second dimension of latent factors 8, and calculate the average of these samples. We can see that the value between this mean (solid red line) and MAP optimized point (dashed blue line) is close. The value of MAP optimized point is obtained from PMF model. That the value of parameters of Bayesian estimation is close to the MAP estimation demonstrates that the expectation rating is valid and the estimation value of parameters is convergence.

F. Analysis of Hyperparameters (RQ5)

The objective of training BLFM is to fit the best approximated density to the target posterior density. Based on a moderate probabilistic model, we observe the training hyperparameters: the number of iterations and the number of samples. By comparing the blue line (iterations of 5000) with red line (iterations of 10000), green line (iterations of 15000) in Figure 5, we can demonstrate that the number of iterations increased can significantly improve the performance, but the effect gradually weak. Figure 5 also shows that the number of samples weakly affect the performance (x axis denotes the number of samples). This experiment shows that our model can achieve good performance without continuous enlarging the number of iteration. In fact, our model has comparable run time to an efficiently implemented SVD.

V. DISCUSSION AND RELATED WORK

Based on conducting extensive experiments, let us first summary the characteristic of our model, as shown in Table IV.

- **PM** denotes the method of parameter estimation. There have two types of PM, MAP (Optimal Estimation) and

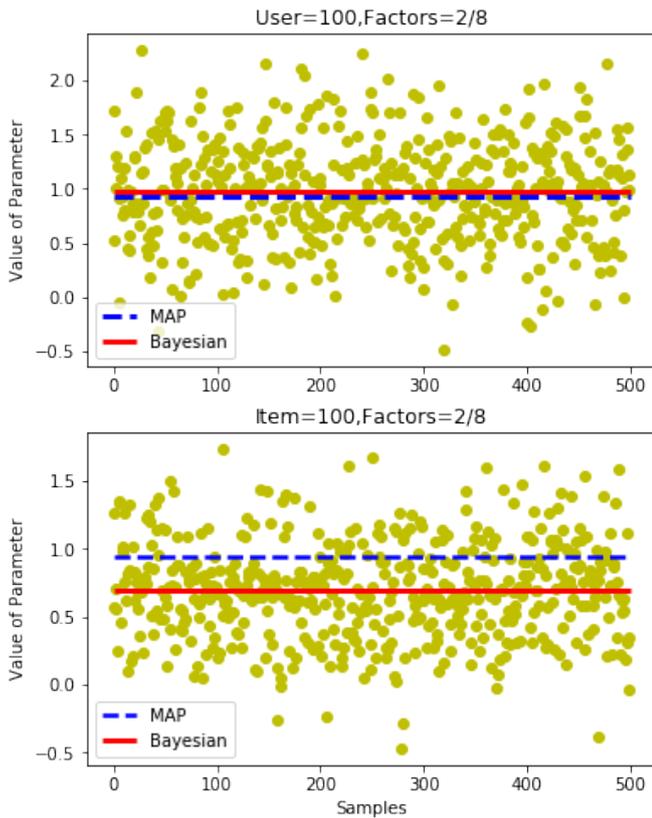


Fig. 4. Estimation of Bayesian and MAP.

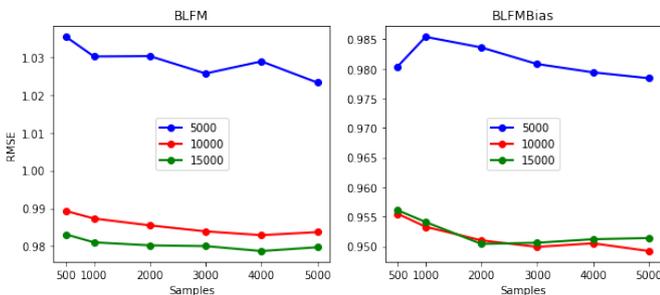


Fig. 5. Setting the number of fit iterations and samples.

TABLE IV
THE CHARACTERISTIC OF THE STATE-OF-THE-ART MODELS

Index	SVD	PMF	BPMF	BLFM
PM	MAP	MAP	Bayesian	<i>Bayesian</i>
RG	Parameter	Parameter	Prior	<i>Prior</i>
RP	medium	lower	higher	<i>medium</i>
FS	Good	Bad	Bad	<i>Good</i>

Bayesian (Expectation Estimation). Compared to the SVD and PMF with MAP method, the BPMF and BLFM applied Bayesian method achieves better performance. The experiment of RQ1 can demonstrate this. The experiment of RQ4 also demonstrate the validity of expectation estimation.

