Hierarchical Bayesian Matrix Factorization with Side Information

Sunho Park[1], **Yong-Deok Kim**[1], **Seungjin Choi**[1,2]
[1] Department of Computer Science and Engineering
[2] Division of IT Convergence Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
{titan,karma13,seungjin}@postech.ac.kr

Abstract

Bayesian treatment of matrix factorization has been successfully applied to the problem of collaborative prediction, where unknown ratings are determined by the predictive distribution, inferring posterior distributions over user and item factor matrices that are used to approximate the user-item matrix as their product. In practice, however, Bayesian matrix factorization suffers from cold-start problems, where inferences are required for users or items about which a sufficient number of ratings are not gathered. In this paper we present a method for Bayesian matrix factorization with side information, to handle cold-start problems. To this end, we place Gaussian-Wishart priors on mean vectors and precision matrices of Gaussian user and item factor matrices, such that mean of each prior distribution is regressed on corresponding side information. We develop variational inference algorithms to approximately compute posterior distributions over user and item factor matrices. In addition, we provide Bayesian Cramér-Rao Bound for our model, showing that the hierarchical Bayesian matrix factorization with side information improves the reconstruction over the standard Bayesian matrix factorization where the side information is not used. Experiments on MovieLens data demonstrate the useful behavior of our model in the case of cold-start problems.

1 Introduction

Matrix factorization refers to a method for uncovering a low-rank latent structure of data, approximating the data matrix as a product of two factor matrices. Matrix factorization is popular 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 [Rennie and Srebro, 2005; Salakhutdinov and Mnih, 2008b; Lim and Teh, 2007; Raiko *et al.*, 2007; Takács *et al.*, 2009; Koren *et al.*, 2009; Singh and Gordon, 2010].

Suppose that $\boldsymbol{X} \in \mathbb{R}^{I\times J}$ is a user-item rating matrix, the (i,j)-entry of which, X_{ij}, represents the rating of user i on item j. Matrix factorization determines factor matrices $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_I] \in \mathbb{R}^{K\times I}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J] \in \mathbb{R}^{K\times J}$ (where K denotes the dimension of latent space), to approximate the rating matrix $\boldsymbol{X}$ by $\boldsymbol{U}^\top \boldsymbol{V}$:

$$\boldsymbol{X} \approx \boldsymbol{U}^\top \boldsymbol{V}. \tag{1}$$

A popular approach is to minimize the regularized squared error loss defined as

$$\sum_{(i,j)\in\Omega} \left[(X_{ij} - \boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 + \lambda(\|\boldsymbol{u}_i\|^2 + \|\boldsymbol{v}_j\|^2)\right], \tag{2}$$

where Ω is a set of indices of observed entries in $\boldsymbol{X}$, and λ is the regularization parameter. Although this works even on large-scale datasets [Takács *et al.*, 2009; Koren *et al.*, 2009], it is prone to overfitting on the training data and require careful tuning of regularization parameter and the number of optimization steps.

Bayesian treatment of matrix factorization successfully alleviates the overfitting problem by integrating out all model parameter, thus allowing for complex models to be learned without requiring much parameter tuning [Lim and Teh, 2007; Raiko *et al.*, 2007; Salakhutdinov and Mnih, 2008a; Yoo and Choi, 2011; 2012; Kim and Choi, 2013]. In practice, however, Bayesian matrix factorization still suffers from *cold-start problem*, where inferences are required for users or items about which a sufficient number of ratings are not gathered. The cold-start problem commonly occurs in practical recommendation system based on collaborative prediction because new users or items are continuously added into the system. Moreover users usually do not rate unpopular items, thus these items will be fixed in long-tail part for a long time but for effective recommendation. To resolve the cold-start problem, it is crucial to use side information, such as user demographic information and item content information.

In this paper we present a method for Bayesian matrix factorization with side information, to handle cold-start problems. To this end, we place Gaussian-Wishart priors on mean vectors and precision matrices of Gaussian user and item factor matrices, such that mean of each prior distribution is regressed on corresponding side information. We develop variational inference algorithms to approximately compute posterior distributions over user and item factor matrices. In addition, we provide Bayesian Cramér-Rao Bound for our model, showing that the hierarchical Bayesian matrix factorization with side information improves the reconstruction over the

standard Bayesian matrix factorization where the side information is not used. Experiments on MovieLens data demonstrate the useful behavior of our model in the case of cold-start problems.

2 Related Work

We briefly review several matrix factorization models which incorporates side information. Suppose that side information is given in the form of feature matrices $\boldsymbol{F} = [\boldsymbol{f}_1, ..., \boldsymbol{f}_I] \in \mathbb{R}^{D_u \times I}$ and $\boldsymbol{G} = [\boldsymbol{g}_1, ..., \boldsymbol{g}_J] \in \mathbb{R}^{D_v \times J}$, where $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$ are feature vector related with user i and item j respectively.

Matrix co-factorization is one one of efficient ways to exploit the side information. Matrix co-factorization jointly decomposes multiple data matrices, where each deposition is coupled by sharing some factor matrices [Singh and Gordon, 2008]. For example, rating matrix $\boldsymbol{X}$, user side information matrix $\boldsymbol{F}$ and item side information matrix $\boldsymbol{G}$ are jointly decomposed as

$$\boldsymbol{F} = \boldsymbol{A}^\top \boldsymbol{U} + \boldsymbol{E}_F,\ \boldsymbol{X} = \boldsymbol{U}^\top \boldsymbol{V} + \boldsymbol{E}_X,\ \boldsymbol{G} = \boldsymbol{B}^\top \boldsymbol{V} + \boldsymbol{E}_G,$$

where factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are shared in the first two and last two decompositions, respectively and $\boldsymbol{E}_F, \boldsymbol{E}_X, \boldsymbol{E}_G$ represents Gaussian noise which reflect uncertainties. Recently Bayesian treatment of matrix co-factorization was developed, where the inference was performed using a sampling method [Singh and Gordon, 2010] or the variational method [Yoo and Choi, 2011; 2012].

