HIERARCHICAL VARIATIONAL BAYESIAN MATRIX CO-FACTORIZATION

Jiho Yoo[1] *and Seungjin Choi*[1,2]

[1] Department of Computer Science, POSTECH, Korea
[2] Division of IT Convergence Engineering, POSTECH, Korea
{zentasis, seungjin}@postech.ac.kr

ABSTRACT

Matrix co-factorization involves jointly decomposing several data matrices to approximate each data matrix as a product of two factor matrices, sharing some factor matrices in the factorization. We have recently developed variational Bayesian matrix co-factorization where factor matrices are inferred by computing variational posterior distributions in the case of Gaussian likelihood with Gaussian prior placed on factor matrices. Empirical Bayesian method was used, so hyperparameters are set to specific values determined by maximizing marginal likelihood. In this paper we present a *hierarchical Bayesian model* for matrix co-factorization in which we derive a variational inference algorithm to approximately compute posterior distributions over factor matrices as well as hyperparameters, placing Gaussian-Wishart prior on hyperparameters. Numerical experiments on MovieLens data demonstrate that the hierarchical variational Bayesian matrix co-factorization alleviates the over-fitting better than the empirical variational Bayesian matrix co-factorization, leading to the improved performance in terms of MAE and RMSE.

Index Terms— Bayesian matrix factorization, cold-start problems, collaborative prediction, matrix co-factorization variational inference

1. INTRODUCTION

Matrix factorization is a method for seeking a low-rank latent structure of data, approximating the data matrix as a product of two or more factor matrices. Matrix factorization is a popular tool for collaborative prediction, where unknown ratings are predicted by user and item factor matrices which are determined to approximate a user-item matrix as their product [2, 5–7, 9, 12]. Probabilistic matrix factorization was introduced in [9], in which a linear model with Gaussian observations was considered to learn user-specific and term-specific latent features, which became equivalent to the minimization of sum-of-squared errors with quadratic regularization terms. Bayesian approaches to matrix factorization are proposed based on the approximate inference such as the variational inference [5] or sampling [8], since the exact inference for the probabilistic model is intractable. Bayesian matrix factorization is preferred over other methods for collaborative filtering, since Bayesian approach alleviates over-fitting by integrating out all model parameters.

Collaborative prediction algorithms suffer from the cold-start problem, where the users or items do not have a sufficient number of ratings. To handle the cold-start problem, an efficient use of side information, such as item content information and user demographic information is crucial. Matrix co-factorization has been developed, as a promising approach to systematically exploiting side information. Matrix co-factorization jointly decomposes multiple data matrices to approximate each data matrix as a product of two factor matrices, sharing some factor matrices in the factorization. Various methods and applications include supervised latent semantic indexing [17], content+link for classification [18], collective factorization [10], group nonnegative matrix factorization [3], nonnegative matrix co-tri-factorization [14], semi-supervised nonnegative matrix factorization [4], and co-factorization on compressed sensing [16], drum source separation [1].

Bayesian matrix co-factorization (BMCF) is often preferred over other methods for collaborative filtering, since Bayesian approach alleviates over-fitting by integrating out all model parameters. BMCF infers factor matrices in the decomposition, by computing posterior distributions approximately in the case of Gaussian likelihood with Gaussian prior placed on factor matrices. A sampling-based BMCF was proposed in [11], where the posterior computation requires storing multiple number of samples which is not appropriate for the large-scale collaborative prediction problems. Variational approximation for BMCF was developed in [13, 15], where empirical variational Bayesian method was applied, i.e., hyperparameters are set to specific values determined by maximizing marginal likelihood.

In this paper we present *hierarchical variational Bayesian matrix co-factorization* (HVBMCF), in which we place Gaussian-Wishart prior on hyperparameters (which are treated as random variables as well) and develop variational inference algorithm in the case of Gaussian likelihood with Gaussian prior on factor matrices. Numerical experiments on MovieLens Data demonstrate that our proposed method, HVBMCF improves the generalization performance in the task of collaborative prediction, compared to the *empirical variational Bayesian matrix co-factorization* (EVBMCF) [15].

