Bayesian Matrix Co-Factorization: Variational Algorithm and Cramér-Rao Bound

Seungjin Choi
(joint work with Jiho Yoo)

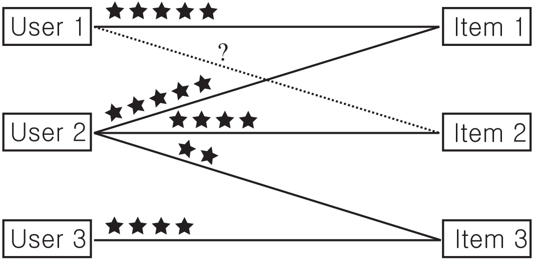
Department of Computer Science
Pohang University of Science and Technology, Korea
seungjin@postech.ac.kr
http://www.postech.ac.kr/∼seungjin

September 6, 2011

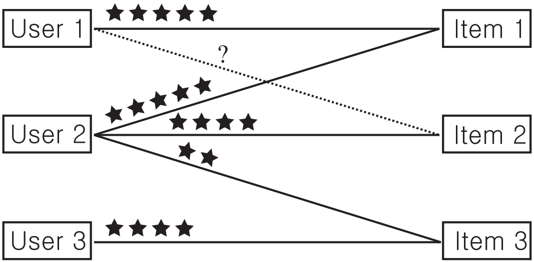
Outline

- Problem of interest
  - Matrix factorization for collaborative prediction
  - Cold-start problem
- Variational Bayesian matrix co-factorization
  - Probabilistic models and variational inference
  - Bayesian Cramér-Rao bound
- Numerical experiments
- Conclusions

Collaborative Prediction



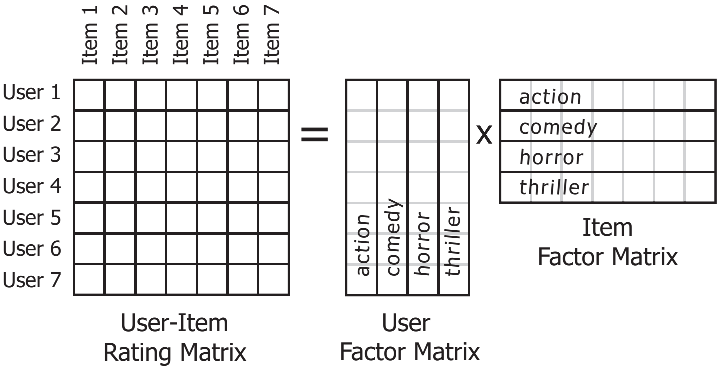
Collaborative Prediction



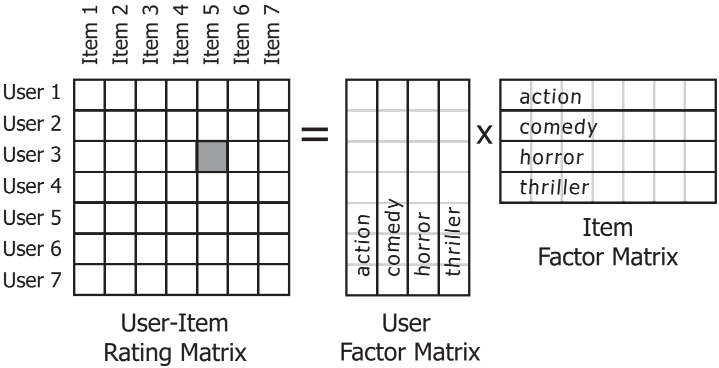
Collaborative prediction

- The task of predicting preferences of users, based on their own available preferences as well as preferences of other users who share similar preferences
- Methods
  - Memory-based methods
  - Model-based methods (matrix factorization)