In contrast to matrix co-factorization model, *Regression-based Latent Factor Model* (RLFM) [Agarwal and Chen, 2009] assumes that the user and item factor matrices are generated from the side information via linear regression,

$$\boldsymbol{U} = \boldsymbol{A}^\top \boldsymbol{F} + \boldsymbol{E}_U, \quad \boldsymbol{V} = \boldsymbol{B}^\top \boldsymbol{G} + \boldsymbol{E}_V, \tag{3}$$

and then rating matrix $\boldsymbol{X}$ is generated by (1). More precisely, RLFM place isotropic Gaussian priors on user and item factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ where mean of each prior is regressed on corresponding side information (see Figure 1-(b)),

$$p(\boldsymbol{U}|\boldsymbol{A}, \lambda_u) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{u}_i \,|\, \boldsymbol{A}^\top \boldsymbol{f}_i, \lambda_u^{-1}), \tag{4}$$

$$p(\boldsymbol{V}|\boldsymbol{B}, \lambda_v) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{v}_j \,|\, \boldsymbol{B}^\top \boldsymbol{g}_j, \lambda_v^{-1}). \tag{5}$$

Baseysian matrix factorization with side information (BMFSI), proposed in [Proteous *et al.*, 2010], also utilizes side information via regression but it is differentiated from RLFM. In BMFSI model, the regression coefficient and side information are augmented in factor matrices such that the rating matrix $\boldsymbol{X}$ is estimated by the sum of the collaborative prediction part and regression part against side information:

$$\boldsymbol{X} = \left[\boldsymbol{U}^\top \boldsymbol{A}^\top \boldsymbol{F}^\top\right] \left[\boldsymbol{V}^\top \boldsymbol{G}^\top \boldsymbol{B}^\top\right]^\top + \boldsymbol{E}_X \tag{6}$$

$$= \boldsymbol{U}^\top \boldsymbol{V} + \boldsymbol{A}^\top \boldsymbol{G} + \boldsymbol{F}^\top \boldsymbol{B} + \boldsymbol{E}_X. \tag{7}$$

Moreover BMFSI utilizes user-item specific features such as N ratings for item j, given by N nearest-neighborhoods of user i, which is not considered in matrix co-factorization and RLFM.

3 Model and Inference

In this section we present our hierarchical Bayesian modeling for matrix factorization which utilizes side information and explain how to our model brides Bayesian matrix factorization (BMF) [Salakhutdinov and Mnih, 2008a] and RLFM. We develop a variational inference algorithm to approximately compute the posterior distributions over factor matrices and provide rating prediction procedures for hold-out and fold-in cases also in the variational inference framework.

3.1 Hierarchical Bayesian Modeling

The graphical model representing our hierarchical Bayesian matrix factorization with side information (HBMFSI) is shown in Figure 1-(c). At first, the observed data X_{ij} is assumed to be generated by a sum of inner product of latent factors, bias terms regressed on the feature vectors, and noise:

$$X_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j + \boldsymbol{a}_0^\top \boldsymbol{f}_i + \boldsymbol{b}_0^\top \boldsymbol{g}_j + \mathcal{E}_{ij}, \tag{8}$$

for $(i, j) \in \Omega$, where Ω is a set of indices of observe entries in $\boldsymbol{X}$. The uncertainty in the model is absorbed by the noise $\mathcal{E}_{ij}$ which is assumed to be Gaussian, i.e., $\mathcal{E}_{ij} \sim \mathcal{N}(\mathcal{E}_{ij} \,|\, 0, \rho^{-1})$ where ρ is the precision (the inverse of variance). Thus, the likelihood is given by

$$p(\boldsymbol{X} \,|\, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{a}_0, \boldsymbol{v}_0, \rho) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \,|\, \boldsymbol{u}_i^\top \boldsymbol{v}_j + \boldsymbol{a}_0^\top \boldsymbol{f}_i + \boldsymbol{b}_0^\top \boldsymbol{g}_j, \rho^{-1}). \tag{9}$$

The prior distributions over factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are assumed to be Gaussian:

$$p(\boldsymbol{U}|\Theta_u) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{u}_i \,|\, \boldsymbol{m}_i, \boldsymbol{\Lambda}_u^{-1}), \tag{10}$$

$$p(\boldsymbol{V}|\Theta_v) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{v}_j \,|\, \boldsymbol{n}_j, \boldsymbol{\Lambda}_v^{-1}), \tag{11}$$

where $\Theta_u = \{\{\boldsymbol{m}_i\}_{i=1}^I, \boldsymbol{\Lambda}_u\}$ and $\Theta_v = \{\{\boldsymbol{n}_j\}_{j=1}^J, \boldsymbol{\Lambda}_v\}$. We further place Gaussian-Wishart priors on the $\Theta_u = \{\{\boldsymbol{m}_i\}_{i=1}^I, \boldsymbol{\Lambda}_u\}$ and $\Theta_v = \{\{\boldsymbol{n}_j\}_{j=1}^J, \boldsymbol{\Lambda}_v\}$:

$$p(\boldsymbol{\Theta}_u|\, \boldsymbol{A}, \boldsymbol{m}_0, \gamma_0, \nu_0, \boldsymbol{W}_0) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{m}_i|\, \boldsymbol{m}_0 + \boldsymbol{A}^\top \boldsymbol{f}_i, (\gamma_0 \boldsymbol{\Lambda}_u)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_u|\boldsymbol{W}_0, \mu_0), \tag{12}$$

$$p(\boldsymbol{\Theta}_v|\, \boldsymbol{B}, \boldsymbol{n}_0, \gamma_0, \nu_0, \boldsymbol{W}_0) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{n}_i|\, \boldsymbol{n}_0 + \boldsymbol{B}^\top \boldsymbol{g}_j, (\gamma_0 \boldsymbol{\Lambda}_v)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_v|\boldsymbol{W}_0, \mu_0), \tag{13}$$

and finally place Gaussian priors on regression coefficient matrices $\boldsymbol{A} \in \mathbb{R}^{D_u \times K}$, $\boldsymbol{B} \in \mathbb{R}^{D_v \times K}$ and vectors $\boldsymbol{a}_0 \in \mathbb{R}^{D^u}$,