2. RELATED WORK

This section briefly reviews our previous work [15] on EVBMCF which is the basis of our proposed method, HVBMCF. Suppose that we are given a set of dyadic data matrices, $\mathcal{X} = \{\boldsymbol{X}^{(a,b)} \in \mathbb{R}^{N_a \times N_b}\}$ for $(a,b) \in \mathcal{R}$, where $\mathcal{R}$ denotes a set of relations between two entities. For example, in the case where (a,b) is the user-item relation, the (i_a, i_b)-entry, denoted by $x_{i_a i_b}^{(a,b)}$, represents the rating of item i_b by user i_a.

EVBMCF assumes a linear Gaussian model, in which $x_{i_a i_b}^{(a,b)}$ is generated by an inner product of two vectors, $\boldsymbol{u}_{i_a}^{(a)}$ and $\boldsymbol{u}_{i_b}^{(b)}$, each of which corresponds to the column vector of factor matrices $\boldsymbol{U}^{(a)} \in \mathbb{R}^{d \times N_a}$ and $\boldsymbol{U}^{(b)} \in \mathbb{R}^{d \times N_b}$ involving entities a and b, respectively:

$$x_{i_a i_b}^{(a,b)} = \boldsymbol{u}_{i_a}^{(a)\top} \boldsymbol{u}_{i_b}^{(b)} + \epsilon_{i_a i_b}^{(a,b)},$$

for $\forall (a,b) \in \mathcal{R}$ and $(i_a, i_b) \in \mathcal{O}^{(a,b)}$, where $\mathcal{O}^{(a,b)}$ is a set of observed entries, and $\epsilon_{i_a i_b}^{(a,b)}$ is Gaussian noise with zero mean

and precision $\rho^{(a,b)}$, which reflects uncertainty in the model, i.e., $\epsilon_{i_a i_b}^{(a,b)} \sim \mathcal{N}(\epsilon_{i_a i_b}^{(a,b)} \,|\, 0, (\rho^{(a,b)})^{-1})$. For instance, in the case of $\mathcal{R} = \{(a,b),(a,c)\}$, where (a,b) is user-item relation and (a,c) is user-demographic information relation, the co-factorization of two matrices $\boldsymbol{X}^{(a,b)}$ and $\boldsymbol{X}^{(a,c)}$ shares the factor matrix $\boldsymbol{U}^{(a)}$, i.e., $\boldsymbol{X}^{(a,b)} \approx \boldsymbol{U}^{(a)\top}\boldsymbol{U}^{(b)}$ and $\boldsymbol{X}^{(a,c)} \approx \boldsymbol{U}^{(a)\top}\boldsymbol{U}^{(c)}$.

Gaussian prior distribution is assumed for factor matrices $\mathcal{U} = \{\boldsymbol{U}^{(a)} \,|\, a \in \mathcal{E}\}$ where $\mathcal{E}$ is a set of entities:

$$p\left(\boldsymbol{U}^{(a)}\right) = \prod_{i_a=1}^{N_a} \mathcal{N}\left(\boldsymbol{u}_{i_a}^{(a)} \middle| \boldsymbol{0}, \left(\boldsymbol{\Lambda}^{(a)}\right)^{-1}\right).$$

where $\boldsymbol{\Lambda}^{(a)}$ is the precision matrix which is the inverse of the covariance matrix. Then the likelihood is given by

$$\begin{aligned} p(\mathcal{X}|\mathcal{U}) &= \prod_{(a,b)\in\mathcal{R}} p\left(\boldsymbol{X}^{(a,b)} \middle| \boldsymbol{U}^{(a)}, \boldsymbol{U}^{(b)}\right) \\ &= \prod_{(a,b)\in\mathcal{R}} \prod_{(i_a,i_b)\in\mathcal{O}} \mathcal{N}\left(x_{i_a i_b}^{(a,b)} \middle| \boldsymbol{u}_{i_a}^{(a)\top}\boldsymbol{u}_{i_b}^{(b)}, \left(\rho^{(a,b)}\right)^{-1}\right). \end{aligned}$$

The graphical representation is illustrated in Fig. 1.