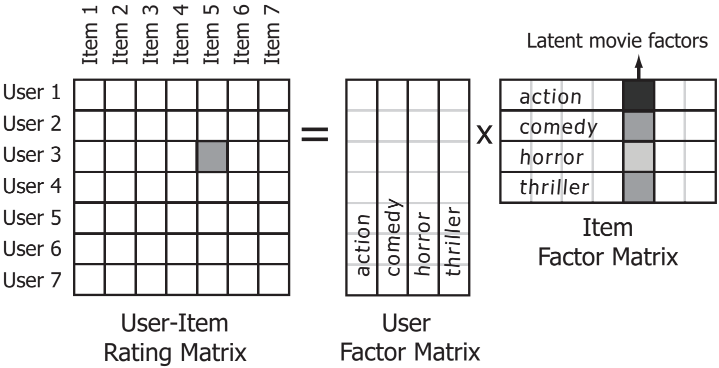
Matrix Factorization



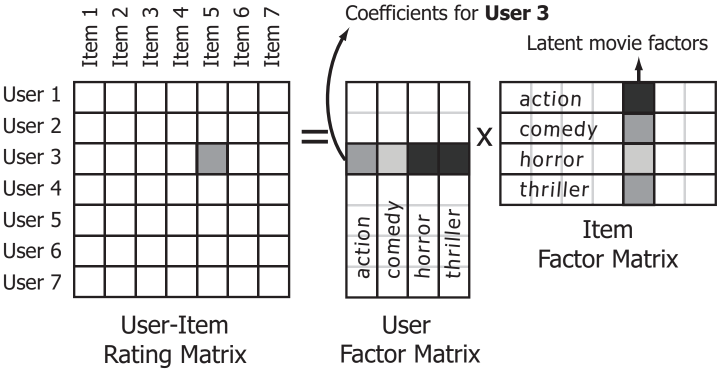
Matrix Factorization



Matrix Factorization



Matrix Factorization

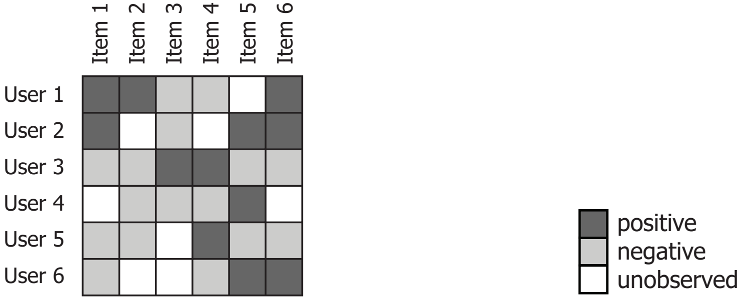


User-Item Rating Matrix

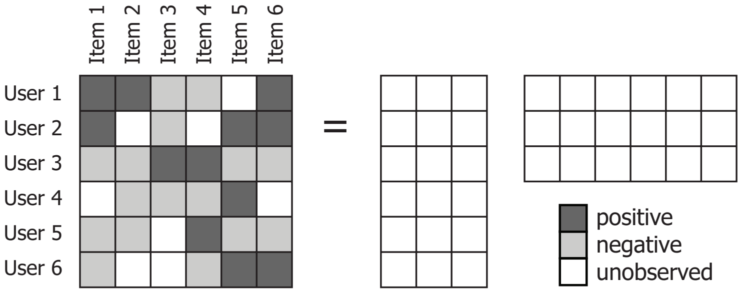
	Item 1	Item 2	Item 3	Item 4	Item 5	...
User 1	5	0	5	0	0	...
User 2	0	0	0	5	0	...
User 3	0	0	2	0	0	...
User 4	2	5	4	0	3	...
User 5	1	0	1	5	0	...
...	...	...	...	...	...	...

