Bayesian Recommendation

Seungjin Choi
(joint work with Jiho Yoo, Sunho Park, and Yong-Deok Kim)

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
seungjin@postech.ac.kr
http://mlg.postech.ac.kr/∼seungjin

November 6, 2015

Acknowledgments: Machine Learning Lab (since 2001)

Currently 6 PhD students and 2 MS students

http://mlg.postech.ac.kr

Bayesian Recommendation?

- Recommendation

Bayesian Recommendation?

- Recommendation

- Collaborative prediction (wisdom of crowd)

	최승진	한준희	황승원	한보형	최재식
광해	★★★★★	★★★★★	★		
관상	★★★★		★	★★★	
아저씨	★★★	★★★	★★★		★★★★★
신세계	★★★★	★★★	★★★★	★★★★	
엽기적인 그녀	★★★	★★	★★★★★	★★	
겨울 왕국	★				★★★★

Bayesian Recommendation?
= Bayesian models for recommendation

- Recommendation

- Collaborative prediction (wisdom of crowd)

	최승진	한준희	황승원	한보형	최재식
광해	★★★★★	★★★★★	★		
관상	★★★★		★	★★★	
아저씨	★★★	★★★	★★★		★★★★★
신세계	★★★★	★★★	★★★★	★★★★	
엽기적인 그녀	★★★	★★	★★★★★	★★	
겨울 왕국	★				★★★★

Collaborative Filtering

	최승진	한준희	황승원	한보형	최재식
광해	★★★★★	★★★★★	★		
관상	★★★★		★	★★★	
아저씨	★★★	★★★	★★★		★★★★★
신세계	★★★★	★★★	★★★★	★★★★	
엽기적인 그녀	★★★	★★	★★★★★	★★	
겨울 왕국	★				★★★★

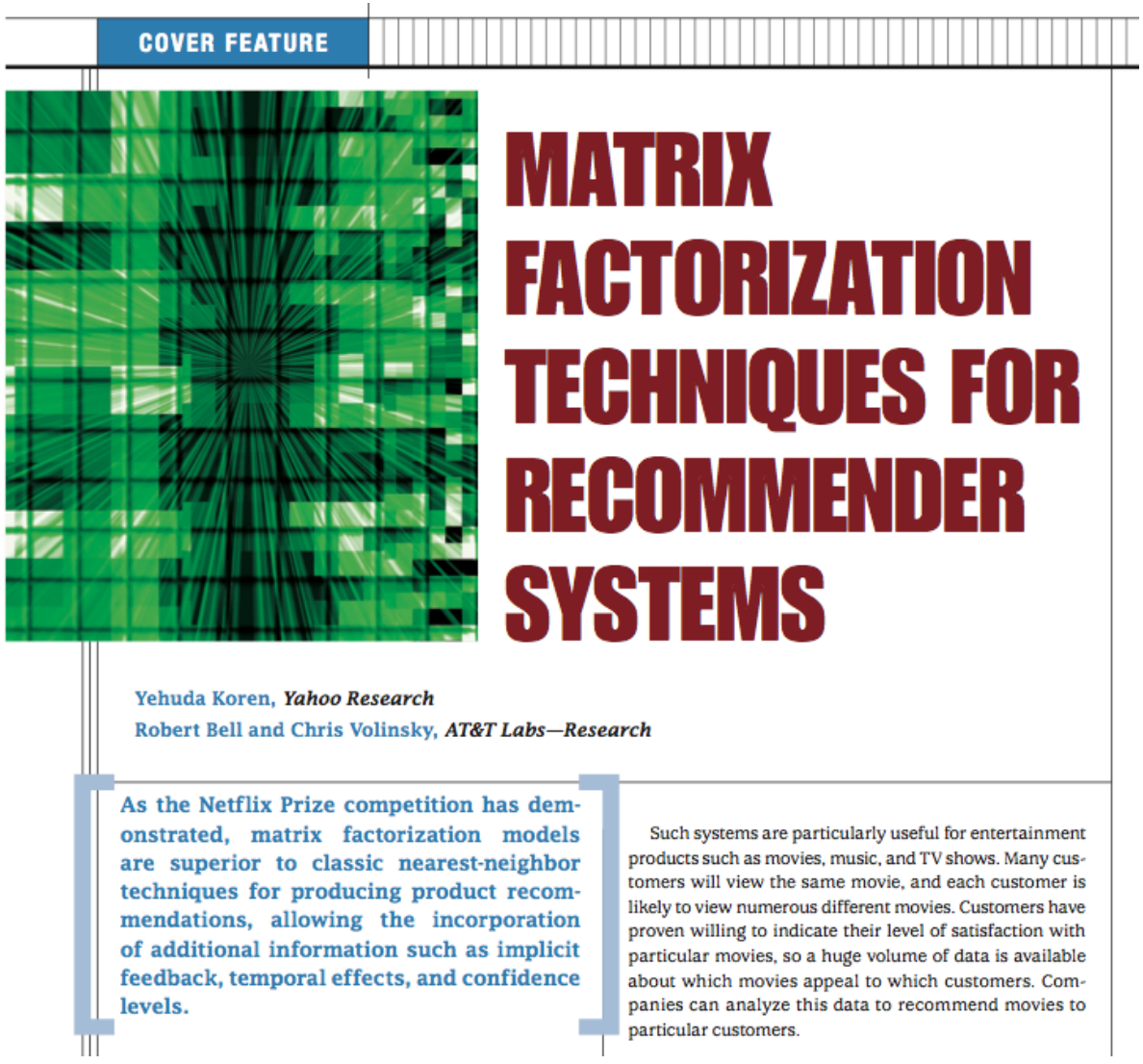
- Memory-based methods
  - User-based nearest neighbor (GroupLens 1994)
  - Item-based nearest neibhbor (Amazon 2001)
- Model-based methods
  - Matrix factorization
  - Clustering models
  - RBM
  - Bayesian models

Collaborative Filtering

	최승진	한준희	황승원	한보형	최재식
광해	★★★★★	★★★★★	★		
관상	★★★★		★	★★★	
아저씨	★★★	★★★	★★★		★★★★★
신세계	★★★★	★★★	★★★★	★★★★	
엽기적인 그녀	★★★	★★	★★★★★	★★	
겨울 왕국	★				★★★★

- Memory-based methods
  - User-based nearest neighbor (GroupLens 1994)
  - Item-based nearest neibhbor (Amazon 2001)
- Model-based methods
  - Matrix factorization
  - Clustering models
  - RBM
  - Bayesian models

Matrix Factorization for Collaborative Filtering (Y. Koren, 2009)



COVER FEATURE

MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

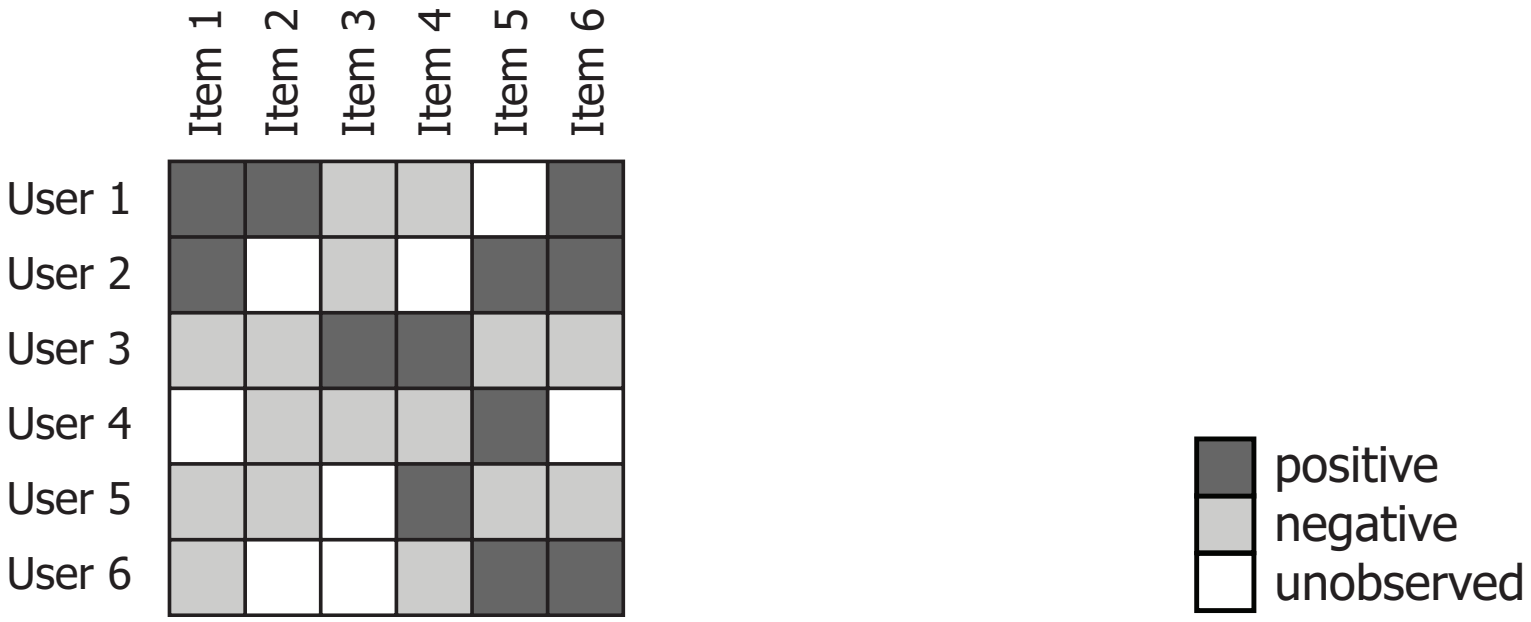
Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Research*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

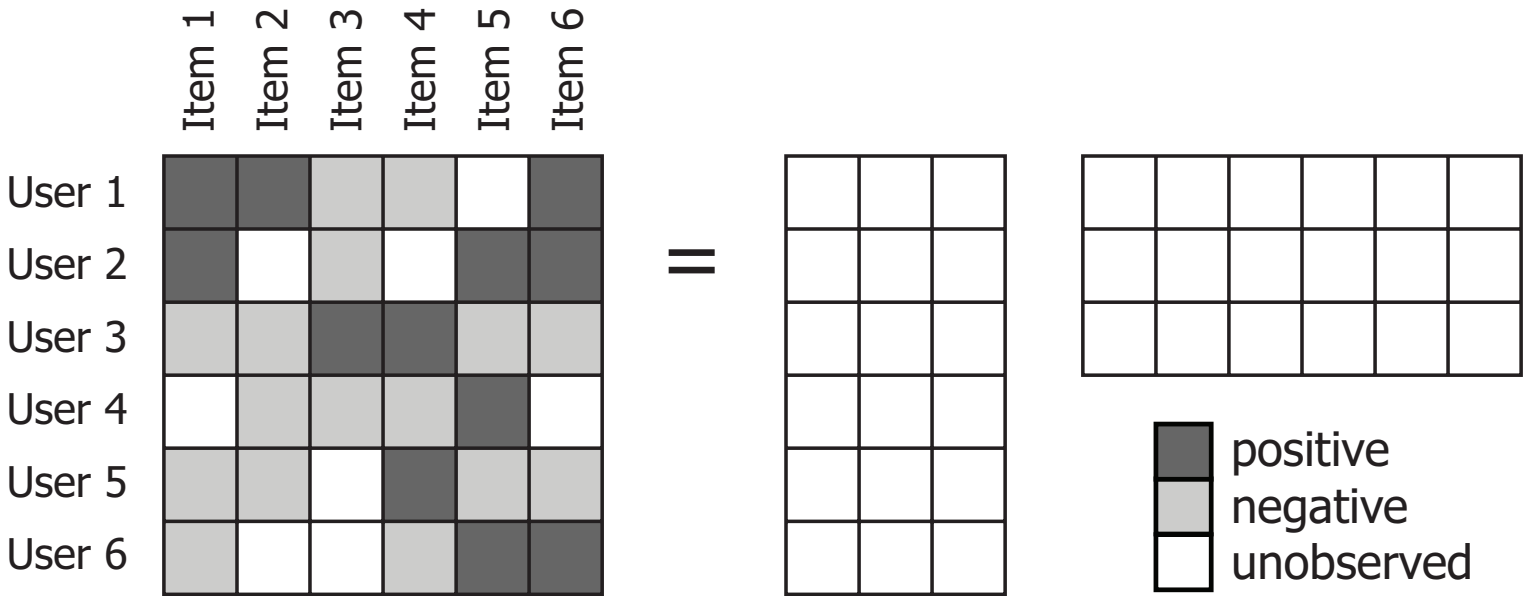
Such systems are particularly useful for entertainment products such as movies, music, and TV shows. Many customers will view the same movie, and each customer is likely to view numerous different movies. Customers have proven willing to indicate their level of satisfaction with particular movies, so a huge volume of data is available about which movies appeal to which customers. Companies can analyze this data to recommend movies to particular customers.

Popularity of matrix factorization were brought by Netflix prize

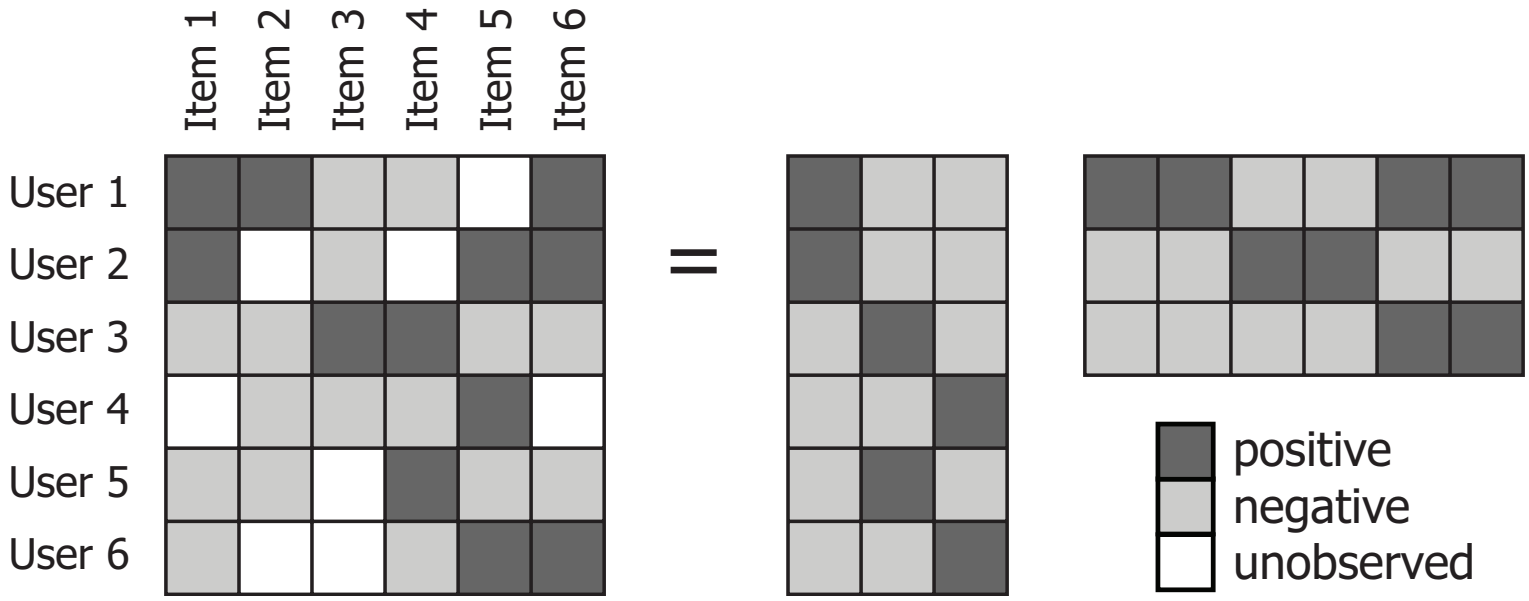
Matrix Factorization for Collaborative Prediction