- **RG** denotes the method of regularization, includes introducing prior constraint and placing parameter control. On alleviating overfitting, the experiment of RQ3 demonstrate the effectiveness of Bayesian method which includes expectation estimation and prior constraint.
- **RT** denotes the run time for achieving result of Table II. The experiment of RQ1 demonstrates that BLFM with VI is faster than BPMF applied MCMC. The experiment of RQ5 demonstrates that the run time of training BLFM model is mainly decided by the number of iterations.
- **FS** denotes flexibility and scalability of model. Due to the model trained on observed rating entry, the experiment of RQ2 demonstrate that BLFM can incorporate biases into model to enhance the performance.

With rating prediction and LFM method on CF have been widely researched in early literature [41], [42], [43], [44], a variety of probabilistic factor models [45], [46], [47], [24] and Bayesian method [48], [22], [49], [28], [29], [40] have been proposed to enhance the performance. For estimating parameters of probabilistic model, Expectation Maximization (EM) [50], [51], [52] and MAP were used in early works. Zhang *et al* introduce a novel adaptation of the EM algorithms to learn the parameters of a prediction model for personalised content-based prediction[53]. Stern *et al* instead used Expectation Propagation and Variational Message Passing to learn a model using both ratings data and content [54]. Related to applying Bayesian method to solve the probabilistic factor model, the early pioneer works are described as follows:

- Salakhutdinov *et al.* [29] proposed BPMF based on MCMC and rating matrix which we have discussed detail in experiments.
- Lim *et al.* proposed a Bayesian approach to alleviate overfitting in SVD, where priors are introduced and all parameters are integrated out using VI [28]. For parameter estimation of low-rank matrix decomposition, this work demonstrated that Bayesian approach is more robust against overfitting than EM and MAP. This work also showed that the effectiveness of VI is not just due to the priors introduced in probabilistic model, but also due to expectation rating. It is trained like as BPMF on rating matrix with extreme data sparsity, this is difference with our work.
- Harvey *et al.* presented a Bayesian latent variable model for rating prediction that models ratings over each user's latent interests and also each item's latent topics[40]. This work constructed a probabilistic topic model with assuming the user-interest and item-topic variables drawn from multinomial distributions. It is different with probabilistic factor model in which map users and items to a

joint latent space. This work used Gibbs sampling [55] which is a MCMC method to estimate its parameters and showed that it is competitive with the gradient descent SVD methods. In fact, the model of this work used the ability of Latent Dirichlet Allocation (LDA) [56] to extrude latent factors from sparse data. By incorporating user-dependent and item-dependent biases into the model, this work make an extension to enhance the performance of rating prediction. We concede that the ideas of performing rating prediction using Bayesian method and the inclusion of user and item biases is motivated by this work. However, the probabilistic model constructed and Bayesian inference method applied is different.

In summary, given observed rating, our BLFM conduct expectation rating with Bayesian method by introducing prior on latent factors of users and items, and have the characteristic of implementing simply, running flexibly, and easy of scalability. Furthermore, we illustrate the characteristics of our model, comparing with the state-of-the-art models. In the future work, we present a exploring research with incorporating additional information [57], [58] into BLFM to solve cold start problem [59], [60]. By combining prior constraint and posterior expectation estimation, BLFM is more robust against overfitting. In addition, how to choose the prior distribution for factor vectors and biases should be taken application and dataset into consideration.

VI. CONCLUSION AND FUTURE WORK

In this paper, given the observed rating, we propose probabilistic factor model with the method of Bayesian parameter estimation to implement LFM. First, we introduce Gaussian prior distribution for latent factors of each user and item. Second, we build likelihood function to establish the link between latent factors and observed rating. After constructing the probabilistic model, we apply VI which is a method of Bayesian approximation inference to make expectation rating prediction. We also extend our model by incorporating biases of users and items to enhance the performance. By extensive experiments on Movielens dataset, we demonstrate the effectiveness of our methods.