Most of the entries are not rated (value 0)

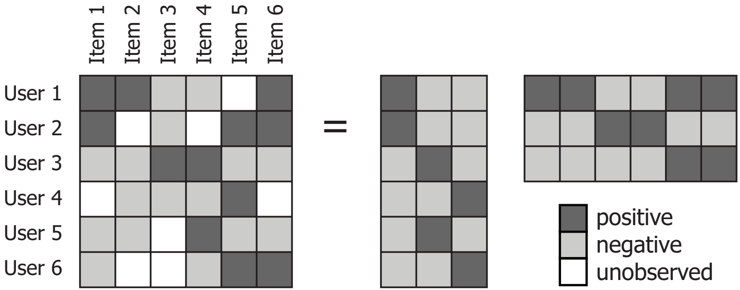
Matrix Factorization for Collaborative Prediction



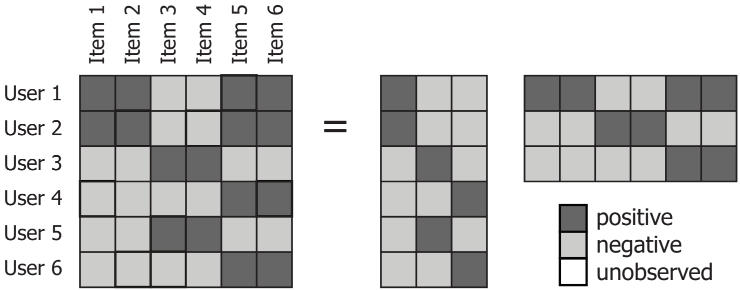
Matrix Factorization for Collaborative Prediction



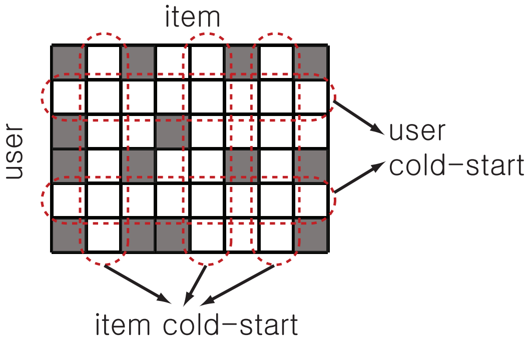
Matrix Factorization for Collaborative Prediction



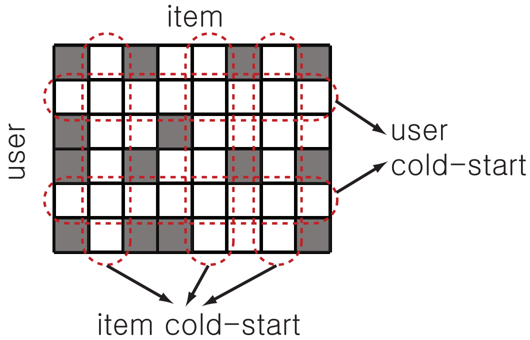
Matrix Factorization for Collaborative Prediction