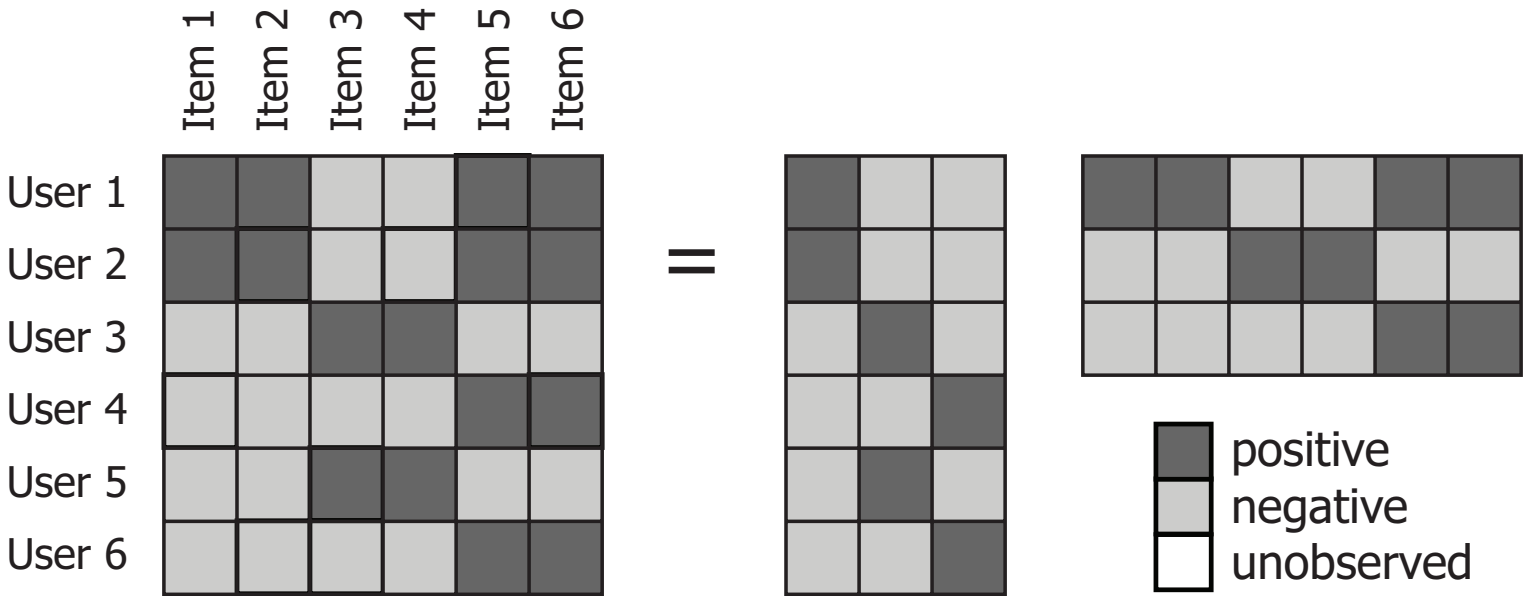
Matrix Factorization for Collaborative Prediction



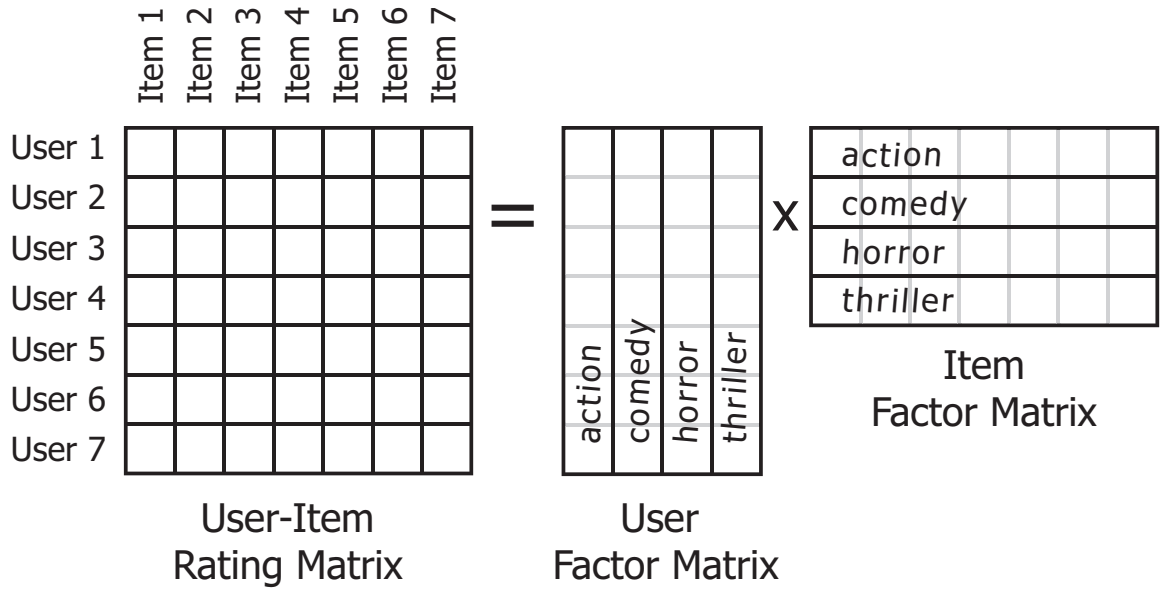
Matrix Factorization for Collaborative Prediction



Matrix Factorization for Collaborative Prediction

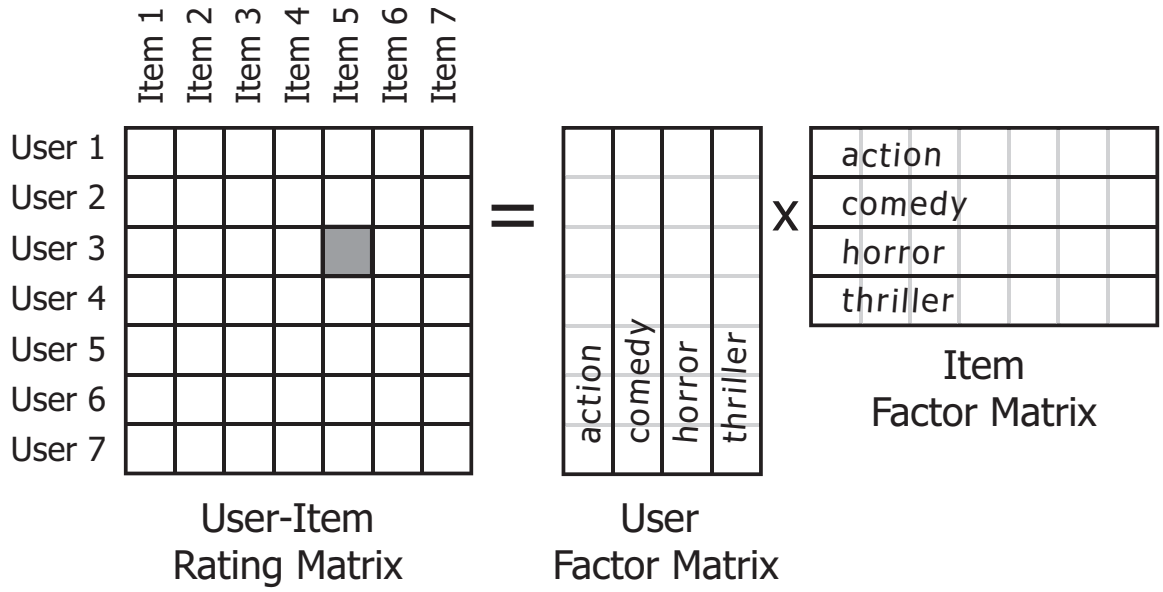


Matrix Factorization for Collaborative Prediction



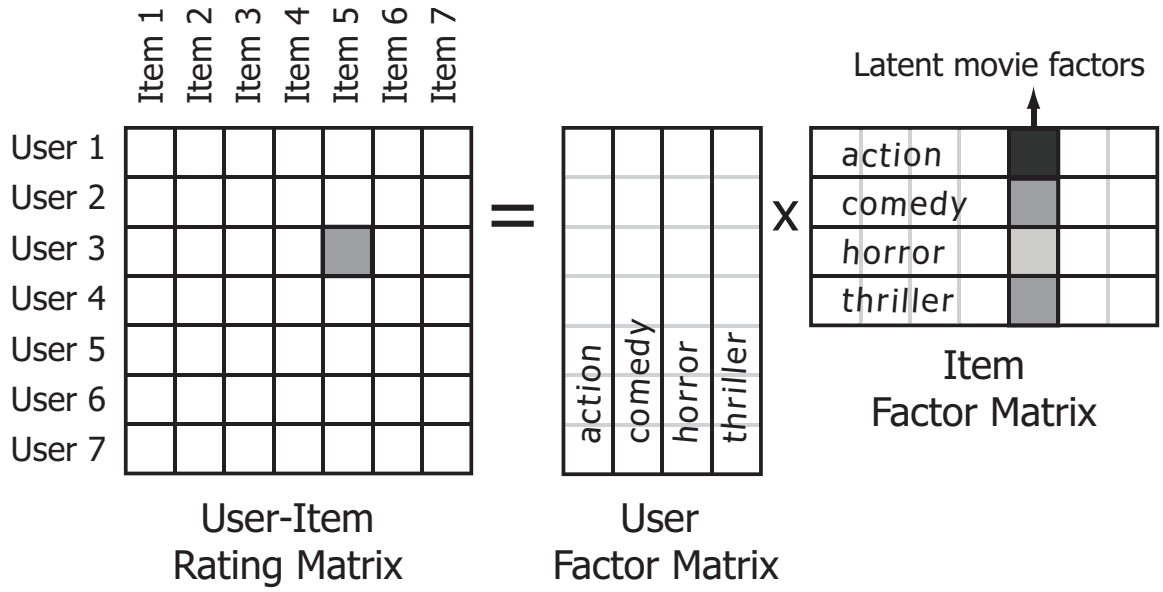
$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j$$

Matrix Factorization for Collaborative Prediction



$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j$$

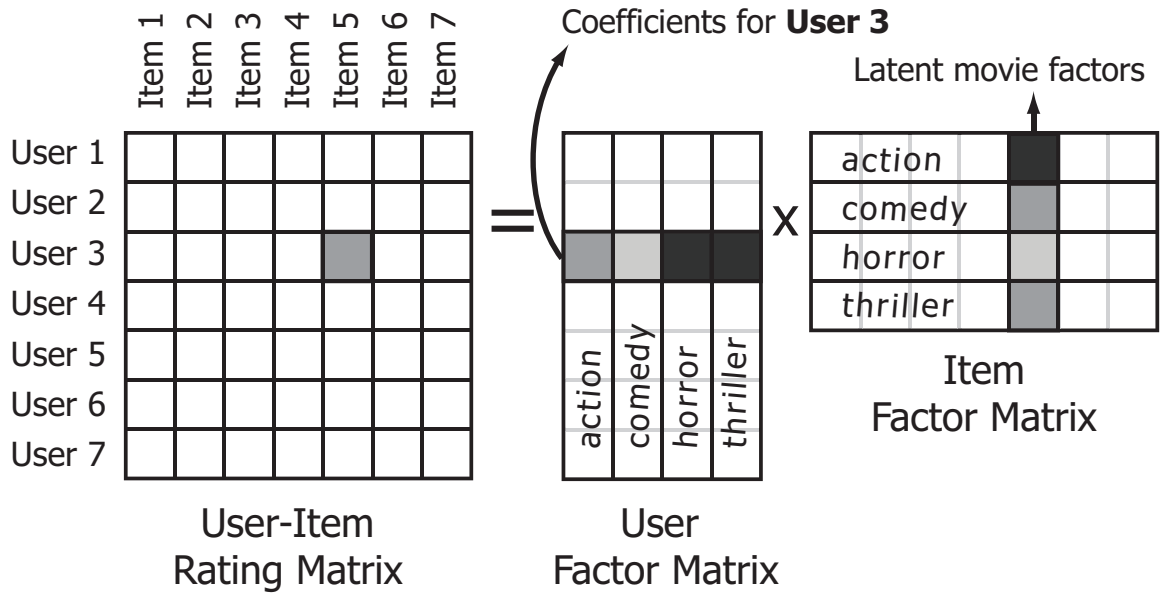
Matrix Factorization for Collaborative Prediction



$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j$$

$\mathbf{v}_j$: item-specific profile

Matrix Factorization for Collaborative Prediction



$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j$$

$\mathbf{u}_i$: user-specific profile $\quad$ $\mathbf{v}_j$: item-specific profile

- Algebraic methods to find a decomposition (for instance, SVD)
- Probabilistic models to infer $p(\mathbf{u}_i|\mathbf{X})$ and $p(\mathbf{v}_j|\mathbf{X})$

Matrix Factorization Methods

- Single data matrix
  - Singular value decomposition (SVD)
  - Nonnegative matrix factorization (NMF)
  - Probabilistic matrix factorization (PMF)
  - Bayesian matrix factorization (BMF)
- Multiple data matrices (rating matrix + side information)
  - Nonnegative matrix co-factorization
  - Bayesian matrix co-factorization
  - Bayesian matrix factorization with side information

Matrix Factorization Methods

- Single data matrix
  - Singular value decomposition (SVD)
  - Nonnegative matrix factorization (NMF)
  - Probabilistic matrix factorization (PMF)
  - Bayesian matrix factorization (BMF)
- Multiple data matrices (rating matrix + side information)
  - Nonnegative matrix co-factorization
  - Bayesian matrix co-factorization
  - Bayesian matrix factorization with side information

Matrix Factorization Methods

- Single data matrix
  - Singular value decomposition (SVD)
  - Nonnegative matrix factorization (NMF)
  - Probabilistic matrix factorization (PMF)
  - Bayesian matrix factorization (BMF)
- Multiple data matrices (rating matrix + side information)
  - Nonnegative matrix co-factorization
  - Bayesian matrix co-factorization
  - Bayesian matrix factorization with side information

Bayesian matrix factorization + variational inference $\Rightarrow$

variational Bayesian matrix factorization

In this talk ...

1. Over-fitting:
2. Scalability:
3. Cold-start problems:
4. Nonrandom missing data:

In this talk ...