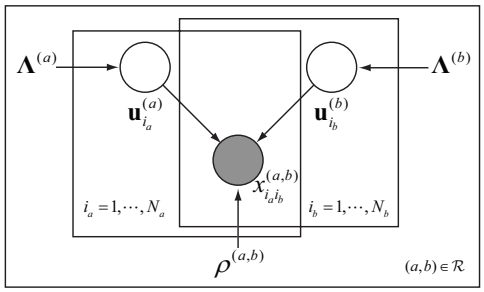

Fig. 1. Graphical representation of the probabilistic model for EVBMCF.

The variational Bayesian inference considers a lower-bound on the log marginal likelihood

$$\begin{aligned} \log p(\mathcal{X}) &= \log \int p(\mathcal{X},\mathcal{U}) d\mathcal{U} \\ &\geq \int q(\mathcal{U}) \log \frac{p(\mathcal{X},\mathcal{U})}{q(\mathcal{U})} d\mathcal{U} \equiv \mathcal{F}(q), \end{aligned}$$

where the Jensen's inequality was used and $\mathcal{F}(q)$ denotes the *variational lower-bound* to be maximized. Assume that the variational distribution $q(\mathcal{U})$ is factorized:

$$q(\mathcal{U}) = \prod_{a\in\mathcal{E}} q\left(\boldsymbol{U}^{(a)}\right).$$

Variational posterior distributions, $q^*(\cdot)$, are determined by maximizing the variational lower bound $\mathcal{F}(q)$, leading to

$$\log q^*\left(\boldsymbol{U}^{(a)}\right) \propto \mathbb{E}_{\mathcal{U}\setminus U^{(a)}}\left\{\log p(\mathcal{X},\mathcal{U})\right\},$$

where the expectation is taken with respect to the variational distributions over all variables excluding $\boldsymbol{U}^{(a)}$. In EVBMCF, hyperparameters, $\{\boldsymbol{\Lambda}^{(a)} \,|\, a \in \mathcal{E}\}$, and noise variance $\{\rho^{(a,b)} \,|\, (a,b) \in \mathcal{R}\}$ are set to specific values determined by maximizing the variational lower bound $\mathcal{F}(q)$. EVBMCF works fairly well for cold-start problems, however, the prediction accuracy often degrades as iterations proceed, due to the over-fitting caused by the point estimates of hyperparameters. Thus, the early-stopping, in general, is required to obtain accurate predictions. See [15] for more details on EVBMCF.

3. HIERARCHICAL VARIATIONAL BAYESIAN MATRIX CO-FACTORIZATION

In this section we present the main contribution of our paper, in which treat hyperparameters as random variables, placing Gaussian-Wishart prior on hyperparameters, in order to construct a hierarchical Bayesian model for matrix co-factorization. We develop a variational inference algorithm where we iteratively compute variational posterior distributions over factor matrices in the co-factorization.

As in EVBMCF described in Section 2, we assume the same linear Gaussian model with Gaussian prior placed on factor matrices. In contrast to EVBMCF, we place Gaussian-Wishart prior distribution on hyperparameters $\boldsymbol{\mu}^{(a)}$ and $\boldsymbol{\Lambda}^{(a)}$:

$$p\left(\boldsymbol{\mu}^{(a)}, \boldsymbol{\Lambda}^{(a)}\right) = \mathcal{N}\left(\boldsymbol{\mu}^{(a)} \middle| \boldsymbol{\mu}_0, \left(\gamma_0 \boldsymbol{\Lambda}^{(a)}\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}^{(a)} \middle| \boldsymbol{\Omega}_0, \nu_0\right),$$

where $\mathcal{W}(\boldsymbol{\Lambda}^{(a)}|\boldsymbol{\Omega}_0,\nu_0)$ is a Wishart distribution over $\boldsymbol{\Lambda}^{(a)}$. We assume Gamma distribution over the noise precision variables:

$$p\left(\rho^{(a,b)}\right) = \mathcal{G}\left(\rho^{(a,b)} \middle| \alpha_0, \beta_0\right),$$

where α_0 and β_0 are shape and inverse scale parameters. The graphical representation is illustrated in Fig. 2.