Cold Start Problems



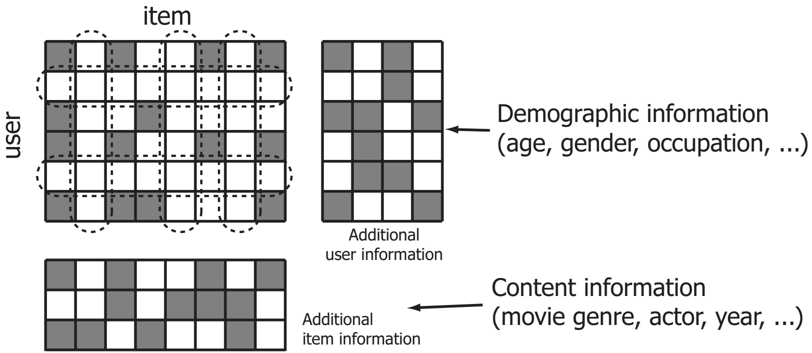
Cold Start Problems



Cold start problems

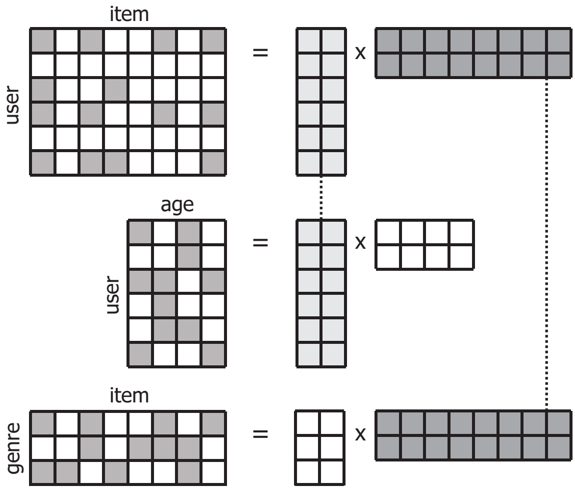
- Extremely small number of ratings or no ratings at all for some users or items
- Not able to accurately predict preferences for cold-start users or cold-start items

Side Information



Matrix Co-Factorization

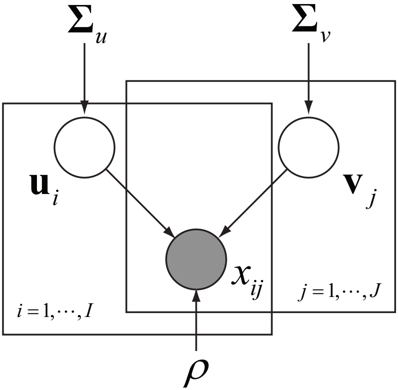
Input matrices are jointly decomposed, sharing some factor matrices.



Related Work on Matrix Co-Factorization

Authors	Side Information	Work
Yu *et al.*, 2005	label	supervised LSI
Zhu *et al.*, 2007	content+link	information retrieval
Singh & Gordon, 2008	relational	collective matrix factorization
Williamson & Ghahramani, 2008	relational	probabilistic models
Lee & Choi, 2009	inter+intra subject	group NMF
Yoo & Choi, 2009	relational	matrix co-tri-factorization
Lee & Choi, 2010	label	semi-supervised NMF
Singh & Gordon, 2010	relational	Bayesian factorization (sampling)
Yoo *et al.*, 2010	drum	drum source separation
Yoo & Choi, 2011	uncompressed	compressed sensing

Bayesian Matrix Factorization: Empirical Variational Bayes



Lim and Teh, 2007
Raiko *et al.*, 2007

- Model

$$\begin{aligned} \mathbf{X} &= \mathbf{U}^\top \mathbf{V} + \mathbf{E}, \\ x_{ij} &= \mathbf{u}_i^\top \mathbf{v}_j + \epsilon_{ij}. \end{aligned}$$

- Gaussian likelihood

$$p(x_{ij} | \mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(x_{ij} |, 0, \rho).$$

- Priors ($\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_v$ are diagonal)

$$\begin{aligned} p(\mathbf{U}) &= \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i | 0, \boldsymbol{\Sigma}_u), \\ p(\mathbf{V}) &= \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j | 0, \boldsymbol{\Sigma}_v). \end{aligned}$$

Variational Inference

Marginal likelihood is given by

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

where the variational lower-bound is given by

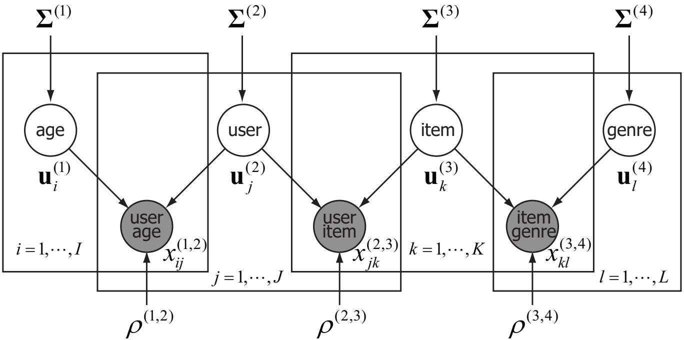
$$\mathcal{I}(q) = \int\int q(\mathbf{U},\mathbf{V}) \log p(\mathbf{X},\mathbf{U},\mathbf{V})d\mathbf{U}d\mathbf{V} - \int\int q(\mathbf{U},\mathbf{V}) \log q(\mathbf{U},\mathbf{V})d\mathbf{U}d\mathbf{V}.$$

Mean field approximation assumes that $q(\mathbf{U},\mathbf{V}) = q(\mathbf{U})q(\mathbf{V})$.