1. Over-fitting:
   Variational Bayesian view of weighted trace norm regularization

2. Scalability:
   Scalable variational Bayesian matrix factorization

3. Cold-start problems:
   Bayesian matrix factorization with side information

4. Nonrandom missing data:
   Bayesian binomial mixture models

Regularized Matrix Factorization

- Goal: Given ratings X_{ij} for $(i,j) \in \Omega$, determine $\mathbf{u}_i$ and $\mathbf{v}_j$ such that

$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j.$$

- How: Solve

$$\arg\min_{\mathbf{u}_i, \mathbf{v}_j} \frac{1}{2} \underbrace{\left\{ \sum_{(i,j)\in\Omega} \|X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j\|^2 \right\}}_{\text{LS fit}} + \frac{\lambda}{2} \underbrace{\left\{ \sum_i \|\mathbf{u}_i\|^2 + \sum_j \|\mathbf{v}_j\|^2 \right\}}_{\text{Regularization}}.$$

Regularized Matrix Factorization

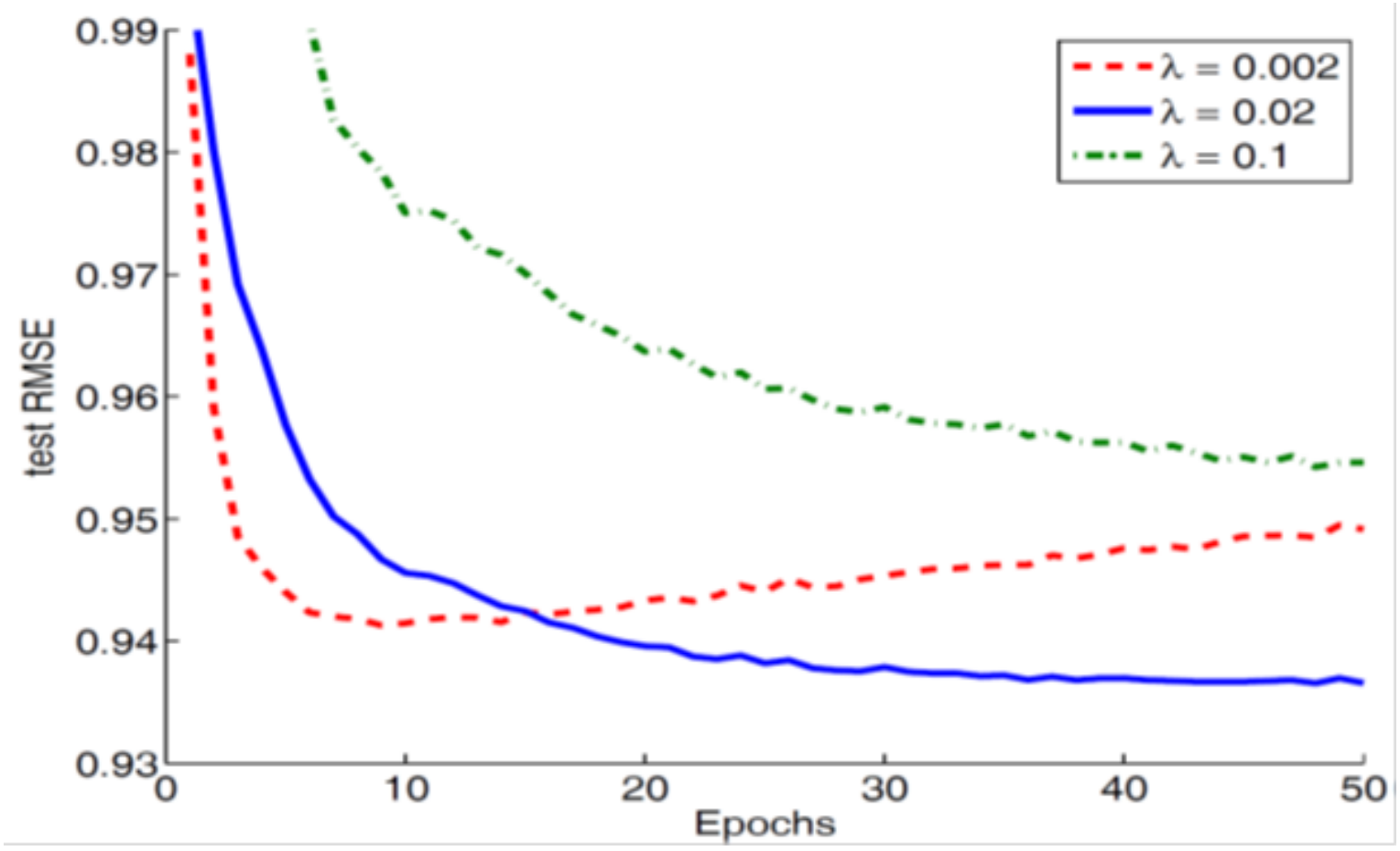
- Goal: Given ratings X_{ij} for $(i,j) \in \Omega$, determine $\mathbf{u}_i$ and $\mathbf{v}_j$ such that

$$X_{ij} \approx \mathbf{u}_i^\top \mathbf{v}_j.$$

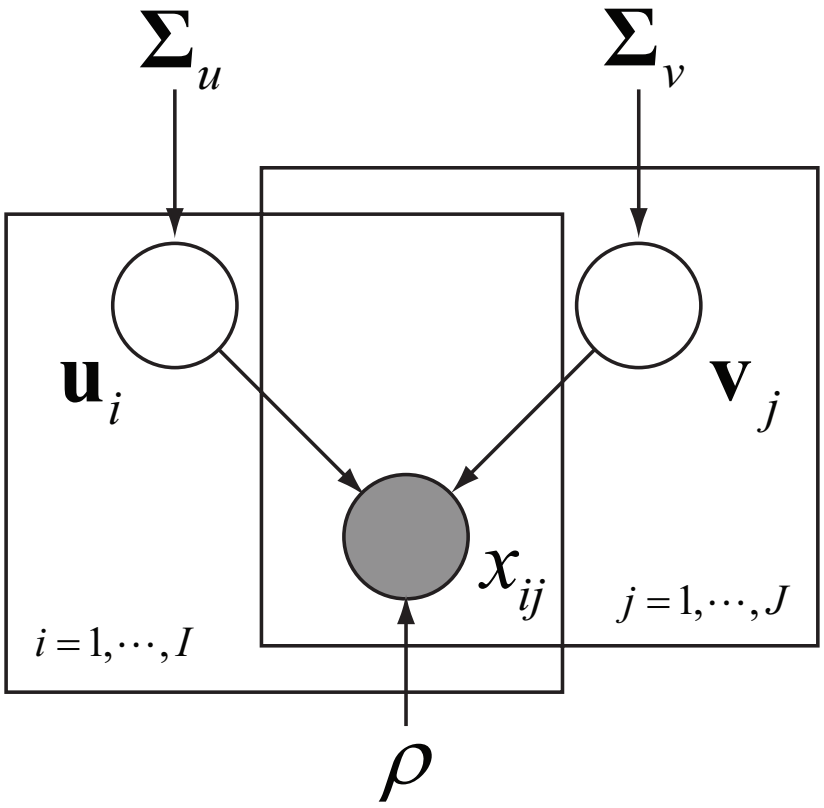
- How: Solve

$$\arg\min_{\mathbf{u}_i, \mathbf{v}_j} \frac{1}{2} \underbrace{\left\{ \sum_{(i,j)\in\Omega} \|X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j\|^2 \right\}}_{\text{LS fit}} + \frac{\lambda}{2} \underbrace{\left\{ \sum_i \|\mathbf{u}_i\|^2 + \sum_j \|\mathbf{v}_j\|^2 \right\}}_{\text{Regularization}}.$$

Regularization: Overfit and Underfit



Probabilistic Matrix Factorization



Salakhutdinov and Mnih, 2007

- Gaussian Likelihood

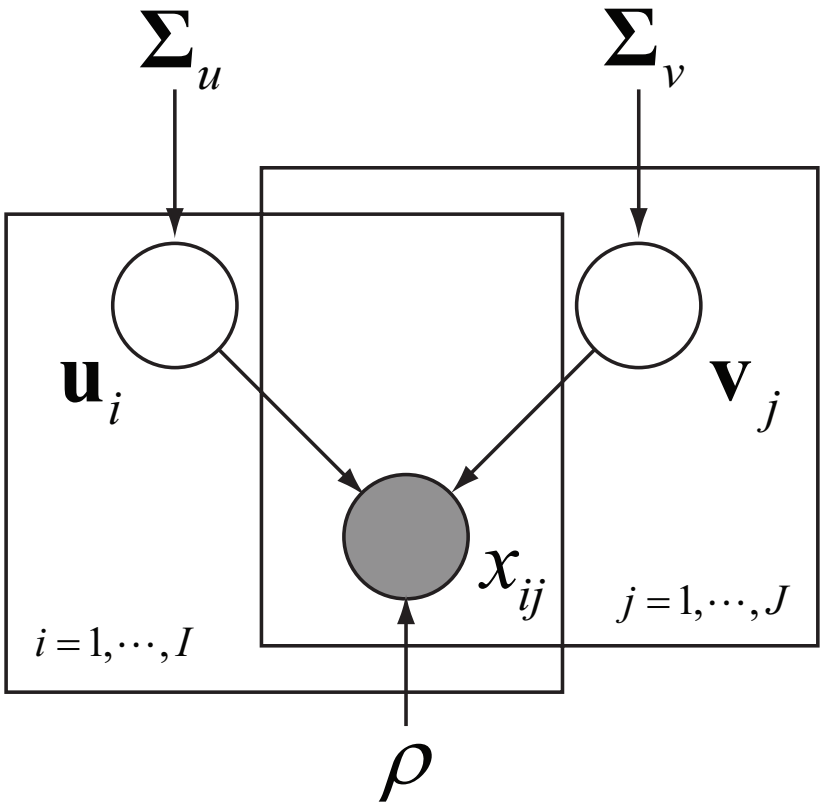
$$p(X_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(X_{ij}|\mathbf{u}_i^\top \mathbf{v}_j, \rho).$$

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Probabilistic Matrix Factorization



Salakhutdinov and Mnih, 2007

- Gaussian Likelihood

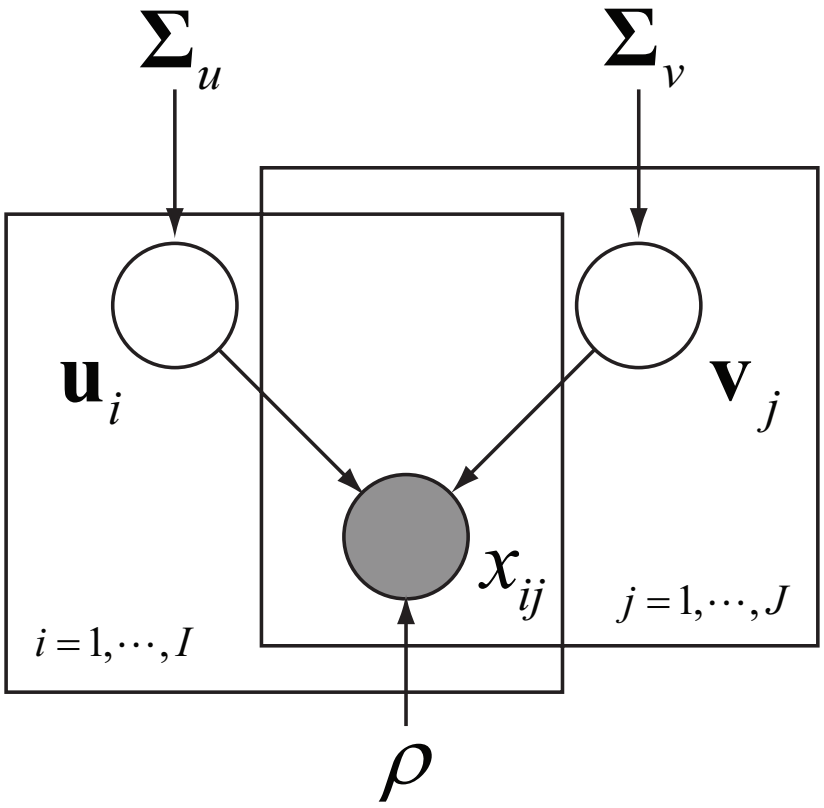
$$p(X_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(X_{ij}|\mathbf{u}_i^\top \mathbf{v}_j, \rho).$$

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Probabilistic Matrix Factorization



- Gaussian Likelihood

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

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

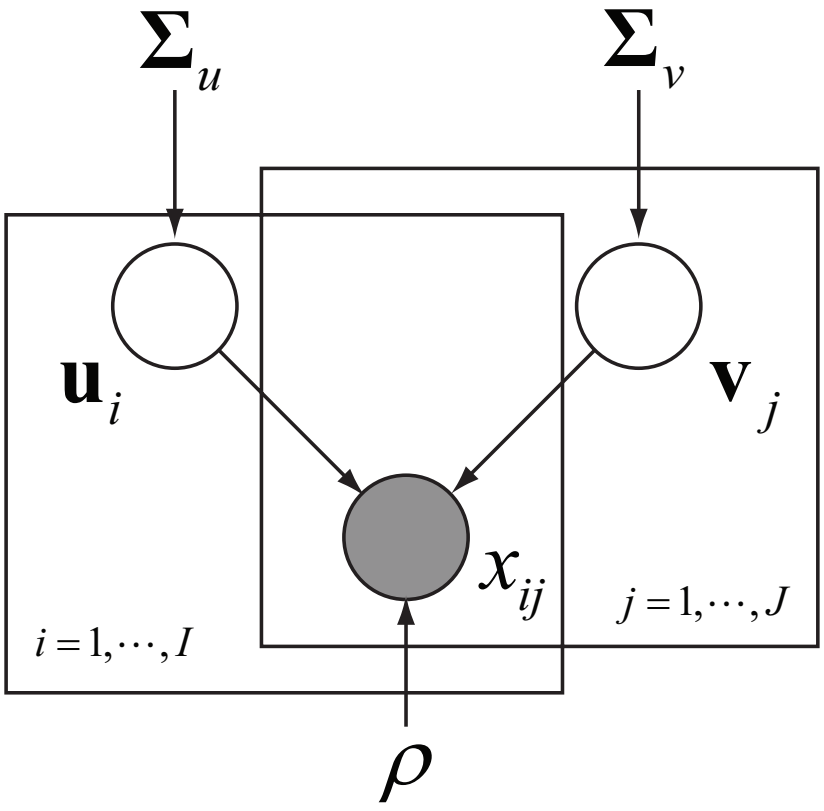
$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Salakhutdinov and Mnih, 2007

MAP = regularized matrix factorization

$$\min \frac{1}{2} \sum_{(i,j)\in\Omega} \|X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j\|^2 + \frac{\lambda_u}{2} \sum_i \|\mathbf{u}_i\|^2 + \frac{\lambda_v}{2} \sum_j \|\mathbf{v}_j\|^2.$$