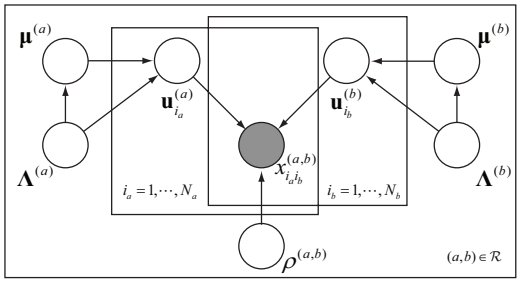

Fig. 2. Graphical representation of the probabilistic model for HVBMCF.

We define $\Theta = \{\rho^{(a,b)}, \boldsymbol{\mu}^{(a)}, \boldsymbol{\Lambda}^{(a)} \,|\, (a,b) \in \mathcal{R}, a \in \mathcal{E}\}$. Then, the marginal likelihood is written as

$$p(\mathcal{X}) = \int\int p(\mathcal{X},\mathcal{U},\Theta) d\mathcal{U} d\Theta.$$

The variational lower-bound $\mathcal{F}(q)$ is given by

$$\log p(\mathcal{X}) \geq \int\int q(\mathcal{U},\Theta) \log \frac{p(\mathcal{X},\mathcal{U},\Theta)}{q(\mathcal{U},\Theta)} d\mathcal{U} d\Theta \equiv \mathcal{F}(q).$$

We assume that the variational distribution $q(\mathcal{U},\Theta)$ is factorized:

$$\begin{aligned} q(\mathcal{U},\Theta) = &\prod_{a\in\mathcal{E}} q\left(\boldsymbol{U}^{(a)}\right) \prod_{(a,b)\in\mathcal{R}} q\left(\rho^{(a,b)}\right) \\ &\prod_{a\in\mathcal{E}} q\left(\boldsymbol{\mu}^{(a)} \middle| \boldsymbol{\Lambda}^{(a)}\right) q\left(\boldsymbol{\Lambda}^{(a)}\right). \end{aligned}$$

Table 1. Variational posterior distributions and corresponding parameter updates are summarized for HVBMCF.