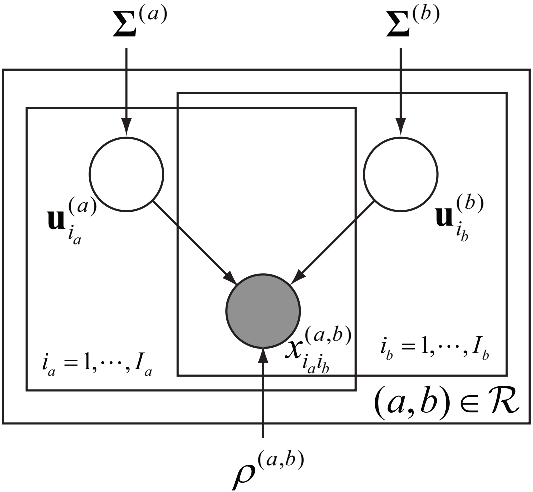
Variational posterior distributions $q(\mathbf{U})$ and $q(\mathbf{V})$ are computed by maximizing $\mathcal{I}(q)$, leading to

$$\begin{aligned}\log q(\mathbf{U}) &\propto \mathbb{E}_{q(V)} \left\{\log p(\mathbf{X},\mathbf{U},\mathbf{V})\right\}, \\ \log q(\mathbf{V}) &\propto \mathbb{E}_{q(U)} \left\{\log p(\mathbf{X},\mathbf{U},\mathbf{V})\right\}.\end{aligned}$$

Probabilistic Model for Matrix Co-Factorization



Probabilistic Model for Matrix Co-Factorization



Variational Inference for Matrix Co-Factorization

- A set of relational data matrix: $\mathcal{X} = \left\{ \mathbf{X}^{(a,b)} \right\}$ for $(a, b) \in \mathcal{R}$.
- A set of model parameters: $\mathcal{U} = \left\{ \mathbf{U}^{(a)} \right\}$ for $a \in \mathcal{E}$.
- Variational lower bound on the log marginal likelihood is given by

$$\log p(\mathcal{X}) \geq \int q(\mathcal{U}) \log \frac{p(\mathcal{X}, \mathcal{U})}{q(\mathcal{U})} d\mathcal{U} = \mathcal{I}(q)$$

- Mean field approximation assumes that $q(\mathcal{U}) = \prod_{a \in \mathcal{E}} q\left(\mathbf{U}^{(a)}\right)$.
- Variational posterior distributions, which maximize $\mathcal{I}(q)$, are computed by

$$q_a\left(\mathbf{U}^{(a)}\right) \propto \exp\left\{ \mathbb{E}_{\mathcal{U} \setminus U^{(a)}} \left[\log p(\mathcal{X}, \mathcal{U})\right] \right\}.$$

Variational Posterior Distributions over Factor Matrices

Variational posterior distribution over factor matrices, $q_a\left(\mathbf{U}^{(a)}\right)$, are Gaussian, which are calculated as:

$$q_a\left(\mathbf{U}^{(a)}\right) = \prod_{i_a} \mathcal{N}\left(\mathbf{u}_{i_a}^{(a)} | \overline{\mathbf{u}}_{i_a}^{(a)}, \mathbf{\Phi}_{i_a}^{(a)}\right),$$