Variational Bayesian Matrix Factorization



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

- Gaussian Likelihood

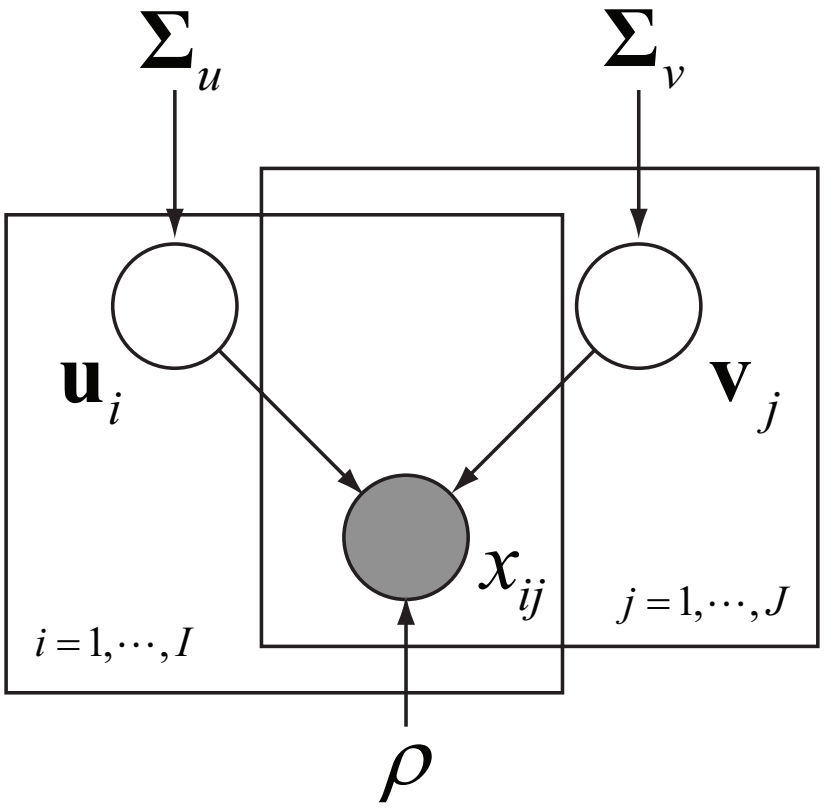
$$p(X_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(X_{ij}|\mathbf{u}_i^\top \mathbf{v}_j, \rho).$$

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Variational Bayesian Matrix Factorization



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

- Gaussian Likelihood

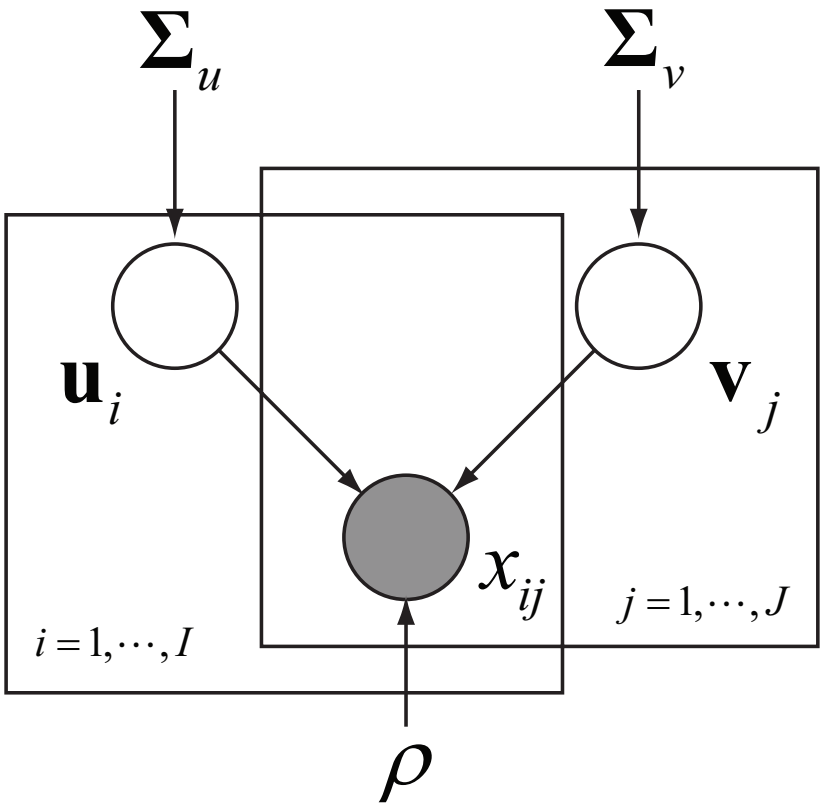
$$p(X_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(X_{ij}|\mathbf{u}_i^\top \mathbf{v}_j, \rho).$$

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Variational Bayesian Matrix Factorization



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

- Gaussian Likelihood

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

- Gaussian Priors ($\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}$ and $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}$)

$$p(\mathbf{U}) = \sum_{i=1}^{I} \mathcal{N}(\mathbf{u}_i|0, \mathbf{\Sigma}_u),$$

$$p(\mathbf{V}) = \sum_{j=1}^{J} \mathcal{N}(\mathbf{v}_j|0, \mathbf{\Sigma}_v).$$

Compute predictive prediction to infer missing entries $X_{i^*j^*}$ (hold-out case)

$$p(X_{i^*j^*}|\mathbf{X}) = \int p(X_{i^*j^*}|\mathbf{U}, \mathbf{V}) q_u^*(\mathbf{U}) q_v^*(\mathbf{V}) d\mathbf{U} d\mathbf{V} = \mathcal{N}(X_{i^*j^*}|\langle \mathbf{u}_{i^*} \rangle^\top \langle \mathbf{v}_{j^*} \rangle, \rho^{(x)}).$$

Matrix Factorization with Trace Norm Regularization

Given a rating matrix $\mathbf{X}$, find its low-rank approximation $\mathbf{Y} = \mathbf{U}^\top \mathbf{V}$, leading to

$$\arg\min_{Y} \sum_{(i,j)\in\Omega} \left[(X_{i,j} - Y_{i,j})^2 \right] + \lambda \|\mathbf{Y}\|_*.$$

Matrix Factorization with Trace Norm Regularization

Given a rating matrix $\mathbf{X}$, find its low-rank approximation $\mathbf{Y} = \mathbf{U}^\top \mathbf{V}$, leading to

$$\arg\min_{Y} \sum_{(i,j)\in\Omega} \left[(X_{i,j} - Y_{i,j})^2 \right] + \lambda \|\mathbf{Y}\|_*.$$

The trace norm $\|\mathbf{Y}\|_* = \mathrm{tr}\left(\sqrt{\mathbf{Y}\mathbf{Y}^\top}\right)$ is also written as

$$\|\mathbf{Y}\|_* = \min_{Y = U^\top V} \frac{1}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right),$$

leading to

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right] + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right).$$

Weighted Trace Norm Regularization

- Trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right] + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right).$$

- Weighted trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right) \right].$$

Weighted Trace Norm Regularization

- Trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right] + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right).$$

- Weighted trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right) \right].$$

Weighted Trace Norm Regularization

- Trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right] + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right).$$

- Weighted trace norm regularization

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 + \lambda \left(\sum_{i=1}^{I} \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} \|\mathbf{v}_j\|^2 \right) \right].$$

Thus, weighted trace norm regularization yields:

$$\arg\min_{U,V} \sum_{(i,j)\in\Omega} \left[(X_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right] + \lambda \left(\sum_{i=1}^{I} N_i^r \|\mathbf{u}_i\|^2 + \sum_{j=1}^{J} N_j^c \|\mathbf{v}_j\|^2 \right).$$

Variational Inference

Consider the marginal likelihood

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

where $\mathcal{F}(q)$ is referred to as variational lower-bound.
Mean-filed approximation leads to

$$q(\mathbf{U}, \mathbf{V}) = q_u(\mathbf{U}) q_v(\mathbf{V}).$$

Variational Inference

Consider the marginal likelihood

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

where $\mathcal{F}(q)$ is referred to as variational lower-bound.
Mean-filed approximation leads to

$$q(\mathbf{U}, \mathbf{V}) = q_u(\mathbf{U}) q_v(\mathbf{V}).$$

Then we solve the following optimization to compute variational posterior distributions q_u and q_v:

$$\arg\max_{q_u, q_v} \mathcal{F}(q_u, q_v) = \int \int q_u(\mathbf{U}) q_v(\mathbf{V}) \log \frac{p(\mathbf{X}, \mathbf{U}, \mathbf{V})}{q_u(\mathbf{U}) q_v(\mathbf{V})} d\mathbf{U} d\mathbf{V}.$$

Variational Bayesian View (Kim and Choi, IEEE SPL 2013)

Assume the likelihood that is of the form

$$p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j, \tau^{-1}),$$

with isotropic Gaussian priors over parameters

$$p(\mathbf{U}) = \prod_{i=1}^{I} \mathcal{N}(\mathbf{u}_i \mid 0, \alpha^{-1}), \qquad p(\mathbf{V}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{v}_j \mid 0, \alpha^{-1}).$$

Variational Bayesian View (Kim and Choi, IEEE SPL 2013)

Assume the likelihood that is of the form

$$p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j, \tau^{-1}),$$

with isotropic Gaussian priors over parameters

$$p(\mathbf{U}) = \prod_{i=1}^{I} \mathcal{N}(\mathbf{u}_i \mid 0, \alpha^{-1}), \qquad p(\mathbf{V}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{v}_j \mid 0, \alpha^{-1}).$$

Assume variational distributions that are of the form

$$q(\mathbf{U}, \mathbf{V}) = \prod_{i=1}^{I} \mathcal{N}(\mathbf{u}_i \mid \overline{\mathbf{u}}_i, \beta^{-1}) \prod_{j=1}^{J} \mathcal{N}(\mathbf{v}_j \mid \overline{\mathbf{v}}_j, \beta^{-1}).$$

Then, the maximization of variational lower-bound $\mathcal{F}_q$ yields

$$\arg\min \mathcal{J}(\overline{\mathbf{U}}, \overline{\mathbf{V}}) = \sum_{(i,j)\in\Omega} \left[(X_{ij} - \overline{\mathbf{u}}_i^\top \overline{\mathbf{v}}_j)^2 \right] + \left(\frac{1}{\beta} W + \frac{\alpha}{\tau} T \right),$$

where

$$\begin{aligned} W &= \sum_{i=1}^{I} N_i^r \|\overline{\mathbf{u}}_i\|^2 + \sum_{j=1}^{J} N_j^c \|\overline{\mathbf{v}}_j\|^2, \\ T &= \|\overline{\mathbf{U}}\|_F^2 + \|\overline{\mathbf{V}}\|_F^2. \end{aligned}$$

When $\alpha \to 0$, we have

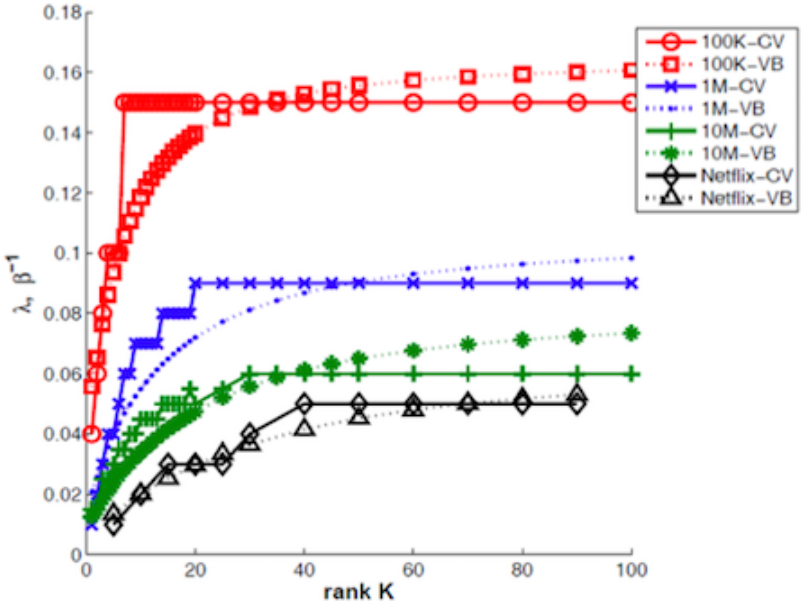
$$\arg\min \mathcal{J}(\overline{\mathbf{U}}, \overline{\mathbf{V}}) = \sum_{(i,j)\in\Omega} \left[(X_{ij} - \overline{\mathbf{u}}_i^\top \overline{\mathbf{v}}_j)^2 \right] + \frac{1}{\beta} W.$$

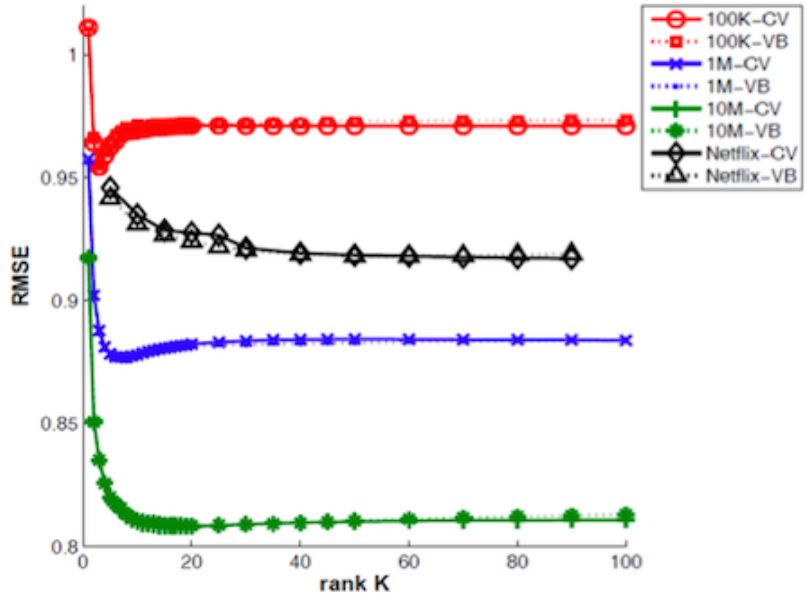
The optimal regularization parameter is learned via

$$\frac{1}{\beta} = \frac{-(\tau W + \alpha d) + \sqrt{(\tau W + \alpha d)^2 + 8NK\tau d}}{4NK\tau} \leq \sqrt{\frac{I+J}{2N\tau}},$$

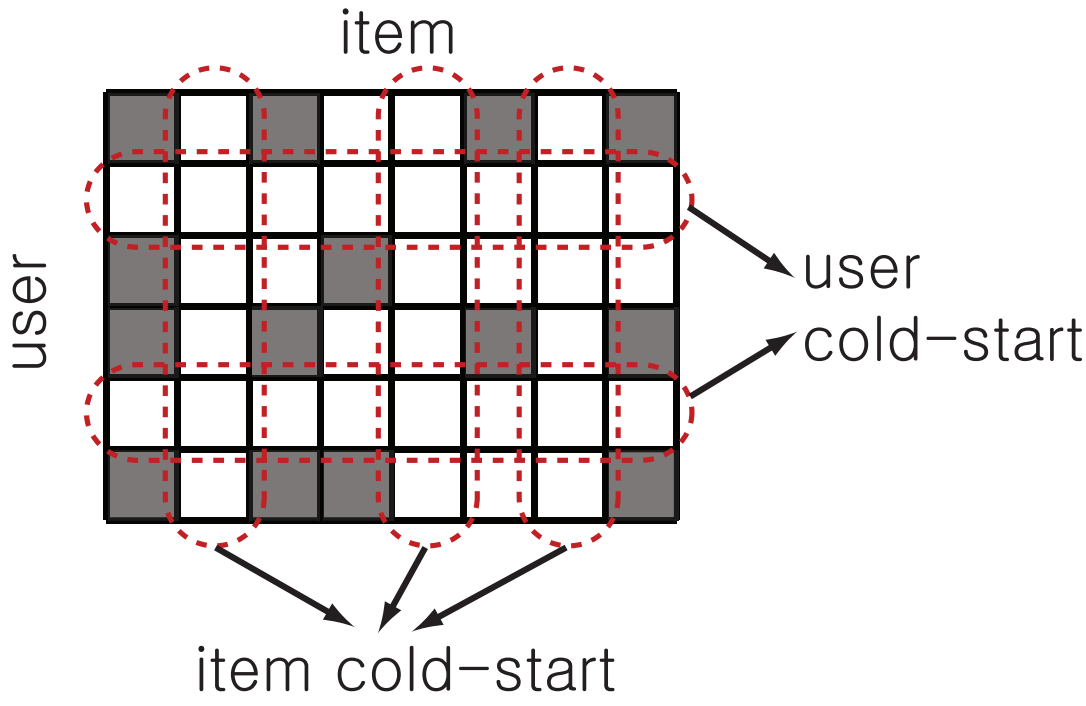
where $d = (I+J)K$.