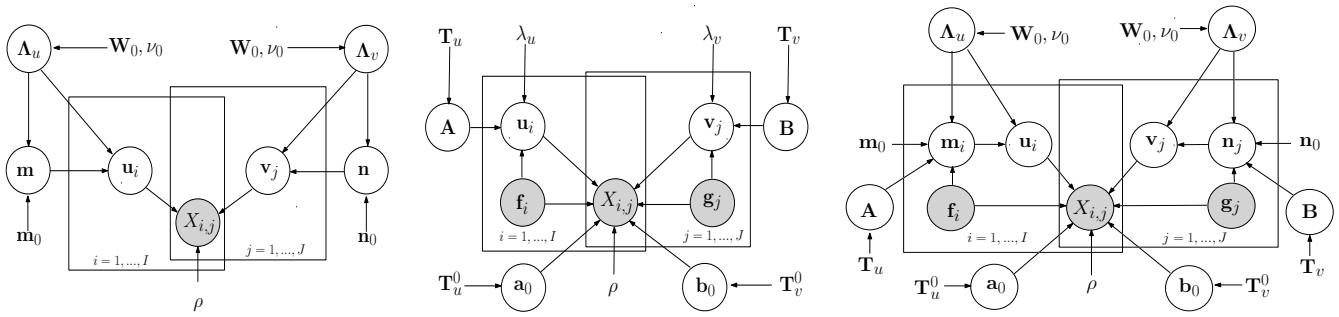

Figure 1: The probabilistic graphical models of (a) Standard Bayesian matrix factorization without side information, (b) Regression-based Latent Factor Model (RLFM) and (c) our hierarchical Bayesian matrix factorization with side information.

$\boldsymbol{b}_0 \in \mathbb{R}^{D_v}$:

$$p(\boldsymbol{A}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{a}_k|0, \boldsymbol{T}_u^{-1}), \quad (14)$$

$$p(\boldsymbol{B}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{b}_k|0, \boldsymbol{T}_v^{-1}), \quad (15)$$

$$p(\boldsymbol{a}) = \mathcal{N}(\boldsymbol{a}_0|0, (\boldsymbol{T}_u^0)^{-1}), \quad (16)$$

$$p(\boldsymbol{b}) = \mathcal{N}(\boldsymbol{b}_0|0, (\boldsymbol{T}_v^0)^{-1}), \quad (17)$$

where the precision matrices $\boldsymbol{T}_u, \boldsymbol{T}_u^0 \in \mathbb{R}^{D_u \times D_u}$ and $\boldsymbol{T}_v^0, \boldsymbol{T}_v \in \mathbb{R}^{D_v \times D_v}$ are assumed to be diagonal.

In standard BMF (Figure1-(a)), means of the Gaussian-Wishart priors, $\boldsymbol{m}$ and $\boldsymbol{n}$, are shared for all users and items and they generated from the non-informative constant vectors, $\boldsymbol{m}_0$ and $\boldsymbol{n}_0$, which are trivially set to zero. However our HBMFSI places individual means for Gaussian-Wishart priors, $\{\boldsymbol{m}_i\}_{i=1}^I$ and $\{\boldsymbol{n}_j\}_{j=1}^J$, by associating with corresponding side information via regression:

$$\boldsymbol{m}_i = \boldsymbol{m}_0 + \boldsymbol{A}^\top \boldsymbol{f}_i + \boldsymbol{e}_i, \quad \text{for } i = 1, \ldots, I, \quad (18)$$

$$\boldsymbol{n}_j = \boldsymbol{n}_0 + \boldsymbol{B}^\top \boldsymbol{g}_j + \boldsymbol{e}_j, \quad \text{for } j = 1, \ldots, J. \quad (19)$$

Note that HBMFSI can be reduced to RLFM (Figure 1-(b)) if we marginalize out $\{\boldsymbol{m}_i\}_{i=1}^I$ and $\{\boldsymbol{n}_j\}_{j=1}^J$ by using the property of the linear Gaussian model. After marginalization, the conditional distributions over factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ become

$$p(\boldsymbol{U}|\boldsymbol{\Lambda}_u, \boldsymbol{A}, \boldsymbol{m}_0) = \prod_{i=1} \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{m}_0 + \boldsymbol{A}^\top \boldsymbol{f}_i, (\gamma \boldsymbol{\Lambda}_u)^{-1}),$$

$$p(\boldsymbol{V}|\boldsymbol{\Lambda}_v, \boldsymbol{B}, \boldsymbol{n}_0) = \prod_{j=1} \mathcal{N}(\boldsymbol{v}_j|\boldsymbol{n}_0 + \boldsymbol{B}^\top \boldsymbol{g}_j, (\gamma \boldsymbol{\Lambda}_v)^{-1}),$$

where $\gamma = \gamma_0/(\gamma_0 + 1)$. These conditional distributions are equivalent to priors over factor matrices in RLFM (4, 5), provided that $\boldsymbol{m}_0 = \boldsymbol{n}_0 = 0$ and precision matrices, $\boldsymbol{\Lambda}_u$ and $\boldsymbol{\Lambda}_v$, are set to isotropic.