In our future work, the model can be extended to include more information, such as the time when each rating was made, the attributes of users and items to tackle cold start problem. Combining the additional information with explicit feedback and implicit feedback is a hot research of CF. The early works applied point estimate for parameters. By our experiment, we hold that the Bayesian method could be applied to solve the LFM model with additional information. The probabilistic model of matrix factorization with Bayesian treatment can be extended to other pattern recognition problems.

REFERENCES

- [1] C. A. Gomez-Urbe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, pp. 1–19, 2015.
- [2] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.
- [3] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. Orgun, "Sequential recommender systems: challenges, progress and prospects," *arXiv preprint arXiv:2001.04830*, 2019.
- [4] R. Glauber and A. Loula, "Collaborative filtering vs. content-based filtering: differences and similarities," *arXiv preprint arXiv:1912.08932*, 2019.
- [5] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer communications*, vol. 41, pp. 1–10, 2014.
- [6] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1–45, 2014.
- [7] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, no. 1, pp. 76–80, 2003.
- [8] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl *et al.*, "Item-based collaborative filtering recommendation algorithms." *Www*, vol. 1, pp. 285–295, 2001.
- [9] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 426–434.
- [10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.
- [11] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.
- [12] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 2008, pp. 263–272.
- [13] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 549–558.
- [14] T. Lake, S. A. Williamson, A. T. Hawk, C. C. Johnson, and B. P. Wing, "Large-scale collaborative filtering with product embeddings," *arXiv preprint arXiv:1901.04321*, 2019.
- [15] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," in *Icml*, vol. 98, 1998, pp. 46–54.
- [16] A. Mongia, N. Jhamb, E. Chouzenoux, and A. Majumdar, "Deep latent factor model for collaborative filtering," *Signal Processing*, vol. 169, p. 107366, 2020.
- [17] J. Sun, Y. Zhang, C. Ma, M. Coates, H. Guo, R. Tang, and X. He, "Multi-graph convolution collaborative filtering," *arXiv preprint arXiv:2001.00267*, 2020.
- [18] W. Huang and R. Y. Da Xu, "Gaussian process latent variable model factorization for context-aware recommender systems," *arXiv preprint arXiv:1912.09593*, 2019.
- [19] S. Verma, P. Patel, and A. Majumdar, "Collaborative filtering with label consistent restricted boltzmann machine," in *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2017, pp. 1–6.
- [20] Z. Sun, Q. Guo, J. Yang, H. Fang, G. Guo, J. Zhang, and R. Burke, "Research commentary on recommendations with side information: A survey and research directions," *Electronic Commerce Research and Applications*, vol. 37, p. 100879, 2019.
- [21] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [22] L. Chen, Y. Xu, F. Xie, M. Huang, and Z. Zheng, "Data poisoning attacks on neighborhood-based recommender systems," *Transactions on Emerging Telecommunications Technologies*, p. e3872, 2019.
- [23] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization," in *Proceedings of the 2010 SIAM international conference on data mining*. SIAM, 2010, pp. 211–222.
- [24] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257–1264.