VB and Regularization

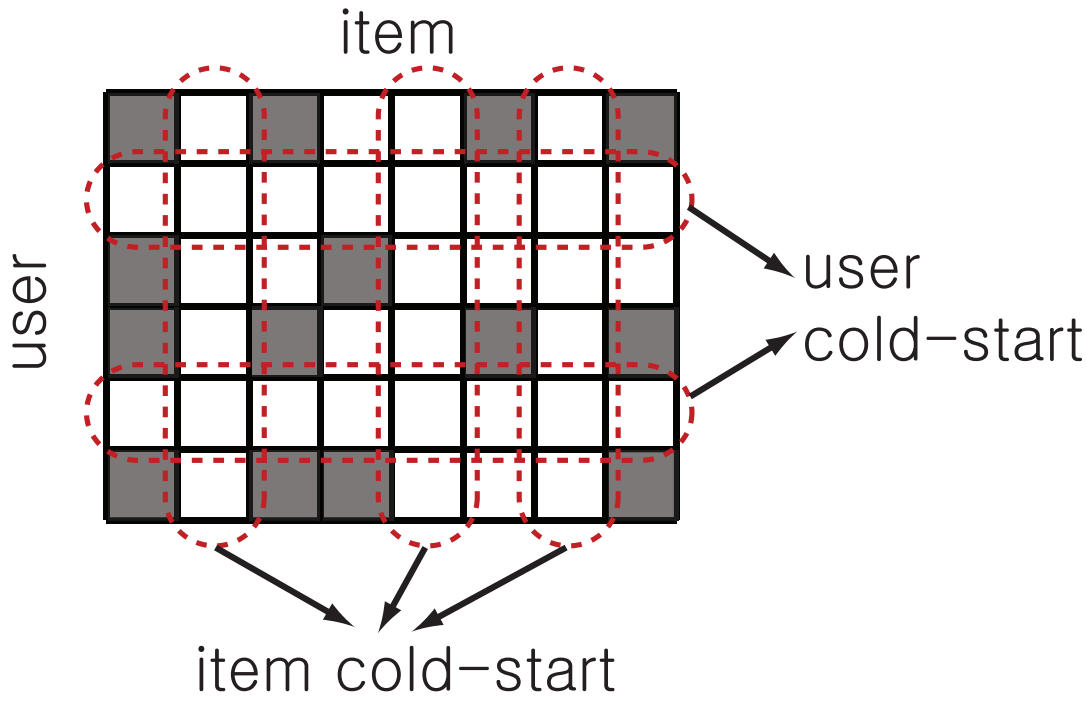




Cold Start Problems

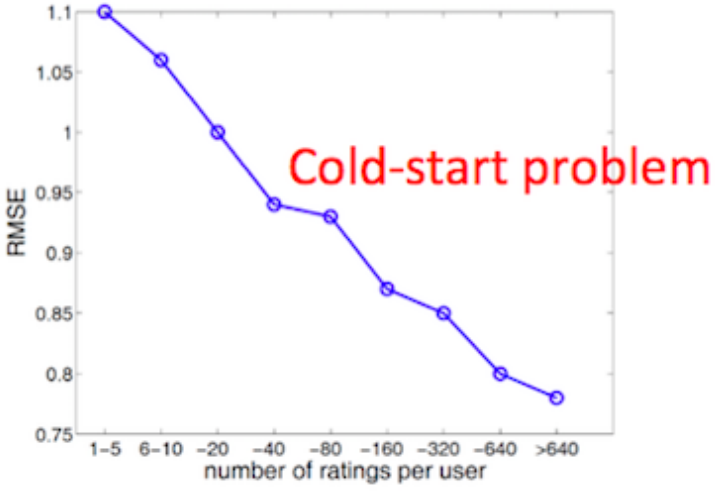


Cold Start Problems



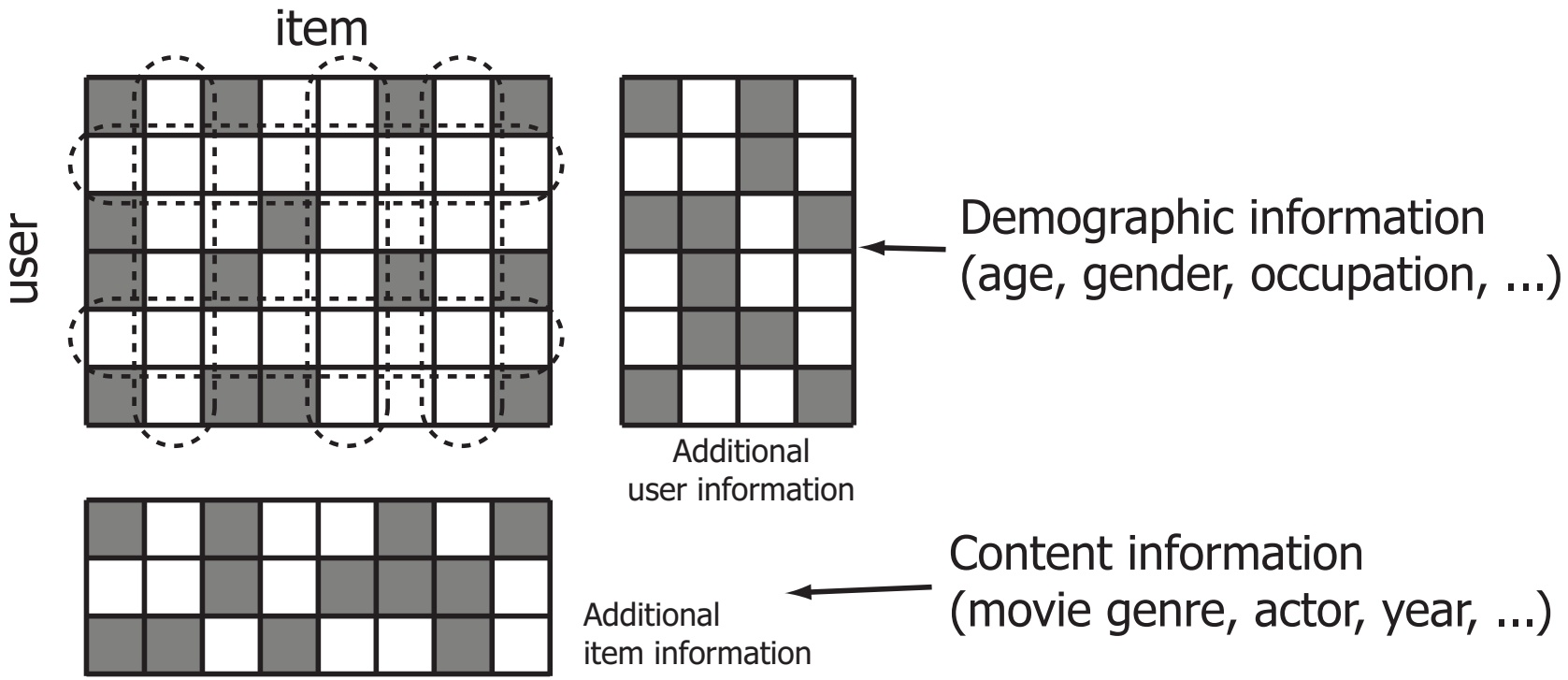
Cold start problems

- 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


Cold-start problem
RMSE
1.1
1.05
1
0.95
0.9
0.85
0.8
0.75
1–5
6–10
–20
–40
–80
–160
–320
–640
>640
number of ratings per user

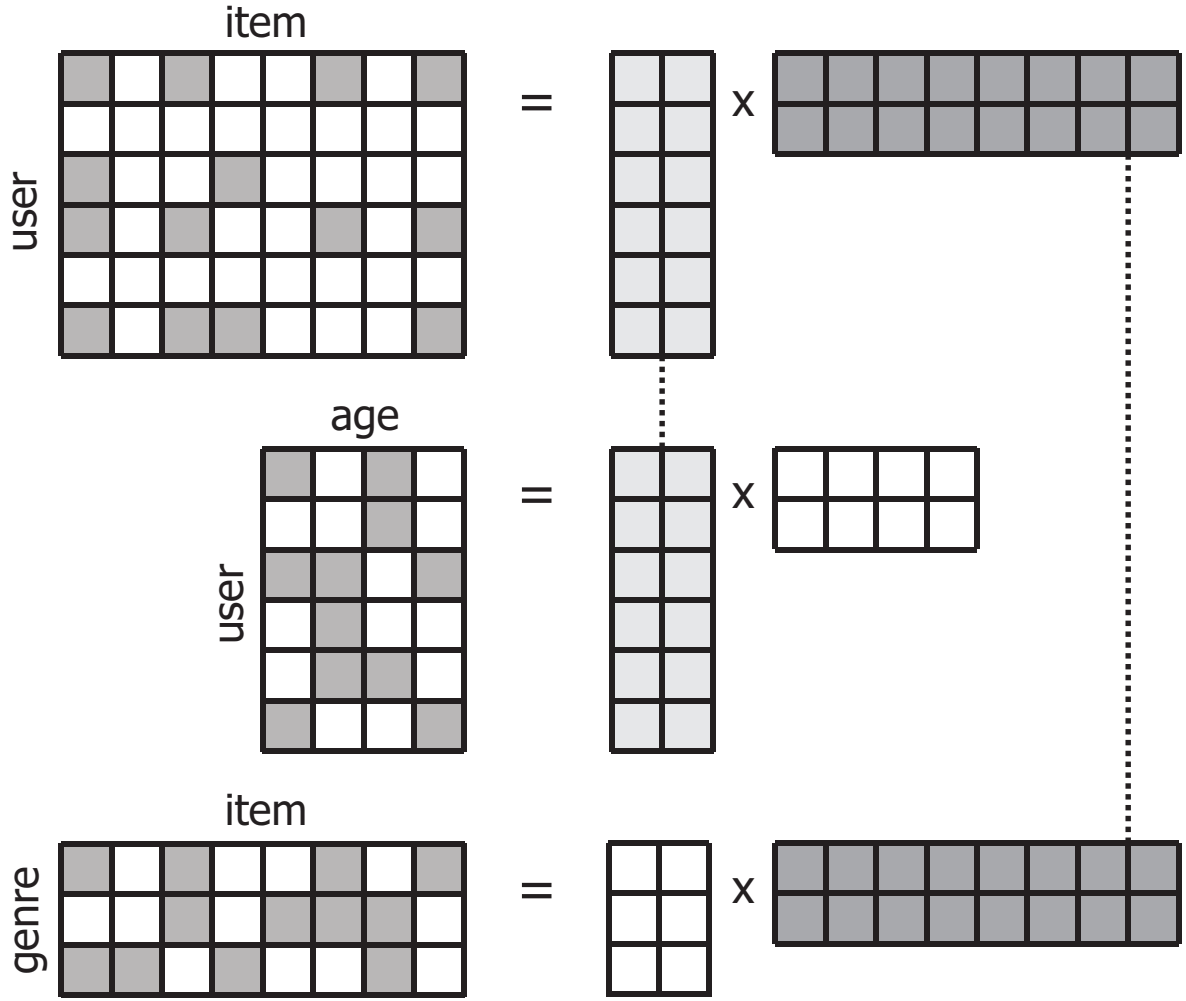

사용자 평점
SNS
영화
1
2
3
4
Fiction
배우
장르

Side Information

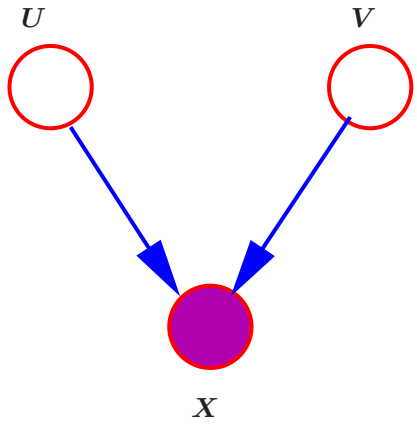


Matrix Co-Factorization (Yoo and Choi, 2009)

Input matrices are jointly decomposed, sharing some factor matrices.

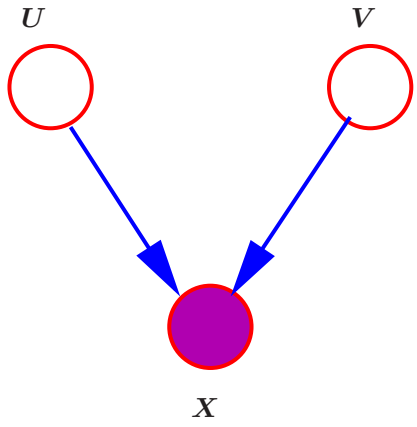


Nonnegative Matrix Co-Factorization

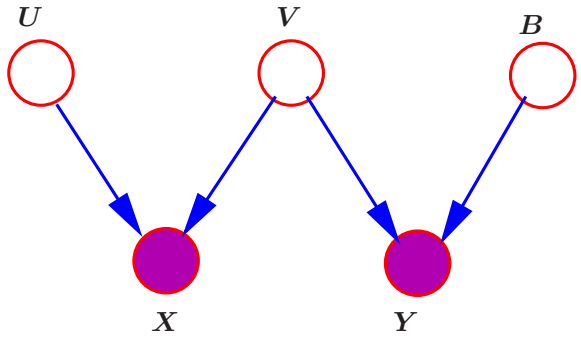


- $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$
- $\min \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|^2$