3.2 Variational Inference

We approximately computes posterior distributions over factor matrices by maximizing a lower-bound on the marginal log-likelihood. Let $\mathcal{Z}$ be a set of all latent variables:

$$\mathcal{Z} = \{\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Lambda}_u, \boldsymbol{\Lambda}_v, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{a}_0, \boldsymbol{b}_0\}. \quad (20)$$

The marginal log-likelihood $\log p(\boldsymbol{X})$ is given by

$$\begin{aligned} \log p(\boldsymbol{X}) &= \log \int p(X, \mathcal{Z}) d\mathcal{Z} \\ &\geq \int q(\mathcal{Z}) \log \frac{p(\boldsymbol{X}, \mathcal{Z})}{q(\mathcal{Z})} d\mathcal{Z} \\ &\equiv \mathcal{F}(q), \end{aligned} \quad (21)$$

where Jensen's inequality was used to obtain the *variational lower-bound* $\mathcal{F}(q)$ and $q(\mathcal{Z})$ is *variational distribution* that is assumed to be factorized:

$$q(\mathcal{Z}) = q(\boldsymbol{U})q(\boldsymbol{V})q(\boldsymbol{\Lambda}_u)q(\boldsymbol{\Lambda}_v)q(\boldsymbol{A})q(\boldsymbol{B})q(\boldsymbol{a}_0)q(\boldsymbol{b}_0). \quad (22)$$

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

$$q^*(\boldsymbol{Z}) \propto \exp\left\{\langle \log p(\boldsymbol{X}, \mathcal{Z}) \rangle_{q(\mathcal{Z} \setminus \boldsymbol{Z})}\right\}, \forall \boldsymbol{Z} \in \mathcal{Z}, \quad (23)$$

where the expectation is taken with respect to the variational distributions over all variables excluding $\boldsymbol{Z}$. Variational posterior distributions and corresponding updating equations for all latent variables $\boldsymbol{Z} \in \mathcal{Z}$ are summarized in Table 1.

We use the empirical Bayes estimation to update hyperparameters $\{\rho, \boldsymbol{m}_0, \boldsymbol{n}_0, \boldsymbol{T}_u^0, \boldsymbol{T}_u, \boldsymbol{T}_v^0, \boldsymbol{T}_v\}$. Setting derivatives of the variational lower-bound $\mathcal{F}(q)$ to zero yields the updating equations for hyperparameters:

$$\rho = |\Omega| / \sum_{(i,j) \in \Omega} \mathcal{E}_{ij}, \quad (24)$$

where

$$\mathcal{E}_{ij} = \left\langle \left(X_{ij} - \boldsymbol{u}_i^\top \boldsymbol{v}_j - \boldsymbol{a}_0^\top \boldsymbol{f}_i - \boldsymbol{b}_0^\top \boldsymbol{g}_j \right)^2 \right\rangle, \quad (25)$$

and

$$\boldsymbol{m}_0 = \frac{1}{I} \sum_{i=1}^{I} \left[\langle \boldsymbol{u}_i \rangle - \langle \boldsymbol{A} \rangle^\top \boldsymbol{f}_i \right], \quad (26)$$

$$\boldsymbol{n}_0 = \frac{1}{J} \sum_{j=1}^{J} \left[\langle \boldsymbol{v}_j \rangle - \langle \boldsymbol{B} \rangle^\top \boldsymbol{g}_j \right], \quad (27)$$

Table 1: Variational posteriors and corresponding updating equations are summarized. The symbol $\langle\cdot\rangle$ is the statistical expectation with respect to the corresponding variational distribution, Ω_i is a set of indices j for which X_{ij} is observed, $\boldsymbol{\Gamma}_u \in \mathbb{R}^{K\times K}$ is a diagonal matrix with K elements, $\boldsymbol{\lambda}_k \in \mathbb{R}^K$ is the kth column vector of $\langle\boldsymbol{\Lambda}_u\rangle$, $\mathbf{1}_u = [1,...,1]^\top \in \mathbb{R}^I$, $\widetilde{X}_{i,j} = X_{i,j} - \langle\boldsymbol{a}_0\rangle^\top \boldsymbol{f}_i - \langle\boldsymbol{b}_0\rangle^\top \boldsymbol{g}_j$, and $\widehat{\boldsymbol{C}}_u = \sum_{i=1}^I \boldsymbol{f}_i\boldsymbol{f}_i^\top$. Variational posterior for latent variables corresponding to items can be computed in similar way.