Variational posterior	Parameter updates	Sufficient statistics
$q^*\left(\boldsymbol{U}^{(a)}\right)$ $= \prod_{i_a} \mathcal{N}\left(\boldsymbol{u}_{i_a}^{(a)} \middle\vert \boldsymbol{m}_{i_a}^{(a)}, \left(\boldsymbol{L}_{i_a}^{(a)}\right)^{-1}\right)$	$\boldsymbol{m}_{i_a}^{(a)} = \left(\boldsymbol{L}_{i_a}^{(a)}\right)^{-1}\left\langle\boldsymbol{\Lambda}^{(a)}\right\rangle\left\langle\boldsymbol{\mu}^{(a)}\right\rangle$ $+\left(\boldsymbol{L}_{i_a}^{(a)}\right)^{-1}\sum_{b\vert(a,b)\in\mathcal{R}}\left\langle\rho^{(a,b)}\right\rangle\sum_{i_b\vert(i_a,i_b)\in\mathcal{O}} x_{i_a i_b}^{(a,b)}\left\langle\boldsymbol{u}_{i_b}^{(b)}\right\rangle$ $\boldsymbol{L}_{i_a}^{(a)} = \left\langle\boldsymbol{\Lambda}^{(a)}\right\rangle + \sum_{b\vert(a,b)\in\mathcal{R}}\left\langle\rho^{(a,b)}\right\rangle\sum_{(i_a,i_b)\in\mathcal{O}}\left\langle\boldsymbol{u}_{i_b}^{(b)}\boldsymbol{u}_{i_b}^{(b)\top}\right\rangle$	$\left\langle\boldsymbol{u}_{i_a}^{(a)}\right\rangle = \boldsymbol{m}_{i_a}^{(a)}$ $\left\langle\boldsymbol{u}_{i_a}^{(a)}\boldsymbol{u}_{i_a}^{(a)\top}\right\rangle$ $= \left(\boldsymbol{L}_{i_a}^{(a)}\right)^{-1}$ $+\boldsymbol{m}_{i_a}^{(a)}\boldsymbol{m}_{i_a}^{(a)\top}$
$q^*\left(\rho^{(a,b)}\right)$ $= \mathcal{G}\left(\rho^{(a,b)} \middle\vert \alpha^{(a,b)}, \beta^{(a,b)}\right)$	$\alpha^{(a,b)} = \alpha_0 + \vert\mathcal{O}^{(a,b)}\vert/2$ $\beta^{(a,b)} = \beta_0 + \frac{1}{2}\sum_{(i_a,i_b)\in\mathcal{O}^{(a,b)}}\left(\left(x_{i_a i_b}^{(a,b)}\right)^2 - 2x_{i_a i_b}^{(a,b)}\left\langle\boldsymbol{u}_{i_a}^{(a)}\right\rangle^\top\left\langle\boldsymbol{u}_{i_b}^{(b)}\right\rangle\right)$ $+\frac{1}{2}\sum_{(i_a,i_b)\in\mathcal{O}^{(a,b)}}\mathrm{tr}\left(\left\langle\boldsymbol{u}_{i_a}^{(a)}\boldsymbol{u}_{i_a}^{(a)\top}\right\rangle\left\langle\boldsymbol{u}_{i_b}^{(b)}\boldsymbol{u}_{i_b}^{(b)\top}\right\rangle\right)$	$\left\langle\rho^{(a,b)}\right\rangle$ $= \alpha^{(a,b)}/\beta^{(a,b)}$
$q^*\left(\boldsymbol{\mu}^{(a)} \middle\vert \boldsymbol{\Lambda}^{(a)}\right)$ $= \mathcal{N}\left(\boldsymbol{\mu}^{(a)} \middle\vert \boldsymbol{\phi}^{(a)}, \left(\gamma^{(a)}\boldsymbol{\Lambda}^{(a)}\right)^{-1}\right)$	$\boldsymbol{\phi}^{(a)} = \frac{\gamma_0}{\gamma_0+N_a}\boldsymbol{\mu}_0 + \frac{1}{\gamma_0+N_a}\sum_{i_a}\left\langle\boldsymbol{u}_{i_a}^{(a)}\right\rangle$ $\gamma^{(a)} = \gamma_0 + N_a$	$\left\langle\boldsymbol{\mu}^{(a)}\right\rangle = \boldsymbol{\phi}^{(a)}$
$q^*(\boldsymbol{\Lambda}^{(a)})$ $= \mathcal{W}(\boldsymbol{\Lambda}^{(a)}\vert\boldsymbol{\Omega}^{(a)}, \nu^{(a)})$	$\left(\boldsymbol{\Omega}^{(a)}\right)^{-1} = \boldsymbol{\Omega}_0^{-1} + \sum_{i_a}\left\langle\boldsymbol{u}_{i_a}^{(a)}\boldsymbol{u}_{i_a}^{(a)\top}\right\rangle$ $+\frac{\gamma_0 N_a}{\gamma_0+N_a}\left(\boldsymbol{\mu}_0-\boldsymbol{\omega}^{(a)}\right)\left(\boldsymbol{\mu}_0-\boldsymbol{\omega}^{(a)}\right)^\top - N_a\boldsymbol{\omega}^{(a)}\boldsymbol{\omega}^{(a)\top}$ $\nu^{(a)} = \nu_0 + N_a$ $\boldsymbol{\omega}^{(a)} = \frac{1}{N_a}\sum_{i_a}\left\langle\boldsymbol{u}_{i_a}^{(a)}\right\rangle$	$\left\langle\boldsymbol{\Lambda}^{(a)}\right\rangle$ $= \nu^{(a)}\boldsymbol{\Omega}^{(a)}$