Variational Posterior Distributions over Factor Matrices

Variational posterior distribution over factor matrices, $q_a\left(\mathbf{U}^{(a)}\right)$, are Gaussian, which are calculated as:

$$q_a\left(\mathbf{U}^{(a)}\right) = \prod_{i_a} \mathcal{N}\left(\mathbf{u}_{i_a}^{(a)} | \overline{\mathbf{u}}_{i_a}^{(a)}, \mathbf{\Phi}_{i_a}^{(a)}\right),$$

where mean vectors and covariance matrices are given by

$$\overline{\mathbf{u}}_{i_a}^{(a)} = \mathbf{\Phi}_{i_a}^{(a)} \left(\sum_{b|(a,b)\in\mathcal{R}} \sum_{i_b|(i_a,i_b)\in\mathcal{O}^{(a,b)}} \frac{1}{\rho^{(a,b)}} x_{i_a i_b}^{(a,b)} \overline{\mathbf{u}}_{i_b}^{(b)} \right),$$

$$\left(\mathbf{\Phi}_{i_a}^{(a)}\right)^{-1} = \left(\mathbf{\Sigma}^{(a)}\right)^{-1} + \sum_{b|(a,b)\in\mathcal{R}} \sum_{i_b|(i_a,i_b)\in\mathcal{O}^{(a,b)}} \frac{\mathbf{\Phi}_{i_b}^{(b)} + \overline{\mathbf{u}}_{i_b}^{(b)} \overline{\mathbf{u}}_{i_b}^{(b)\top}}{\rho^{(a,b)}}.$$

Hyperparameters $\rho^{(a,b)}$ and $\boldsymbol{\Sigma}^{(a)}$ are estimated by maximizing the variational lower bound $\mathcal{I}(q)$.

$$\begin{aligned}
\rho^{(a,b)} &= \frac{1}{N^{(a,b)}} \sum_{(i_a,i_b)\in\mathcal{O}^{(a,b)}} \left\{ \left(x_{i_a i_b}^{(a,b)} \right)^2 - 2 x_{i_a i_b}^{(a,b)} \overline{\mathbf{u}}_{i_a}^{(a)\top} \overline{\mathbf{u}}_{i_b}^{(b)} \right\} \\
&+ \frac{1}{N^{(a,b)}} \sum_{(i_a,i_b)\in\mathcal{O}^{(a,b)}} \operatorname{tr}\left\{ \left(\boldsymbol{\Phi}_{i_a}^{(a)} + \overline{\mathbf{u}}_{i_a}^{(a)} \overline{\mathbf{u}}_{i_a}^{(a)\top} \right) \left(\boldsymbol{\Phi}_{i_b}^{(b)} + \overline{\mathbf{u}}_{i_b}^{(b)} \overline{\mathbf{u}}_{i_b}^{(b)\top} \right) \right\}, \\
\boldsymbol{\Sigma}^{(a)} &= \frac{1}{I^{(a)}} \operatorname{ddiag}\left(\sum_{i_a} \left[\boldsymbol{\Phi}_{i_a}^{(a)} + \overline{\mathbf{u}}_{i_a}^{(a)} \overline{\mathbf{u}}_{i_a}^{(a)\top} \right] \right).
\end{aligned}$$

Predictive Distribution

Predictive distribution is computed by

$$\begin{aligned} p(x_{i_a^* i_b^*}) &= \int\int p\left(x_{i_a^* i_b^*} \mid \mathbf{U}^{(a)}, \mathbf{U}^{(b)}\right) q_a^*\left(\mathbf{U}^{(a)}\right) q_b^*\left(\mathbf{U}^{(b)}\right) d\mathbf{U}^{(a)} d\mathbf{U}^{(b)}, \\ &= \mathcal{N}(x_{i_a^* i_b^*} \mid \overline{\mathbf{u}}_{i_a^*}^{(a)\top} \overline{\mathbf{u}}_{i_b^*}^{(b)}, \rho^{(a,b)}), \end{aligned}$$

which is Gaussian.