Variational distributions	Updating equations for variational parameters	Sufficient statistics
$q^*(\boldsymbol{U})$ $= \prod_{i=1}^I \mathcal{N}(\boldsymbol{u}_i\|\boldsymbol{\psi}_i, \boldsymbol{L}_i^{-1})$	$\boldsymbol{L}_i = \gamma\langle\boldsymbol{\Lambda}\rangle + \rho\sum_{j\in\Omega_i}\langle\boldsymbol{v}_j\boldsymbol{v}_j^\top\rangle$ $\boldsymbol{\psi}_i = \boldsymbol{L}_i^{-1}\Big\{\gamma\langle\boldsymbol{\Lambda}^u\rangle\Big(\boldsymbol{m}_0 + \langle\boldsymbol{A}\rangle^\top\boldsymbol{f}_i\Big) + \rho\sum_{j\in\Omega_i}\widetilde{X}_{i,j}\langle\boldsymbol{v}_j\rangle\Big\}$	$\langle\boldsymbol{u}_i\rangle = \boldsymbol{\psi}_i$ $\langle\boldsymbol{u}_i\boldsymbol{u}_i^\top\rangle = \boldsymbol{L}_i^{-1} + \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top$
$q^*(\boldsymbol{\Lambda}_u)$ $= \mathcal{W}(\boldsymbol{\Lambda}_u\|\boldsymbol{W}_u, \nu_u)$	$\nu_u = \nu_0 + I$ $\boldsymbol{W}_u^{-1} = \boldsymbol{W}_0^{-1} + \gamma\Big\{\Big(\langle\boldsymbol{U}\rangle - \langle\boldsymbol{A}\rangle^\top\boldsymbol{F} - \boldsymbol{m}_0\mathbf{1}_u^\top\Big)$ $\cdot\Big(\langle\boldsymbol{U}\rangle - \langle\boldsymbol{A}\rangle^\top\boldsymbol{F} - \boldsymbol{m}_0\mathbf{1}_u^\top\Big)^\top + \sum_{i=1}^I\boldsymbol{L}_i^{-1} + \boldsymbol{\Gamma}_u\Big\}$	$\langle\boldsymbol{\Lambda}_u\rangle = \nu_u\boldsymbol{W}_u$ $\Big([\boldsymbol{\Gamma}_u]_{k,k} = \mathrm{Tr}\big(\widehat{\boldsymbol{C}}_u\boldsymbol{H}_k^{-1}\big)\Big)$
$q^*(\boldsymbol{A})$ $= \prod_{k=1}^K \mathcal{N}(\boldsymbol{a}_k\|\boldsymbol{\varphi}_k, \boldsymbol{H}_k^{-1})$	$\boldsymbol{H}_k = \boldsymbol{T}_u + \gamma[\langle\boldsymbol{\Lambda}_u\rangle]_{k,k}\widehat{\boldsymbol{C}}_u$ $\boldsymbol{\varphi}_k = \gamma\boldsymbol{H}_k^{-1}$ $\cdot\Big\{\sum_{i=1}^I\boldsymbol{\lambda}_k^\top(\langle\boldsymbol{u}_i\rangle - \boldsymbol{m}_0)\boldsymbol{f}_i - \widehat{\boldsymbol{C}}_u\sum_{k'\neq k}[\boldsymbol{\lambda}_k]_{k'}\langle\boldsymbol{a}_{k'}\rangle\Big\}$	$\langle\boldsymbol{a}_k\rangle = \boldsymbol{\varphi}_k$, $\langle\boldsymbol{a}_k\boldsymbol{a}_k^\top\rangle = \boldsymbol{H}_k^{-1} + \boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^\top$
$q^*(\boldsymbol{a}_0)$ $= \mathcal{N}(\boldsymbol{a}_0\|\boldsymbol{\varphi}_0, \boldsymbol{H}_0^{-1})$	$\boldsymbol{H}_0 = \boldsymbol{T}_u^0 + \rho\sum_{i=1}^I\Big[\|\Omega_i\|\boldsymbol{f}_i\boldsymbol{f}_i^\top\Big]$ $\boldsymbol{\varphi}_0 = \rho\boldsymbol{H}_0^{-1}$ $\cdot\Big\{\sum_{i=1}^I\boldsymbol{f}_i\sum_{j\in\Omega_i}\Big(X_{i,j} - \langle\boldsymbol{u}_i\rangle^\top\langle\boldsymbol{v}_j\rangle - \langle\boldsymbol{b}_0\rangle^\top\boldsymbol{g}_j\Big)\Big\}$	$\langle\boldsymbol{a}_0\rangle = \boldsymbol{\varphi}_0$, $\langle\boldsymbol{a}_0\boldsymbol{a}_0^\top\rangle = \boldsymbol{H}_0^{-1} + \boldsymbol{\varphi}_0\boldsymbol{\varphi}_0^\top$

and

$$(\boldsymbol{T}_u^0)_{d,d} = 1/\left\langle\boldsymbol{a}_0\boldsymbol{a}_0^\top\right\rangle_{d,d}, \quad (28)$$

$$(\boldsymbol{T}_u)_{d,d} = K/\sum_{k=1}^K\left\langle\boldsymbol{a}_k\boldsymbol{a}_k^\top\right\rangle_{d,d}, \quad (29)$$

for $d = 1,\ldots,D_u$ and

$$(\boldsymbol{T}_v^0)_{d,d} = 1/\left\langle\boldsymbol{b}_0\boldsymbol{b}_0^\top\right\rangle_{d,d}, \quad (30)$$

$$(\boldsymbol{T}_v)_{d,d} = K/\sum_{k=1}^K\left\langle\boldsymbol{b}_k\boldsymbol{b}_k^\top\right\rangle_{d,d}, \quad (31)$$

for $d = 1,\ldots,D_v$.

3.3 Prediction

We consider two kinds of prediction tasks in the collaborative filtering: *hold-out* prediction and *fold-in* prediction. In the hold-out prediction, we aim to predict a missing entry X_{i^o,j^o} in the rating matrix $\boldsymbol{X}$, where $(i^o,j^o) \notin \Omega$. To do this, we simply make use of the posterior means of each variable:

$$\widehat{X}_{i^0,j^0} = \langle\boldsymbol{u}_{i^o}\rangle^\top\langle\boldsymbol{v}_{j^o}\rangle + \langle\boldsymbol{a}_0\rangle^\top\boldsymbol{f}_{i^o} + \langle\boldsymbol{b}_0\rangle^\top\boldsymbol{g}_{j^o}. \quad (32)$$

In the fold-in prediction, we want to predict the rating value of a new user or new item. We assume that a user i^+ is newly added into the system and some rating entries $\boldsymbol{x}_{i^+}$ made by the user i^+ are available. Then the prediction of an unknown entry of the user i^+ can be made similarly to (32):

$$\widehat{X}_{i^+,j} = \langle\boldsymbol{u}_{i^+}\rangle^\top\langle\boldsymbol{v}_j\rangle + \langle\boldsymbol{a}_0\rangle^\top\boldsymbol{f}_{i^+} + \langle\boldsymbol{b}_0\rangle^\top\boldsymbol{g}_j, \quad (33)$$

where $\langle\boldsymbol{u}_{i^+}\rangle$ is a mean of the predictive distribution over a new factor $\boldsymbol{u}_{i^+}$ given $\boldsymbol{x}_{i^+}$, which can be defined by

$$p(\boldsymbol{u}_{i^+}|\boldsymbol{x}_{i^+},\boldsymbol{X}) = \int p(\boldsymbol{u}_{i^+}|\boldsymbol{x}_{i^+},\mathcal{Z})p(\mathcal{Z}|\boldsymbol{X})d\mathcal{Z}, \quad (34)$$

where $\mathcal{Z}$ is the set of latent variables defined in (20). We approximate the predictive distribution (34) by maximizing a lower-bound on marginal log-posterior $p(\boldsymbol{x}_{i^+}|\boldsymbol{X})$:

$$\begin{aligned}\log p(\boldsymbol{x}_{i+}|\boldsymbol{X}) &= \log\int\int p(\boldsymbol{x}_{i^+},\boldsymbol{u}_{i^+},\mathcal{Z}|\boldsymbol{X})d\boldsymbol{u}_{i^+}d\mathcal{Z}\\ &\geq \int\int\tilde{q}(\boldsymbol{u}_{i^+},\mathcal{Z})\log\frac{p(\boldsymbol{x}_{i^+},\boldsymbol{u}_{i^+},\mathcal{Z}|\boldsymbol{X})}{\tilde{q}(\boldsymbol{u}_{i^+},\mathcal{Z})}d\boldsymbol{u}_{i^+}d\mathcal{Z}\\ &\equiv \mathcal{F}(\tilde{q}). \end{aligned} \quad (35)$$