Variational posterior distributions over factor matrices and hyperparameters are determined by maximizing the variational lower bound $\mathcal{F}(q)$, leading to

$$\begin{aligned}
\log q^*\left(\boldsymbol{U}^{(a)}\right) &\propto \mathbb{E}_{\mathcal{U}\setminus U^{(a)},\Theta}\Big\{\log p(\mathcal{X},\mathcal{U},\Theta)\Big\},\\
\log q^*\left(\rho^{(a,b)}\right) &\propto \mathbb{E}_{\mathcal{U},\Theta\setminus\rho^{(a,b)}}\Big\{\log p(\mathcal{X},\mathcal{U},\Theta)\Big\},\\
\log q^*\left(\boldsymbol{\mu}^{(a)},\boldsymbol{\Lambda}^{(a)}\right) &\propto \mathbb{E}_{\mathcal{U},\Theta\setminus\mu^{(a)},\Lambda^{(a)}}\Big\{\log p(\mathcal{X},\mathcal{U},\Theta)\Big\},
\end{aligned}$$

where their functional forms and corresponding parameter updates are summarized in Table 1.

4. NUMERICAL EXPERIMENTS

We applied the proposed HVBMCF to the collaborative prediction in the cold-start situations, and compared the performance with EVBMCF [15]. In addition, we compared the performance of the empirical variational Bayesian matrix factorization (EVBMF), which is a special case of EVBMCF exploiting only the user-item rating matrix. The MovieLens data, which consists of the 5-star ratings of 943 users for the 1682 movies was used. In the EVBMCF and HVBMCF, additional user information (age, gender, and occupation) and movie information (genre) were used in the matrix co-factorization. Additional information is coded with the binary values, for example, movie genre data is coded by a vector of length 18, where each element indicates one of the 18 movie categories, and the value 1 represents the movie belongs to the corresponding genre.

Since the user cold-start problem occurs in the situation that the test users do not provide sufficient number of ratings, we randomly selected 200 test users and took out most of their ratings, to remain s ratings for each user. We generated the datasets for the five different values of s, which were 0, 5, 10, 15, and 20. We used mean absolute error (MAE) and root mean squared error (RMSE) as the performance measures, which are computed as

$$\begin{aligned}
\text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|r_i - \widehat{r}_i|,\\
\text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \widehat{r}_i)^2},
\end{aligned}$$

where N is the total number of test data points, $\widehat{r}_i$ and r_i are the predicted rating and the true rating of the i-th test data, respectively. For each value of s, we ran the algorithms 100 times with different set of test users and initial values of sufficient statistics. Number of latent factors d is set to 20, and hyperparameters are set to $\alpha_0 = 1$, $\beta_0 = 1$, $\nu_0 = 1$, and $W_0 = \boldsymbol{I}$. The averaged MAE and RMSE were summarized in Table 2(a). We also simulated the item and user cold-start cases, where we also eliminated all the ratings for the 100 randomly selected movies from each dataset generated for the user cold-start cases (Table 2(b)). Both EVBMCF and HVBMCF showed better performance than EVBMF for all cases, showing the benefit of the matrix co-factorization over the single matrix factorization in the cold-start situations. HVBMCF showed comparable performance to the EVBMCF in user cold-start cases, and showed better performance in the item and user cold-start cases. HVBMCF worked better than EVBMCF in the situation with less available information, which shows better generalization performance of HVBMCF.

Moreover, HVBMCF showed much stable evolvement pattern of the prediction accuracy than EVBMCF, while the prediction accuracy of EVBMCF degrades as iterations proceed because of the over-fitting. Fig. 3 illustrates an exemplary behavior in terms of RMSE, for the case of item and user cold start with $s = 0$.

5. CONCLUSIONS

In this paper, we have presented hierarchical variational Bayesian matrix co-factorization (HVBMCF) which treats hyperparameters as random variables, and places Gaussian-Wishart prior on them.

Table 2. MAE and RMSE results for different number of available ratings for each test user. (a) Simulation of user cold-start case. (b) Simulation of user and item cold-start case.