- *Hold-out* prediction

$$x_{i_a^* i_b^*} = \overline{\mathbf{u}}_{i_a^*}^{(a)\top} \overline{\mathbf{u}}_{i_b^*}^{(b)}.$$

- *Fold-in* prediction

$$\begin{aligned} \overline{\mathbf{u}}_{i_a^*}^{(a)} &= \mathbf{\Phi}_{i_a^*}^{(a)} \left(\sum_{c \mid (a,c) \in \mathcal{R}} \sum_{i_c \mid (i_a^*, i_c) \in \mathcal{O}^{(a,c)}} \frac{1}{\rho^{(a,c)}} x_{i_a^* i_c}^{(a,c)} \overline{\mathbf{u}}_{i_c}^{(c)} \right), \\ \left(\mathbf{\Phi}_{i_a^*}^{(a)}\right)^{-1} &= \left(\mathbf{\Sigma}^{(a)}\right)^{-1} + \sum_{c \mid (a,c) \in \mathcal{R}} \sum_{i_c \mid (i_a^*, i_c) \in \mathcal{O}^{(a,c)}} \frac{\mathbf{\Phi}_{i_c}^{(c)} + \overline{\mathbf{u}}_{i_c}^{(c)} \overline{\mathbf{u}}_{i_c}^{(c)\top}}{\rho^{(a,c)}}. \end{aligned}$$

Bayesian Cramér-Rao Bound

Cramér-Rao Bound

- A lower-bound on the variance of unbiased estimators

$$\mathbb{E}\left\{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\right\} \geq \mathcal{I}^{-1}.$$

- Fisher Information Matrix is computed by

$$\mathcal{I}_{ij} = \mathbb{E}_{\mathbf{x}}\left\{-\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right\}.$$

Bayesian Cramér-Rao Bound

Cramér-Rao Bound

- A lower-bound on the variance of unbiased estimators

$$\mathbb{E}\left\{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\right\} \geq \mathcal{I}^{-1}.$$

- Fisher Information Matrix is computed by

$$\mathcal{I}_{ij} = \mathbb{E}_{\mathbf{x}}\left\{-\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right\}.$$

Bayesian Cramér-Rao Bound

- A lower-bound on the variance of *any* estimators

$$\mathcal{I}_{ij} = \mathbb{E}_{\mathbf{x},\theta}\left\{-\frac{\partial^2 \log p(\mathbf{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right\}.$$

Fisher Information Matrices

Fisher Information Matrix in the case of Bayesian Matrix Co-Factorization

- Fisher information matrix turns out to be a diagonal matrix.
- Each diagonal entry becomes larger when more relational matrices are involved.

- Matrix Factorization

$$\mathbb{E}_{X,U}\left\{-\frac{\partial^2 \log p(\mathcal{X},\mathcal{U})}{\partial u_{i_a k}^{(a)} \partial u_{i_a k}^{(a)}}\right\} = \frac{N_{i_a}^{(a,c)} \rho_k^{(c)}}{\rho^{(a,c)}} + \frac{1}{\rho_k^{(a)}},$$

- Matrix Co-factorization

$$\mathbb{E}_{X,U}\left\{-\frac{\partial^2 \log p(\mathcal{X},\mathcal{U})}{\partial u_{k i_a}^{(a)} \partial u_{k i_a}^{(a)}}\right\} = \sum_{c|(a,c)\in\mathcal{R}} \frac{N_{i_a}^{(a,c)} \rho_k^{(c)}}{\rho^{(a,c)}} + \frac{1}{\rho_k^{(a)}},$$

where $N^{(a,c)} = \left|\mathcal{O}^{(a,c)}\right|$ and $N_{i_a}^{(a,c)} = \left|\left\{ i_a \,|\, \mathcal{O}^{(a,c)} \right\}\right|$.