Variational distribution $\tilde{q}(\boldsymbol{u}_{i^+},\mathcal{Z})$ is assumed to be factorized:

$$\tilde{q}(\boldsymbol{u}_{i^+},\mathcal{Z}) = \tilde{q}(\boldsymbol{u}_{i^+})\tilde{q}^*(\mathcal{Z}), \quad (36)$$

where $\tilde{q}^*(\mathcal{Z})$ is fixed to the pre-trained variational distribution in Table 1. For given pre-fixed $\tilde{q}^*(\mathcal{Z})$, the optimal variational posterior $\tilde{q}^*(\boldsymbol{u}_{i^+})$ maximizing (35) is computed with single update,

$$\tilde{q}^*(\boldsymbol{u}_{i^+}) \propto \exp\Big\{\big\langle\log p(\boldsymbol{x}_{i^+}|\boldsymbol{u}_{i^+},\mathcal{Z})p(\boldsymbol{u}_i^+|\mathcal{Z})\big\rangle_{\tilde{q}^*(\mathcal{Z})}\Big\}, \quad (37)$$

yielding to Gaussian distribution such that

$$\tilde{q}^*(\boldsymbol{u}_{i^+}) = \mathcal{N}(\boldsymbol{u}_{i^+}|\boldsymbol{\psi}_{i^+},\boldsymbol{L}_{i^+}^{-1}), \quad (38)$$

where

$$\begin{aligned}\boldsymbol{L}_{i^+} &= \gamma\langle\boldsymbol{\Lambda}_u\rangle + \langle\rho\rangle\sum_{j\in\Omega_{i^+}}\langle\boldsymbol{v}_j\boldsymbol{v}_j^\top\rangle, \quad (39)\\ \boldsymbol{\psi}_{i^+} &= \boldsymbol{L}_{i^+}^{-1}\Big\{\gamma\langle\boldsymbol{\Lambda}_u\rangle\Big(\boldsymbol{m}_0 + \langle\boldsymbol{A}\rangle^\top\boldsymbol{f}_{i^+}\Big)\\ &\quad + \langle\rho\rangle\sum_{j\in\Omega_{i^+}}\Big(X_{i^+,j} - \langle\boldsymbol{a}_0\rangle^\top\boldsymbol{f}_{i^+} - \langle\boldsymbol{b}_0\rangle^\top\boldsymbol{g}_j\Big)\langle\boldsymbol{v}_j\rangle\Big\}.\end{aligned}$$

Thus the prediction of unknown rating entry of the new user, $X_{i^+,j}$ can be made based on (33) and (38):

$$\widehat{X}_{i^+,j} = \boldsymbol{\psi}_{i^+}^\top \boldsymbol{v}_j + \langle \boldsymbol{a}_0 \rangle^\top \boldsymbol{f}_{i^+} + \langle \boldsymbol{b}_0 \rangle^\top \boldsymbol{g}_j. \quad (40)$$

The aforementioned prediction method can be similarly applied to predict missing ratings in the presence of a new item.

4 Bayesian Cramér-Rao Bound

Motivated by [Yoo and Choi, 2011], we provide Bayesian Cramér-Rao Bound for our model, showing that our HMBFSI improves the reconstruction over the standard BMF where the side information is not used.

The Bayesian Cramér-Rao bound or Posterior Cramér-Rao bound places a lower bound on the variance of any parametric estimator [Tichavsky *et al.*, 1998], as the inverse of the Fisher information matrix $\mathcal{F}$, which is written by,

$$\left\langle (\boldsymbol{\theta} - \boldsymbol{\theta}(\boldsymbol{x}))(\boldsymbol{\theta} - \boldsymbol{\theta}(\boldsymbol{x}))^\top \right\rangle \succeq \mathcal{F}^{-1}, \quad (41)$$

where $\boldsymbol{\theta}(\boldsymbol{x})$ is the estimate of the $\boldsymbol{\theta}$. Each element of the Fisher information matrix is defined by

$$\mathcal{F}_{i,j} = -\left\langle \frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\rangle_{p(\boldsymbol{x},\boldsymbol{\theta})}, \quad (42)$$

where $p(\boldsymbol{x}, \boldsymbol{\theta})$ is the joint probability between the observations and the parameters. The benefit of using BCRB over classical Cramér-Rao bound is that the BCRB is known to provide a lower bound on the variance of any parametric estimators, even for the unbiased ones [Tichavsky *et al.*, 1998].