(a)	EVBMF		EVBMCF		HVBMCF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0	2.5403	2.7767	0.8238	**1.0140**	**0.8182**	**1.0140**
5	0.8281	1.0618	0.7895	**0.9941**	**0.7856**	0.9983
10	0.8032	1.0205	**0.7446**	**0.9424**	0.7485	0.9499
15	0.7474	0.9558	0.7426	0.9314	**0.7315**	**0.9288**
20	0.7421	0.9496	0.7348	**0.9254**	**0.7318**	0.9328
(b)	EVBMF		EVBMCF		HVBMCF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0	2.5098	2.7584	0.8843	1.0857	**0.8399**	**1.0437**
5	0.9333	1.2412	0.8332	1.0550	**0.7930**	**1.0046**
10	0.8956	1.1863	0.7778	0.9857	**0.7686**	**0.9743**
15	0.8991	1.1948	0.7716	0.9789	**0.7556**	**0.9589**
20	0.8618	1.1535	0.7527	0.9555	**0.7394**	**0.9418**

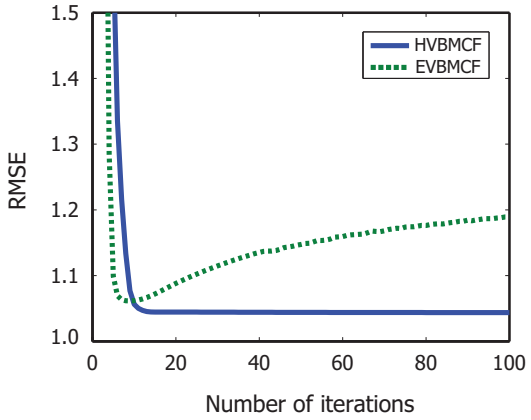

Fig. 3. Performance comparison between EVBMCF and HVBMCF is shown in terms of RMSE.

A variational inference algorithm, which iteratively computes variational posterior distributions over factor matrices and hyperparameters is presented for the hierarchical model. Numerical experiments showed that the HVBMCF alleviates the over-fitting better than EVBMCF, leading to the improved performance.

Acknowledgments: This work was supported by MEST Converging Research Center Program (No. 2011K000673), NIPA ITRC Program (NIPA-2012-C1090-1231-0009), Microsoft Research Asia, and NRF World Class University Program (R31-10100).

6. REFERENCES

[1] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[3] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida, 2009.

[4] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010.

[5] Y. J. Lim and Y. W. Teh, "Variational Bayesian approach to movie rating prediction," in *Proceedings of KDD Cup and Workshop*, San Jose, CA, 2007.

[6] T. Raiko, A. Ilin, and J. Karhunen, "Principal component analysis for large scale problems with lots of missing values," in *Proceedings of the European Conference on Machine Learning (ECML)*, Warsaw, Poland, 2007, pp. 691–698.

[7] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005.

[8] R. Salakhutdinov and A. Mnih, "Bayesian probablistic matrix factorization using MCMC," in *Proceedings of the International Conference on Machine Learning (ICML)*, Helsinki, Finland, 2008.

[9] ——, "Probablistic matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 20. MIT Press, 2008.

[10] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Las Vegas, Nevada, 2008.

[11] ——, "A Bayesian matrix factorization model for relational data," in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, CA, 2010.

[12] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Scalable collaborative filtering approaches for large recommender systems," *Journal of Machine Learning Research*, vol. 10, pp. 623–656, 2009.

[13] S. Williamson and Z. Ghahramani, "Probabilistic models for data combination in recommender systems," in *NIPS-2008 Workshop on Learning from Multiple Sources*, Whistler, Canada, 2010.

[14] J. Yoo and S. Choi, "Weighted nonnegative matrix co-tri-factorization for collaborative prediction," in *Proceedings of the First Asian Conference on Machine Learning (ACML)*, Nanjing, China, 2009.

[15] ——, "Bayesian matrix co-factorization: Variational algorithm and Cramér-Rao bound," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Athens, Greece, 2011.

[16] ——, "Matrix co-factorization on compressed sensing," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain, 2011.

[17] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[18] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Amsterdam, The Netherlands, 2007.