Nonnegative Matrix Co-Factorization

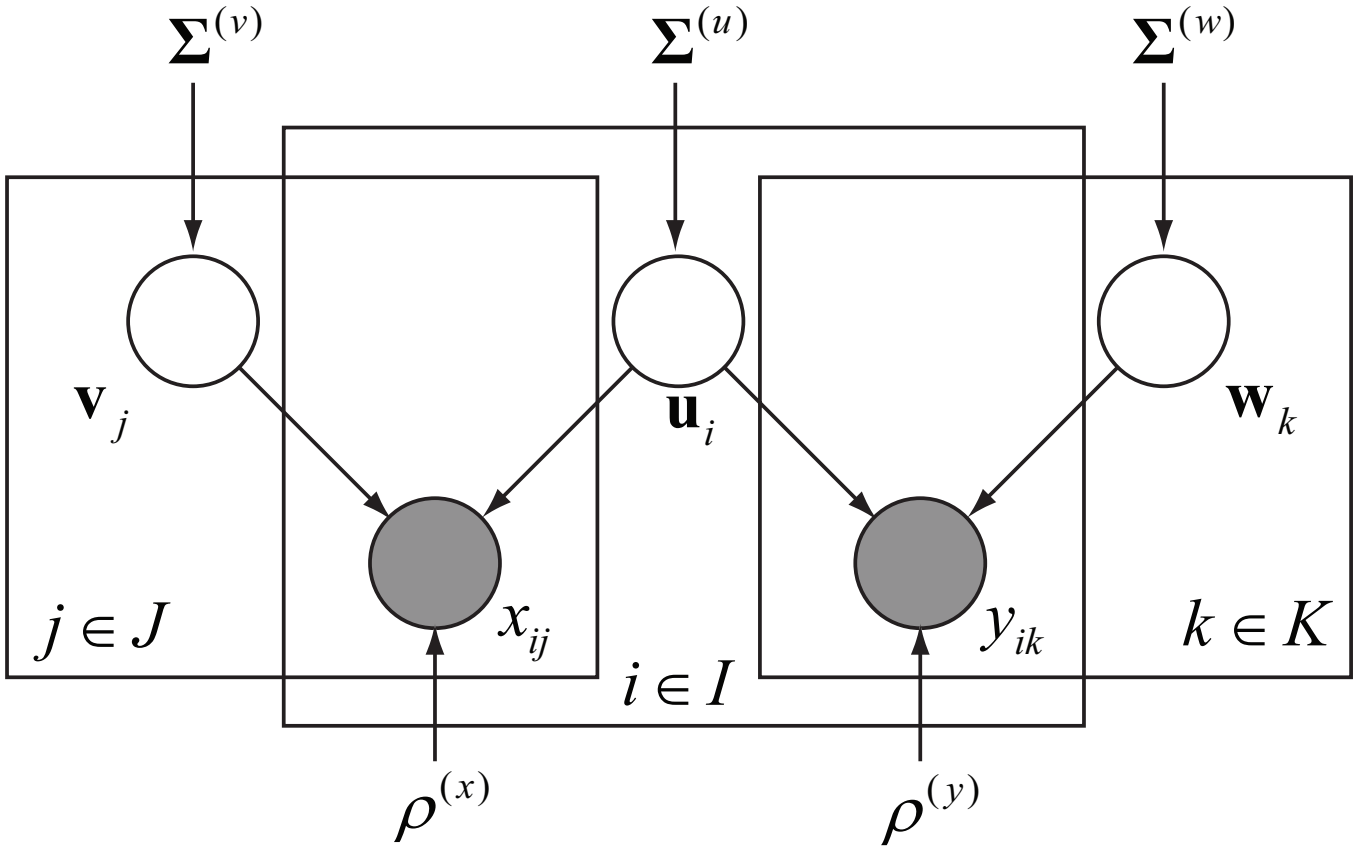


- $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$
- $\min \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|^2$

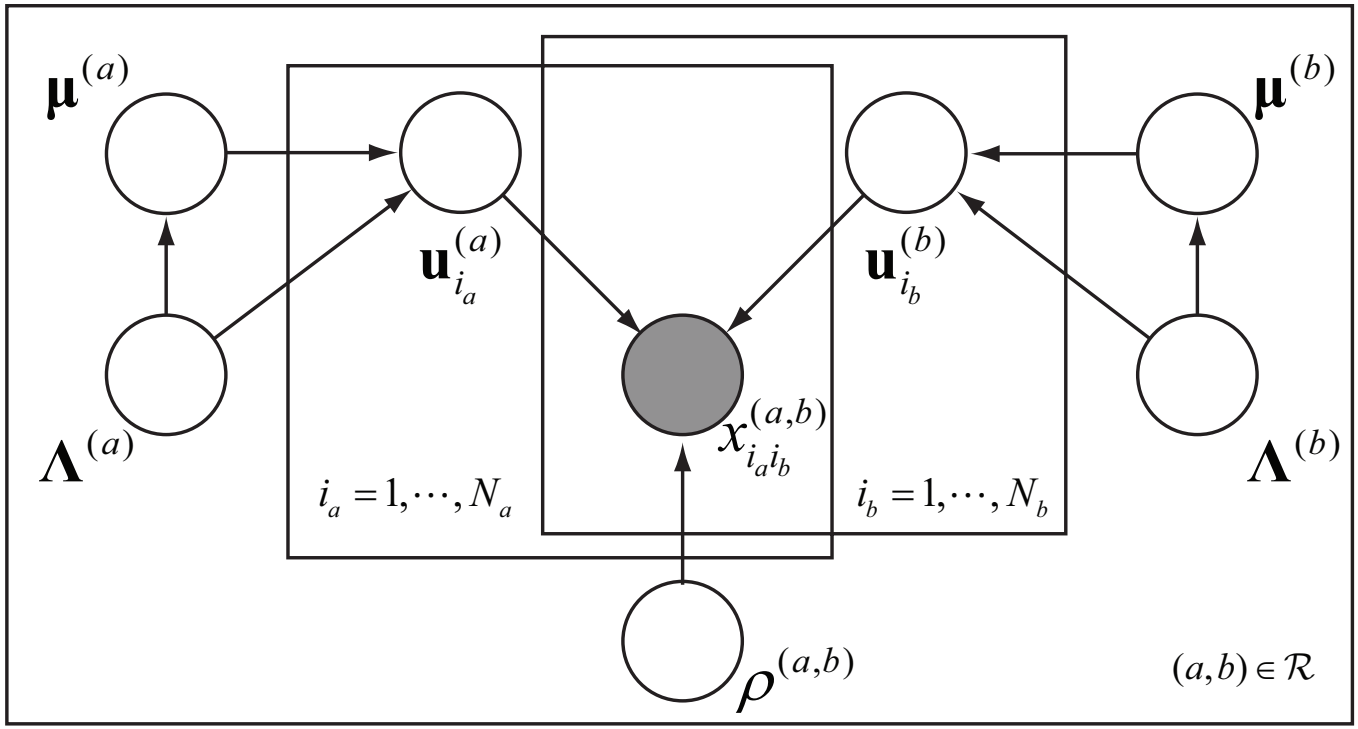


- $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$ and $\mathbf{Y} \approx \mathbf{B}\mathbf{V}^\top$
- $\min \beta\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|^2 + (1 - \beta)\|\mathbf{Y} - \mathbf{B}\mathbf{V}^\top\|^2$

Bayesian Matrix Co-Factorization (Yoo and Choi, ECML-2011)



Hierarchical Bayesian Matrix Co-Factorization (Yoo and Choi, ICASSP-2012)



Experiments: MovieLens Data

- Ratings: 5-star ratings of 943 users for the 1682 movies.
- Side information: user information (age, gender, and occupation) and movie information (genre). For instance, 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.
- User cold-start: Randomly selected 200 test users and took out most of their ratings, to remain s ratings for each user, for $s =, 5, 10, 15, 20$.

Performance Measures: MAE and RMSE

Mean absolute error (MAE) and root mean squared error (RMSE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|r_i - \widehat{r}_i|,$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \widehat{r}_i)^2},$$

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.

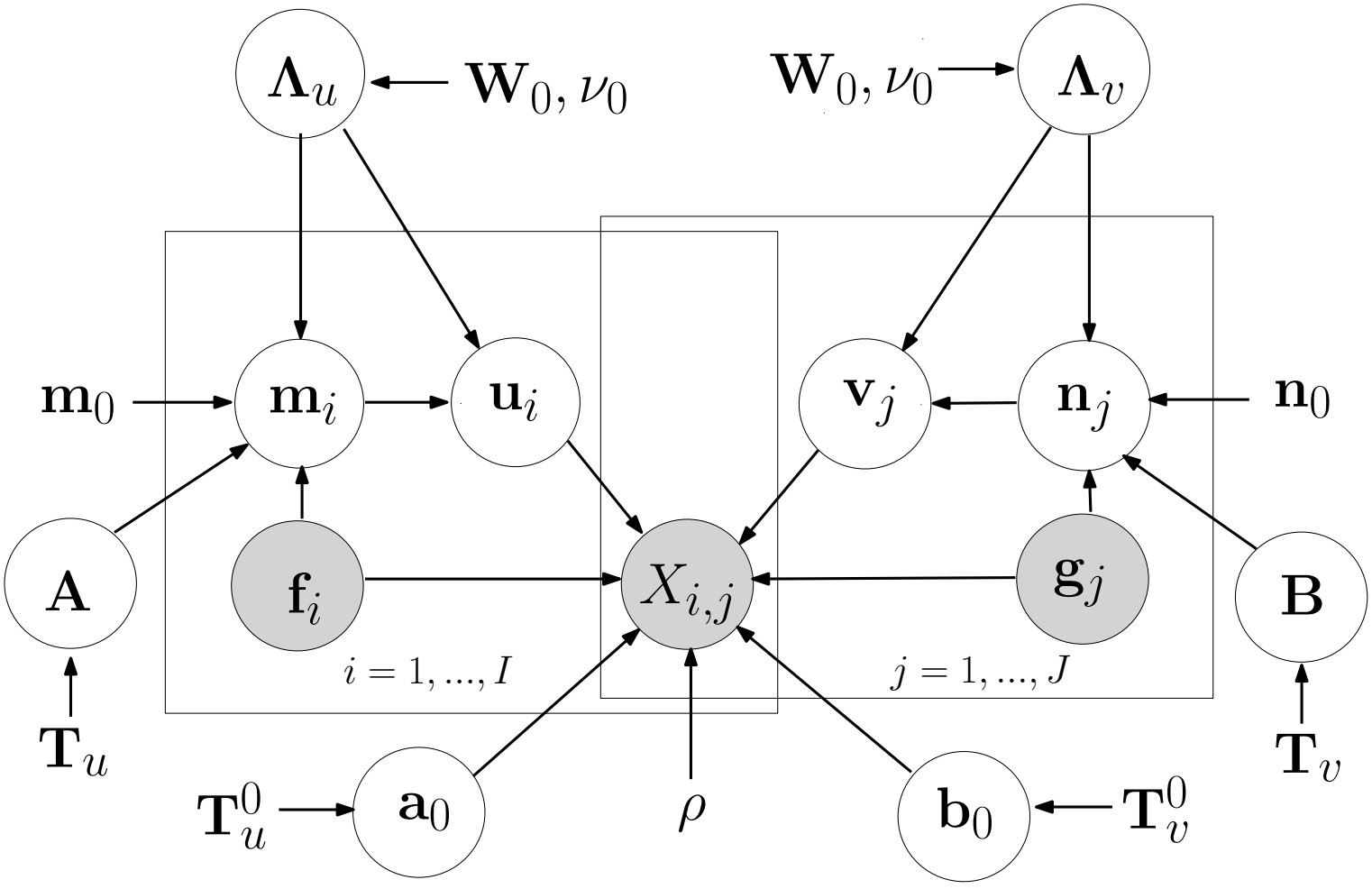
Experiments: MovieLens Data

Table : 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.

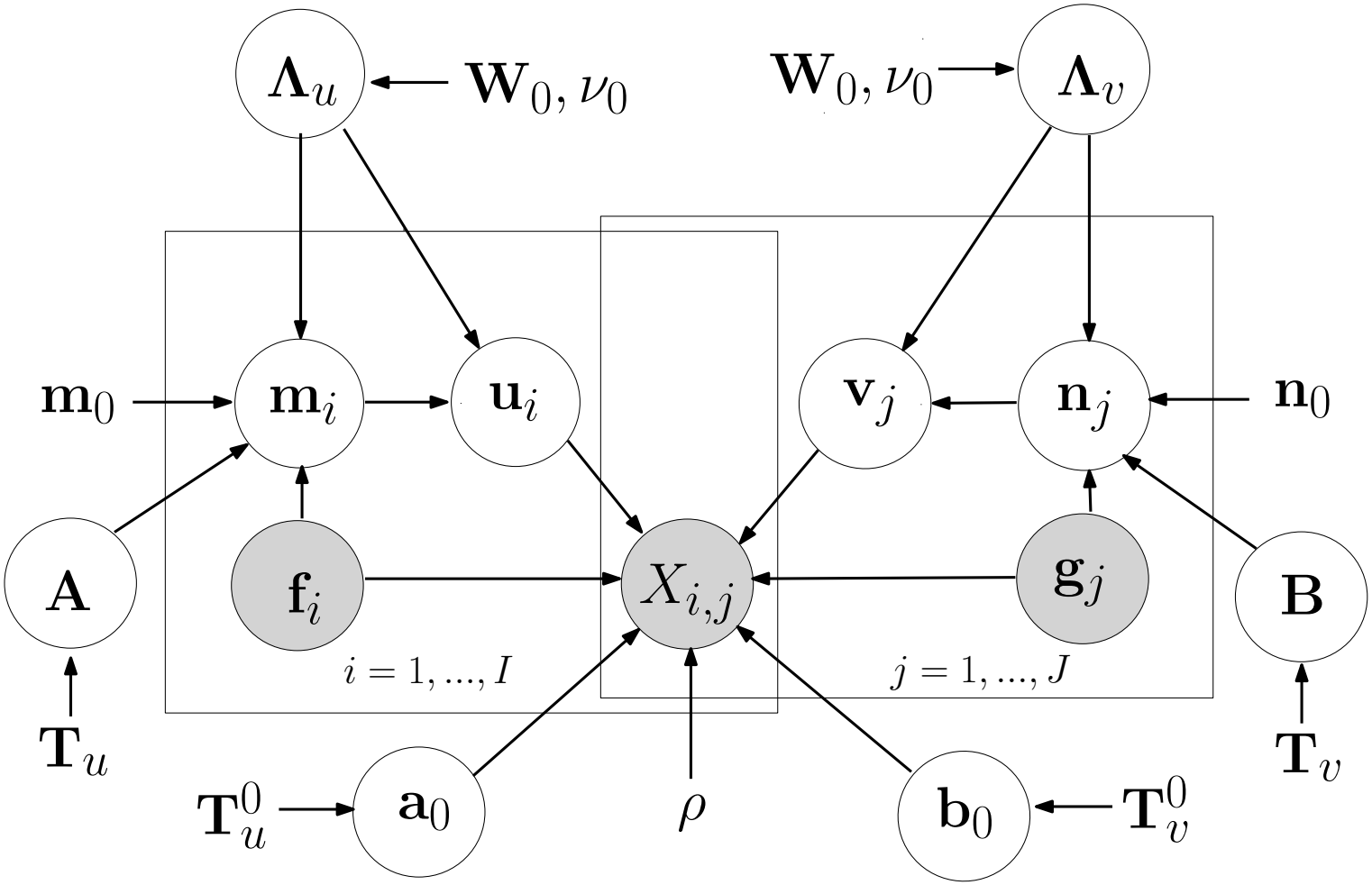
(a)	EVBMF		EVBMCF		HVBMCF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0	2.5403	2.7767	0.8238	**1.0140**	**0.8182**	**1.0140**
5	0.8281	1.0618	0.7895	**0.9941**	**0.7856**	0.9983
10	0.8032	1.0205	**0.7446**	**0.9424**	0.7485	0.9499
15	0.7474	0.9558	0.7426	0.9314	**0.7315**	**0.9288**
20	0.7421	0.9496	0.7348	**0.9254**	**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**

Hierarchical VBMF with Side information (Park, Kim, and Choi, IJCAI-2013)



Hierarchical VBMF with Side information (Park, Kim, and Choi, IJCAI-2013)



Likelihood:

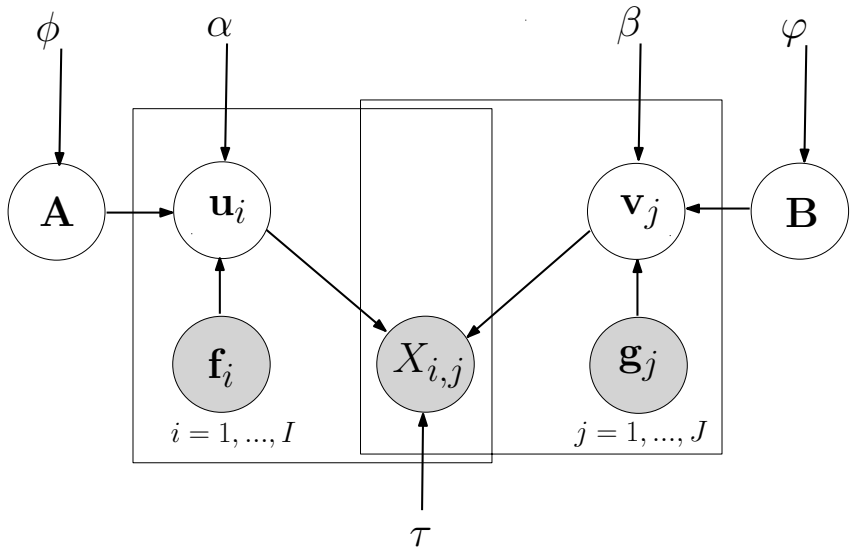
$$p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \mathbf{a}_0, \mathbf{v}_0, \rho) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j + \mathbf{a}_0^\top \mathbf{f}_i + \mathbf{b}_0^\top \mathbf{g}_j, \rho^{-1}).$$

Experiments

Table : 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)**

VBMF with Side Information: A Simpler Model



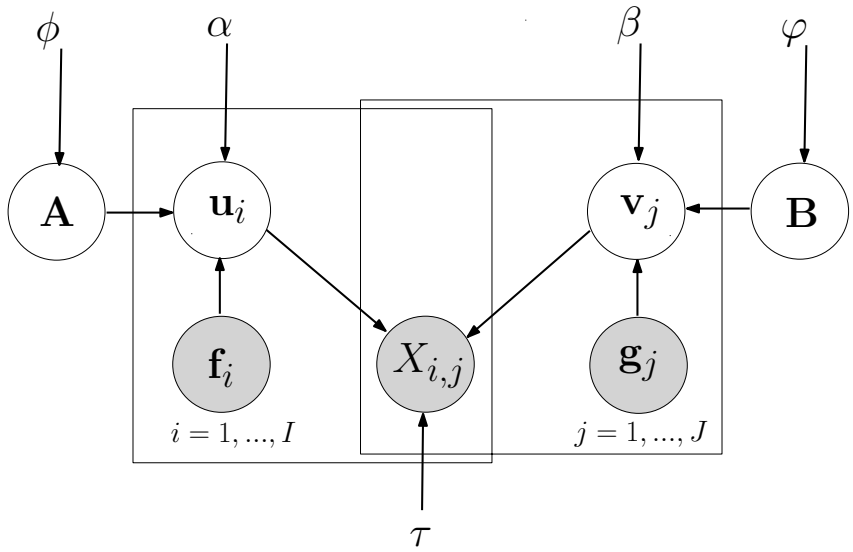
- Gaussian likelihood $p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j, \tau^{-1})$.
- Priors are

$$p(\mathbf{U}|\mathbf{A}, \mathbf{F}, \alpha) = \prod_{k=1}^{K} \prod_{i=1}^{I} \mathcal{N}(u_{ki} \mid \mathbf{a}_k^\top \mathbf{f}_i, \alpha_k^{-1}),$$

$$p(\mathbf{V}|\mathbf{B}, \mathbf{G}, \beta) = \prod_{k=1}^{K} \prod_{j=1}^{J} \mathcal{N}(v_{kj} \mid \mathbf{b}_k^\top \mathbf{g}_j, \beta_k^{-1}),$$

$$p(\mathbf{A}|\phi) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}(a_{mk} \mid 0, \phi_k^{-1}), \quad p(\mathbf{B}|\varphi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(b_{nk} \mid 0, \varphi_k^{-1})$$

VBMF with Side Information: A Simpler Model



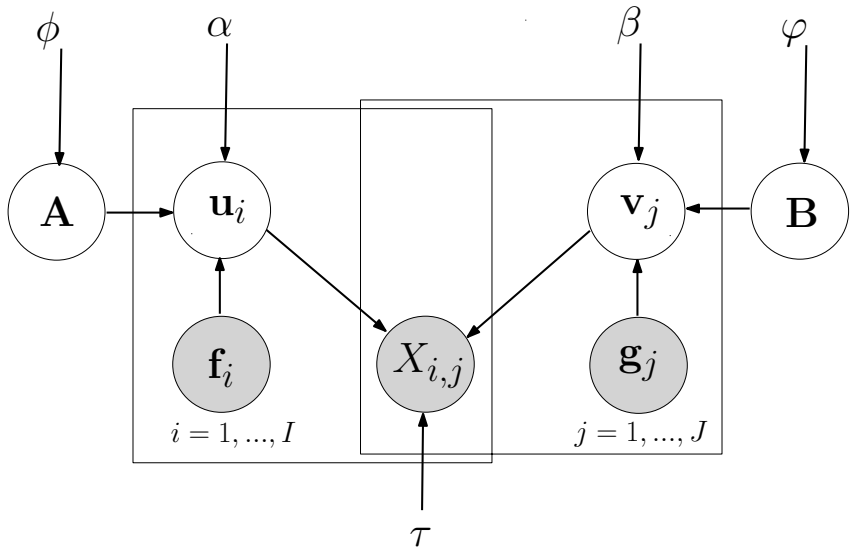
- Gaussian likelihood $p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j, \tau^{-1})$.
- Priors are

$$p(\mathbf{U}|\mathbf{A}, \mathbf{F}, \alpha) = \prod_{k=1}^{K} \prod_{i=1}^{I} \mathcal{N}(u_{ki} \mid \mathbf{a}_k^\top \mathbf{f}_i, \alpha_k^{-1}),$$

$$p(\mathbf{V}|\mathbf{B}, \mathbf{G}, \beta) = \prod_{k=1}^{K} \prod_{j=1}^{J} \mathcal{N}(v_{kj} \mid \mathbf{b}_k^\top \mathbf{g}_j, \beta_k^{-1}),$$

$$p(\mathbf{A}|\phi) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}(a_{mk} \mid 0, \phi_k^{-1}), \quad p(\mathbf{B}|\varphi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(b_{nk} \mid 0, \varphi_k^{-1})$$

VBMF with Side Information: A Simpler Model



- Gaussian likelihood $p(\mathbf{X} \mid \mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_i^\top \mathbf{v}_j, \tau^{-1})$.
- Priors are

$$p(\mathbf{U}|\mathbf{A}, \mathbf{F}, \boldsymbol{\alpha}) = \prod_{k=1}^{K} \prod_{i=1}^{I} \mathcal{N}(u_{ki} \mid \mathbf{a}_k^\top \mathbf{f}_i, \alpha_k^{-1}),$$

$$p(\mathbf{V}|\mathbf{B}, \mathbf{G}, \boldsymbol{\beta}) = \prod_{k=1}^{K} \prod_{j=1}^{J} \mathcal{N}(v_{kj} \mid \mathbf{b}_k^\top \mathbf{g}_j, \beta_k^{-1}),$$

$$p(\mathbf{A}|\boldsymbol{\phi}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}(a_{mk} \mid 0, \phi_k^{-1}), \quad p(\mathbf{B}|\boldsymbol{\varphi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(b_{nk} \mid 0, \varphi_k^{-1})$$

Complexities

Variational distributions	Updating rules	Space complexity	Time complexity
$q(\mathbf{U})$ $= \prod_{i=1}^{I} \mathcal{N}(\mathbf{u}_i \mid \bar{\mathbf{u}}_i, \mathbf{\Lambda}_i^u)$	$\mathbf{\Lambda}_i^u = \left(\text{diag}(\boldsymbol{\alpha}) + \tau \sum_{j \in \Omega_i} (\bar{\mathbf{v}}_j \bar{\mathbf{v}}_j^\top + \mathbf{\Lambda}_j^v)\right)^{-1}$, $\bar{\mathbf{u}}_i = \mathbf{\Lambda}_i^u \left(\text{diag}(\boldsymbol{\alpha}) \mathbf{A}^\top \mathbf{f}_i + \tau \sum_{j \in \Omega_i} X_{ij} \bar{\mathbf{v}}_j\right)$	$I(K + K^2)$	$\lvert\Omega\rvert K^2 + IK^3$
$q(\mathbf{V})$ $= \prod_{j=J}^{J} \mathcal{N}(\mathbf{v}_j \mid \bar{\mathbf{v}}_j, \mathbf{\Lambda}_j^v)$	$\mathbf{\Lambda}_j^v = \left(\text{diag}(\boldsymbol{\beta}) + \tau \sum_{i \in \Omega_j} (\bar{\mathbf{u}}_i \bar{\mathbf{u}}_i^\top + \mathbf{\Lambda}_i^u)\right)^{-1}$, $\bar{\mathbf{v}}_j = \mathbf{\Lambda}_j^v \left(\text{diag}(\boldsymbol{\beta}) \mathbf{B}^\top \mathbf{g}_i + \tau \sum_{i \in \Omega_j} X_{ij} \bar{\mathbf{u}}_i\right)$	$J(K + K^2)$	$\lvert\Omega\rvert K^2 + JK^3$
$q(\mathbf{A})$ $= \prod_{k=1}^{K} \mathcal{N}(\mathbf{a}_k \mid \bar{\mathbf{a}}_k, \mathbf{\Lambda}_k^a)$	$\mathbf{\Lambda}_k^a = \left(\phi_k \mathbf{I}_M + \alpha_k \mathbf{F}\mathbf{F}^\top\right)^{-1}$, $\bar{\mathbf{a}}_k = \mathbf{\Lambda}_k^a \left(\alpha_k \mathbf{F} \bar{U}_k^\top\right)$	$K(M + M^2)$	$K(IM + M^3)$
$q(\mathbf{B})$ $= \prod_{k=1}^{K} \mathcal{N}(\mathbf{b}_k \mid \bar{\mathbf{b}}_k, \mathbf{\Lambda}_k^a)$	$\mathbf{\Lambda}_k^b = \left(\varphi_k \mathbf{I}_N + \beta_k \mathbf{G}\mathbf{G}^\top\right)^{-1}$, $\bar{\mathbf{b}}_k = \mathbf{\Lambda}_k^b \left(\beta_k \mathbf{G} \bar{V}_k^\top\right)$	$K(N + N^2)$	$K(JN + N^3)$

cubic time complexity and quadratic space complexity

Scalable VBMF (Kim and Choi, AISTATS 2014)

- Assume element-wise independence in variational distributions:

$$q(\mathbf{U}, \mathbf{V}, \mathbf{A}, \mathbf{B}) = \prod_{k,i} q(u_{ki}) \prod_{k,j} q(v_{kj}) \prod_{m,k} q(a_{mk}) \prod_{n,k} q(b_{nk}).$$

- Define residuals $R_{ij} = X_{ij} - \sum_{k=1}^{K} \overline{u}_{ki}\overline{v}_{kj}$. When $\overline{u}_{ki}$ is changed to $\overline{u}'_{ki}$, R_{ij} can be easily updated to R'_{ij} for all $j \in \Omega_i$ by:

$$R'_{ij} = R_{ij} - (\overline{u}'_{ki} - \overline{u}_{ki})\overline{v}_{kj}.$$

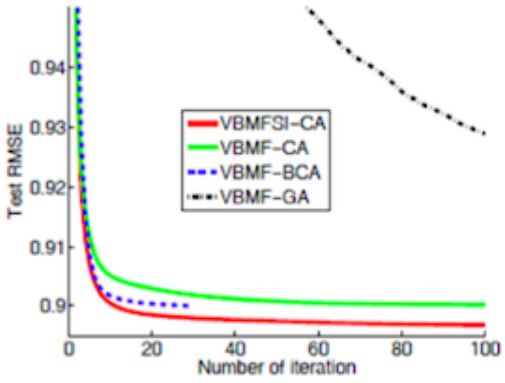
- With these assumptions and tricks, time complexity reduces to linear w.r.t. K.

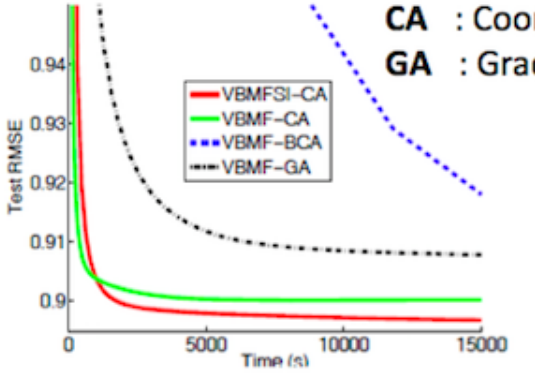
Experiments

Numerical experiments

Netflix prize, $K = 50$

Side information: binary implicit feed back



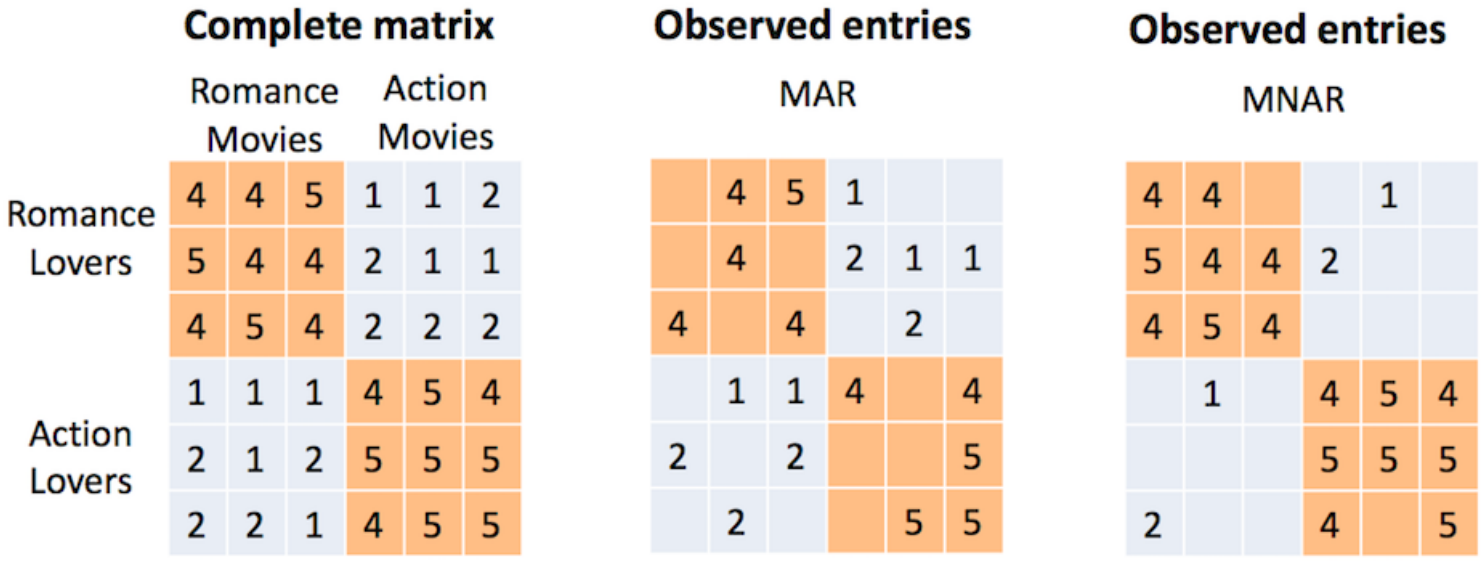


BCA : Block coordinate ascent
CA : Coordinate ascent
GA : Gradient ascent

	Time per iter.
VBMFSI-CA	158 sec.
VBMF-CA	70 sec
VBMF-BCA	66 min.
VBMF-GA	29 sec.