To compute the BCRB for HBMFSI, we rearrange the factor matrices and other parameters to be a vector form:

$$\boldsymbol{\theta} = [\boldsymbol{u}_1^\top \cdots \boldsymbol{u}_I^\top \boldsymbol{v}_1^\top \cdots \boldsymbol{v}_J^\top \boldsymbol{m}_1^\top \cdots \boldsymbol{m}_I^\top \boldsymbol{n}_1^\top \cdots \boldsymbol{n}_J^\top \\ \mathrm{vec}(\boldsymbol{\Lambda}_u)^\top \mathrm{vec}(\boldsymbol{\Lambda}_v)^\top \mathrm{vec}(\boldsymbol{A})^\top \mathrm{vec}(\boldsymbol{B})^\top \boldsymbol{a}_0^\top \boldsymbol{b}_0^\top]^\top. \quad (43)$$

Each element of the Fisher information matrix is computed as (42). The second derivatives of the log joint probability with respect to elements of factor matrices are given by

$$\begin{aligned}
\frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial^2 U_{k,i}} &= -[\boldsymbol{\Lambda}_u]_{k,k} - \rho \sum_{j \in \Omega_i} V_{k,j}^2, \\
\frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial U_{k,i} \partial U_{k^+,i}} &= -[\boldsymbol{\Lambda}_u]_{k,k^+} - \rho \sum_{j \in \Omega_i} V_{k,j} V_{k^+,j}, \\
\frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial U_{k,i} \partial U_{k,i^+}} &= \frac{\partial^2 \log p(\boldsymbol{X}, \boldsymbol{\theta})}{\partial U_{k,i} \partial U_{k^+,i^+}} = 0, \\
\frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial U_{k,i} \partial V_{k,j}} &= -\rho \left\{ (X_{i,j} - \boldsymbol{u}_i^\top \boldsymbol{v}_j) - U_{k,i} V_{k,j} \right\}, \\
\frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial U_{k,i} \partial V_{k^+,j}} &= -\rho U_{k^+,i} V_{k,j}.
\end{aligned}$$

In standard BMF where side information is not used, if we set $\boldsymbol{m}_0$, $\boldsymbol{n}_0$ to zero vector and $\boldsymbol{W}_0$ to an identity matrix as in [Salakhutdinov and Mnih, 2008a], the most of the expectation above second derivatives are vanished and only nonzero values come from the differentiating with the same parameter,

$$-\left\langle \frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial^2 U_{k,i}} \right\rangle_{p(\boldsymbol{x},\boldsymbol{\theta})} = \nu_0 + \frac{\rho |\Omega_i| (\gamma_0 + 1) \nu_0}{\gamma_0 (\nu_0 - K - 1)}, \quad (44)$$

hence the Fisher information matrix becomes diagonal.

The Fisher information matrix of HBMFSI is also diagonal but larger than (44) because of extra term involved with side information,

$$\begin{aligned}
&-\left\langle \frac{\partial^2 \log p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial^2 U_{k,i}} \right\rangle_{p(\boldsymbol{x},\boldsymbol{\theta})} \quad (45) \\
&= \frac{\nu_0 + \rho |\Omega_i| (\gamma_0 + 1) \nu_0}{\gamma_0 (\nu_0 - K - 1)} + \rho \sum_{j \in \Omega_i} \boldsymbol{g}_j^\top \boldsymbol{T}_v^{-1} \boldsymbol{g}_j,
\end{aligned}$$

hence has lower BCRB (the inverse of the fisher information matrix) and also lower reconstruction error [Yoo and Choi, 2011].

5 Numerical Experiments

We evaluated the proposed HBMFSI for the collaborative prediction problem in both warm-start and cold-start situations, and compared the performance with that of BMF [Salakhutdinov and Mnih, 2008a], BMCF [Yoo and Choi, 2012], BMFSI [Proteous *et al.*, 2010] and RLFM [Agarwal and Chen, 2009] to show the benefit of the HBMFSI.

Datasets: We conducted experiments on two MovieLens datasets: MovieLens-100K, which consists of $100K$ ratings with 943 users and 1682 movies, and MovieLens-1M, which consisits of $1M$ ratings with 6040 users and 3706 movies. MovieLens datasets provide side information for both users and items. The user information which consists of the user's age, gender and occupation was encoded into a binary valued vector with length 28. Similarity, the item information which consists of the 18 category of movie genre was ecoded into a binary valued vector with length 18. Ratings were normalized to be zero-mean.

Evaluation Metris: The performance of the collaborative predction algorithm was evaluated in terms of the widely used root mean squared error (RMSE) defined by RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^N (x_i - \hat{x}_i)^2}$, and mean absolute error (MAE) defined by MAE $= \frac{1}{N}\sum_{i=1}^N |x_i - \hat{x}_i|$, where N is the total number of test data points, x_i and $\hat{x}_i$ are the true rating and predcited rating of the i-th test data, respectively.

Model Training: Because each model was trained by different inferece algortihm such as Markov Chain Mornte Carlo (BMF, BMFSI), Monte Carlo Expectation Maximization (RLFM), and the variational method (BMCF, HBMFSI), it is difficult to directly compare the effectiveness of the modeling for side information. For this reason, we decided to train all models by the variational method. We set the latent dimension K to 10 (compare to the Netflix data, relatively small K is sufficient for MovieLens datasets), and hyperparameters of Gaussian-Wishart priors to $\nu_0 = K$, $\gamma_0 = 1$ and $\boldsymbol{W}_0 = \boldsymbol{I}$.

Warm-start Performance: Firstly we investigate the predictive performance on the common warm-start situation with MovieLens-1M data. The data is randomly split around 50% taring and 50% test. Experiments are repeated ten times, and we report their averages and standard deviations in Table 2. We confirmed that HBMFSI outperforms other methods in terms of both MAE and RMSE. Interestingly, BMFSI and BMCF showed slightly worse performance than BMF, even

Table 2: MAE and RMSE on MovieLens-1M data in warm-start situation. The number in parenthesis represents the standard deviation.

Method	MAE	RMSE
BMF	0.6761(0.0005)	0.8621(0.0005)
BMFSI	0.6785(0.0005)	0.8636(0.0005)
BMCF	0.6818(0.0009)	0.8666(0.0008)
RLMF	0.6718(0.0004)	0.8561(0.0004)
HBMFSI	**0.6698(0.0005)**	**0.8536(0.0005)**

though side information is utilized. In the case of BMFSI, similar result was reported in [Proteous *et al.*, 2010], where movie metadata which extracted from Wikipedia (e.g. movie director, actors, languages) do not improve performance measured by RMSE in Netflix Prize data.