- [25] B. M. Marlin, "Modeling user rating profiles for collaborative filtering," in *Advances in neural information processing systems*, 2004, pp. 627–634.
- [26] E. d. S. da Silva, T. Kuśmierczyk, M. Hartmann, and A. Klami, "Prior specification via prior predictive matching: Poisson matrix factorization and beyond," *arXiv preprint arXiv:1910.12263*, 2019.
- [27] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.
- [28] Y. J. Lim and Y. W. Teh, "Variational bayesian approach to movie rating prediction," in *Proceedings of KDD cup and workshop*, vol. 7, 2007, pp. 15–21.
- [29] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 880–887.
- [30] P. Yang, S. Varadharajan, L. A. Wilson, I. Smith, D. Don, J. A. Lockman III, V. Gundecha, and Q. Ta, "Parallelized training of restricted boltzmann machines using markov-chain monte carlo methods," *arXiv preprint arXiv:1910.05885*, 2019.
- [31] W. Gong, S. Tschitschek, R. Turner, S. Nowozin, and J. M. Hernández-Lobato, "Icebreaker: Element-wise active information acquisition with bayesian deep latent gaussian model," *arXiv preprint arXiv:1908.04537*, 2019.
- [32] G. Gautier, R. Bardenet, and M. Valko, "Zonotope hit-and-run for efficient sampling from projection dpps," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1223–1232.
- [33] Z. Meng, R. McCreadie, C. Macdonald, and I. Ounis, "Variational bayesian context-aware representation for grocery recommendation," *arXiv preprint arXiv:1909.07705*, 2019.
- [34] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [35] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.
- [36] S. M. Stigler *et al.*, "The epic story of maximum likelihood," *Statistical Science*, vol. 22, no. 4, pp. 598–620, 2007.
- [37] R. Bassett and J. Deride, "Maximum a posteriori estimators as a limit of bayes estimators," *Mathematical Programming*, vol. 174, no. 1-2, pp. 129–144, 2019.
- [38] M. J. Beal *et al.*, *Variational algorithms for approximate Bayesian inference*. university of London London, 2003.
- [39] F. Pérez-Cruz, "Kullback-leibler divergence estimation of continuous distributions," in *2008 IEEE international symposium on information theory*. IEEE, 2008, pp. 1666–1670.
- [40] M. Harvey, M. J. Carman, I. Ruthven, and F. Crestani, "Bayesian latent variable models for collaborative item rating prediction," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 699–708.
- [41] A. Gogna and A. Majumdar, "Balancing accuracy and diversity in recommendations using matrix completion framework," *Knowledge-Based Systems*, vol. 125, pp. 83–95, 2017.
- [42] Y. Lin, P. Ren, Z. Chen, Z. Ren, D. Yu, J. Ma, M. de Rijke, and X. Cheng, "From personalization to privatization: Meta matrix factorization for private rating predictions," *arXiv preprint arXiv:1910.10086*, 2019.
- [43] J. Wang, A. P. De Vries, and M. J. Reinders, "Unified relevance models for rating prediction in collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 16, 2008.
- [44] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 713–719.
- [45] J. Strahl, J. Peltonen, H. Mamitsuka, and S. Kaski, "Scalable probabilistic matrix factorization with graph-based priors," *arXiv preprint arXiv:1908.09393*, 2019.
- [46] S. Zhang, L. Liu, Z. Chen, and H. Zhong, "Probabilistic matrix factorization with personalized differential privacy," *Knowledge-Based Systems*, vol. 183, p. 104864, 2019.
- [47] R. P. Monti and A. Hyvärinen, "A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data," *arXiv preprint arXiv:1805.09567*, 2018.
- [48] T. J. Evans, R. Harper, and S. T. Flammia, "Scalable bayesian hamiltonian learning," *arXiv preprint arXiv:1912.07636*, 2019.
- [49] O. Kampman and C. Heaukulani, "Variational inference for neural network matrix factorization and its application to stochastic blockmodeling," *arXiv preprint arXiv:1905.04502*, 2019.
- [50] H. Ahn and T. Moon, "Iterative channel estimation for discrete denoising under channel uncertainty," *arXiv preprint arXiv:1902.08921*, 2019.
- [51] J. A. Gillenwater, A. Kulesza, E. Fox, and B. Taskar, "Expectation-maximization for learning determinantal point processes," in *Advances in Neural Information Processing Systems*, 2014, pp. 3149–3157.
- [52] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 720–727.
- [53] Y. Zhang and J. Koren, "Efficient bayesian hierarchical user modeling for recommendation system," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 47–54.
- [54] D. H. Stern, R. Herbrich, and T. Graepel, "Matchbox: large scale online bayesian recommendations," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 111–120.
- [55] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
- [56] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022.
- [57] V. W. Anelli, T. Di Noia, E. Di Sciascio, A. Ragone, and J. Trotta, "Local popularity and time in top-n recommendation," in *European Conference on Information Retrieval*. Springer, 2019, pp. 861–868.
- [58] Y. Koren, "Collaborative filtering with temporal dynamics," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 447–456.
- [59] L. Uyangoda, S. Ahangama, and T. Ranasinghe, "User profile feature-based approach to address the cold start problem in collaborative filtering for personalized movie recommendation," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. IEEE, 2018, pp. 24–28.
- [60] G. Guo, J. Zhang, and N. Yorke-Smith, "Trustsvd: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.