Reconstruction Error: BCRB

We evaluate a lower bound on the reconstruction error using BCRB.

$$
\begin{aligned}
\mathcal{E}_{ij} &= \mathbb{E}\left\{(\widehat{x}_{ij} - x_{ij})^2\right\} \\
&= \mathbb{E}\left\{(\overline{\mathbf{u}}_i^\top \overline{\mathbf{v}}_j - \mathbf{u}_i^\top \mathbf{v}_j)^2\right\} \\
&\geq \mathbf{v}_j^\top \left[\mathcal{I}^{-1}\right]_{u_i} \mathbf{v}_j + \mathrm{tr}\left(\left[\mathcal{I}^{-1}\right]_{u_i} \left[\mathcal{I}^{-1}\right]_{v_j}\right) + \mathbf{u}_i^\top \left[\mathcal{I}^{-1}\right]_{v_j} \mathbf{u}_i .
\end{aligned}
$$

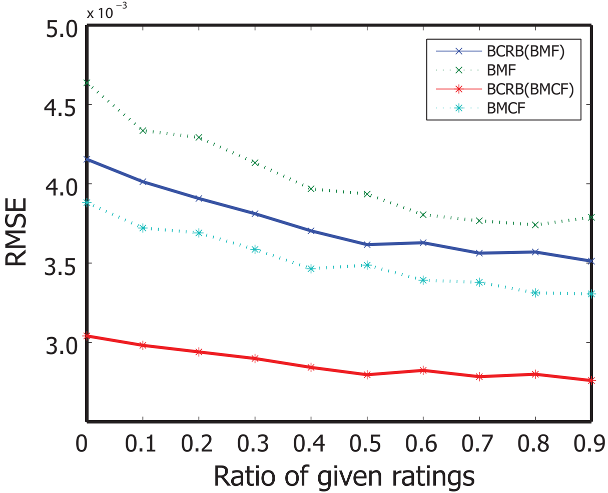
Numerical Experiments

- Experiment 1: BCRB Comparison
  - $\mathcal{E} = \{1, 2, 3, 4\}$
  - $\mathcal{R} = \{(1, 2), (2, 3), (3, 4)\}$
  - $\mathbf{U}^{(a)} \in \mathbb{R}^{5 \times 100}$ and $[\mathbf{U}^{(a)}]_{ij} \sim \mathcal{N}(\mathbf{U}^{(a)} \,|\, 0, 1)$.
  - Ratio of observed entries: 0% $\sim$ 90%

- Experiment 2: Collaborative Prediction
  - MovieLens data: 943 users, 1682 movies
  - User information: age(5), gender(2), and occupation(21)
  - Movie information: genre(18)

BCRB Comparison on Synthetic Data

- BMCF had lower bound and performance compared to the BMF



Collaborative Prediction in the Cold-Start Situation

- BMCF performs better than BMF, especially in the cold-start situations

User Cold Start

	BMF		BMCF	
	MAE	RMSE	MAE	RMSE
0	2.5403	2.7767	0.8238	1.0140
5	0.8281	1.0618	0.7895	0.9941
10	0.8032	1.0205	0.7446	0.9424
15	0.7474	0.9558	0.7426	0.9314
20	0.7421	0.9496	0.7348	0.9254

User and Item Cold Start (200 items out of 1682 are missing)

	BMF		BMCF	
	MAE	RMSE	MAE	RMSE
0	2.5098	2.7584	0.8843	1.0857
5	0.9333	1.2412	0.8332	1.0550
10	0.8956	1.1863	0.7778	0.9857
15	0.8991	1.1948	0.7716	0.9789
20	0.8618	1.1535	0.7527	0.9555

Conclusions

- Matrix co-factorization provides a principled approach to systematically exploiting side information.
- We have presented a Bayesian matrix co-factorization (BMCF) where we used variational Bayesian inference for collaborative prediction.
- We have also provided Bayesian Cramér-Rao bound (BCRB) for both BMF and BMCF, emphasizing that BMCF indeed yielding the smaller Cramér-Rao bound.
- Numerical experiments confirmed the useful behavior of BMCF in the case of user/item cold start.