Cold-start Performance: Next we investigate the predictive performance on the cold-start situation with MovieLens-100K data. We considered the cold-start problems which can occur in the hold-out prediction (prediction of missing ratings of the existing users) and also in the fold-in prediction (prediction of unknown ratings of the new users).

In the case of hold-out prediction, we randomly select 20% ratings as the test data and train each model on ten taring sets, which consist of 10%, 20%,. . .,100% of the remained ratings. In the case of fold-in prediction, we randomly divide the users into two sets: a test set containing 20% of the users and training set containing the rest of the users. First each model is trained on all ratings by the training users. For each of the test users we train the model varying with the number of given ratings from 0 to 20 according to the procedure outlined in Section 3.3. Experiments are repeated ten times for both hold-out and fold-in predictions.

Figure 2 shows that HBMFSI outperforms BMF both in hold-out and fold-in prediction. Especially HBMFSI significantly improves the performance in fold-in prediction which reflects practical cold-start situation more than hold-out prediction. Experimental results also showed that side information modeling used in RLFM and HBMFSI, where side information is used to modeling the means of priors over factor matrices is more efficient to utilize side information than other methods including BMCF and BMFSI. Finally HBMFSI showed slightly better performance than RLFM, which is a special case of our model.

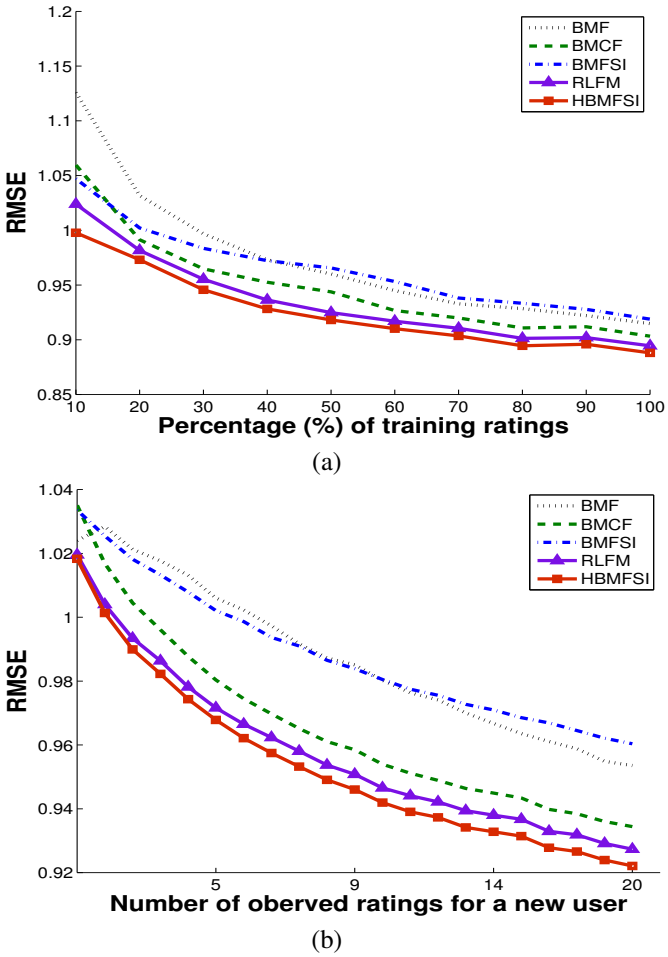

Figure 2: RMSE on MovieLens-100k data in cold-start situdation: (a) hold-out prediction case, (b) fold-in prediction.

6 Conclusions

We have presented hierarchical Bayesian matrix factorization with side information, to handle cold-start problems. We used side information to properly regularize user and item factor matrices by placing Gaussian-Wishart priors on mean vectors and precision matrices of Gaussian user and item factor matrices, such that mean of each prior distribution is regressed on corresponding side information. We developed variational inference algorithms to approximately compute posterior distributions over user and item factor matrices. In addition, we provided Bayesian Cramér-Rao Bound for our model, showing that the hierarchical Bayesian matrix factorization with side information improves the reconstruction over the standard Bayesian matrix factorization where the side information is not used. Experiments on MovieLens data demonstrated the useful behavior of our model in the case of cold-start and also warm-start problems.

7 Acknowledgments

This work was supported by National Research Foundation (NRF) of Korea (2012-0005032), NIPA ITRC Support Program (NIPA-2013-H0301-13-3002), POSTECH Rising Star Program, and NRF World Class University Program (R31-10100).

References

[Agarwal and Chen, 2009] D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Paris, France, 2009.

[Kim and Choi, 2013] Y.-D. Kim and S. Choi. Variational Bayesian view of weighted trace norm regularization for

matrix factorization. *IEEE Signal Processing Letters*, 20(3):261–264, 2013.

[Koren *et al.*, 2009] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[Lim and Teh, 2007] 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.

[Proteous *et al.*, 2010] I. Proteous, A. Asuncion, and M. Welling. Bayesian matrix factorization with side information and Dirichlet process mixtures. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*, Atlanta, Georgia, USA, 2010.

[Raiko *et al.*, 2007] 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)*, pages 691–698, Warsaw, Poland, 2007.

[Rennie and Srebro, 2005] 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.

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

[Salakhutdinov and Mnih, 2008b] R. Salakhutdinov and A. Mnih. Probablistic matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20. MIT Press, 2008.

[Singh and Gordon, 2008] 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.

[Singh and Gordon, 2010] A. P. Singh and G. J. Gordon. A Bayesian matrix factorization model for relational data. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, CA, 2010.

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

[Tichavsky *et al.*, 1998] P. Tichavsky, C. H. Muravchik, and A. Nehorai. Posterior Cramér-Rao bounds for discrete-time nonlinear filtering. *IEEE Transactions on Signal Processing*, 46(5):1386–1395, 1998.

[Yoo and Choi, 2011] J. Yoo and S. Choi. 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.

[Yoo and Choi, 2012] J. Yoo and S. Choi. Hierarchical variational Bayesian matrix co-factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.