RMSE	VBMF-BCA		VBMF-CA	
	Iter.	Time	Iter.	Time
0.9005	19	21 h	63	74 m
0.9004	21	23 h	70	82 m
0.9003	22	24 h	84	98 m
0.9002	25	28 h	108	2 h
0.9001	27	31 h	680	13 h
0.9000	30	33 h		

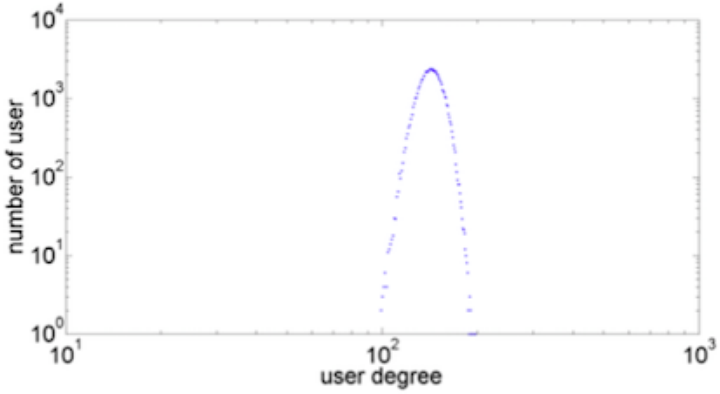
Missing Data



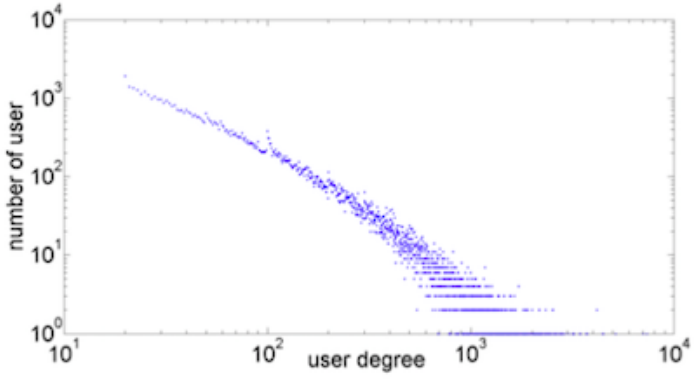
- Missing at random (MAR)
- Missing at non-random (MANR)

Non-Random Missing World

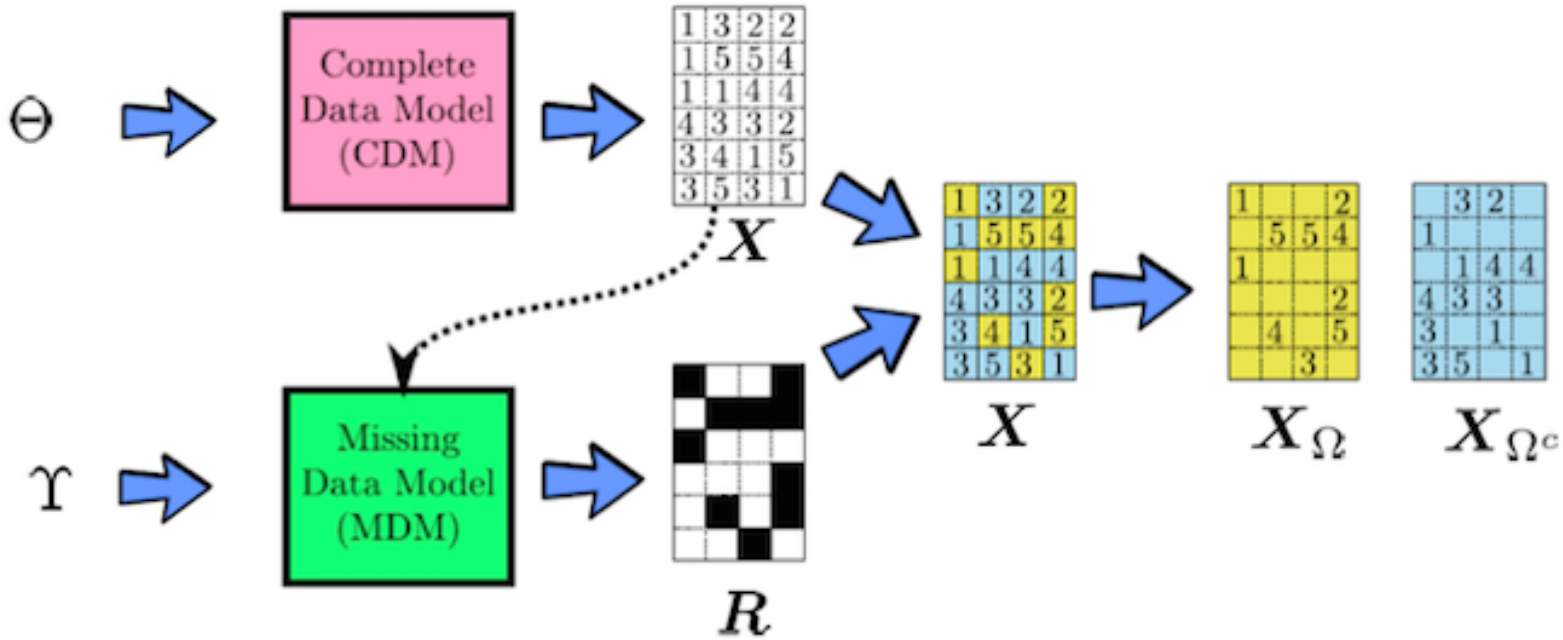
$$p(X_{ij} \text{ is observed}) = p$$



Power-law distribution



Data + Missing Models

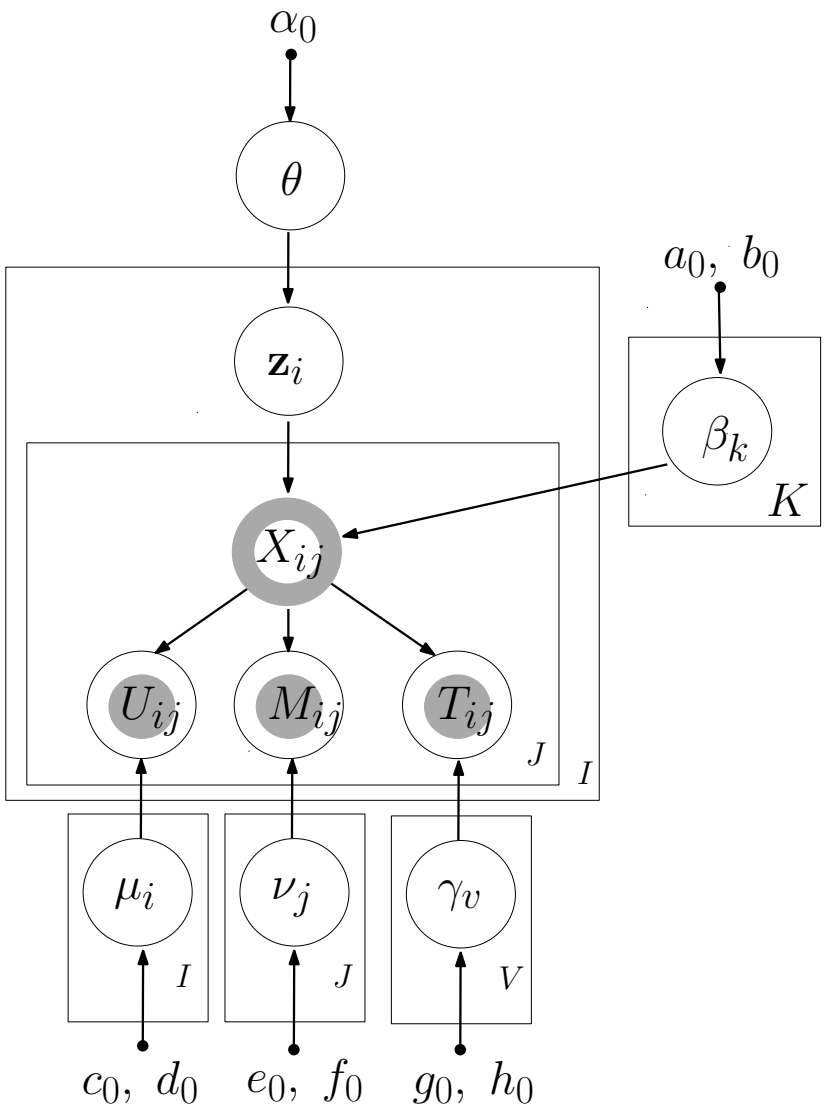


Joint distribution over $\mathbf{X}$ and $\mathbf{R}$ given $\mathbf{\Theta}$ and $\mathbf{\Upsilon}$,

$$p(\mathbf{X}, \mathbf{R}|\mathbf{\Theta}, \mathbf{\Upsilon}) = p(\mathbf{R}|\mathbf{X}, \mathbf{\Upsilon})p(\mathbf{X}|\mathbf{\Theta}).$$

Missing data model: Selection based on user activity, item popularity, and rating value.

Bayesian Binomial Mixture Model (Kim and Choi, RecSys-2014)



1. Choose the number of clusters K.
2. Choose mixing proportions $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\theta}|\alpha_0, K)$,
3. Choose parameters for binomial distributions

$$\boldsymbol{\beta} \sim \prod_{k=1}^{K}\prod_{j=1}^{J} \text{Beta}(\beta_{kj}|a_0, b_0)$$

4. For each user $i \in \{1, \ldots, I\}$
   Choose a cluster indicator:
   $\mathbf{z}_i|\boldsymbol{\theta} \sim \text{Mult}(\mathbf{z}_i|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k{}^{Z_{ki}}$.
   For each item $j \in \{1, \ldots, J\}$, Choose a rating
   $X_{ij}|\mathbf{z}_i, \boldsymbol{\beta} \sim \prod_{k=1}^{K} \text{Bin}(X_{ij}-1|\beta_{kj}, V-1)^{Z_{ki}}$.
5. Choose per-user, per-item, and per-value parameters for missing data model

$$\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma} \sim \prod_{i=1}^{I} \text{Beta}(\mu_i|c_0, d_0) \prod_{j=1}^{J} \text{Beta}(\nu_j|e_0, f_0) \prod_{v=1}^{V} \text{Beta}(\gamma_v|g_0, h_0)$$

6. For each $(i, j) \in \{1, \ldots, I\} \times \{1, \ldots, J\}$, Choose
   $U_{ij}|X_{ij}, \boldsymbol{\mu} \sim \text{Bern}(U_{ij}|\mu_i)$
   Choose $M_{ij}|X_{ij}, \boldsymbol{\nu} \sim \text{Bern}(M_{ij}|\nu_j)$
   Choose $T_{ij}|X_{ij}, \boldsymbol{\gamma} \sim \text{Bern}(T_{ij}|\gamma_{X_{ij}})$
   $R_{ij} = 1 - (1 - U_{ij})(1 - M_{ij})(1 - T_{ij})$.
   X_{ij} is observed if $R_{ij} = 1$, otherwise X_{ij} is missing.

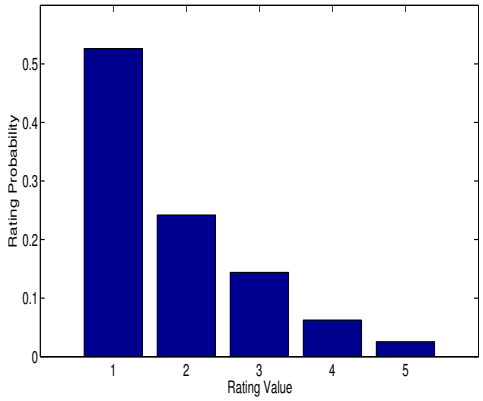
Missing Data Model

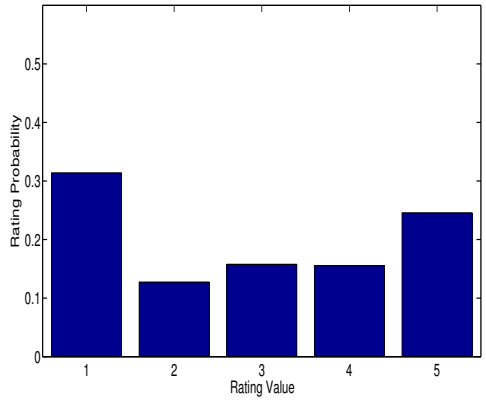
The probability of observation is assumed to be

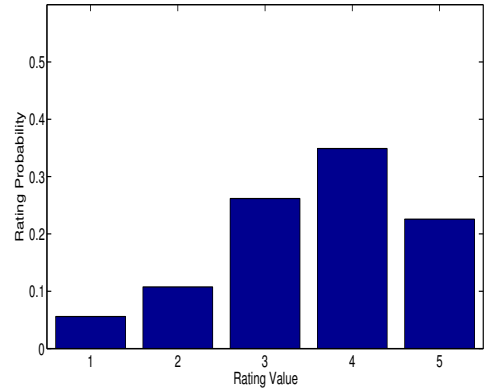
$$\begin{aligned} p(R_{ij} = 1 | X_{ij} = v) &= 1 - p(R_{ij} = 0 | X_{ij} = v) \\ &= 1 - p(U_{ij} = 0, M_{ij} = 0, T_{ij} = 0 | X_{ij} = v) \\ &= 1 - (1 - \mu_i)(1 - \nu_j)(1 - \gamma_v). \end{aligned}$$

μ_i	ν_j	γ_v	$p(R_{ij} = 1 \mid X_{ij} = v)$
0.01	0.01	0.01	0.029
0.01	0.01	0.1	0.118
0.01	0.1	0.1	0.198
0.1	0.1	0.1	0.271
0.9	0.1	0.1	0.919
0.9	0.5	0.1	0.955

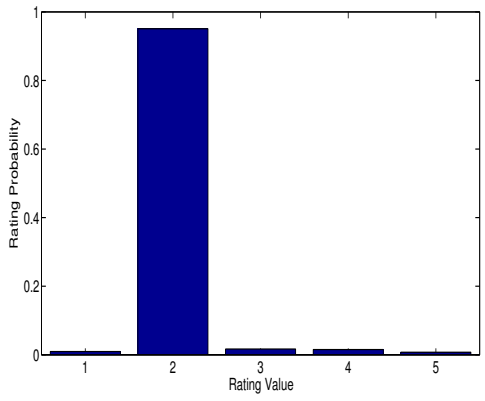
Distribution over rating values for Yahoo! Random, Yahoo! User, and MovieLens:

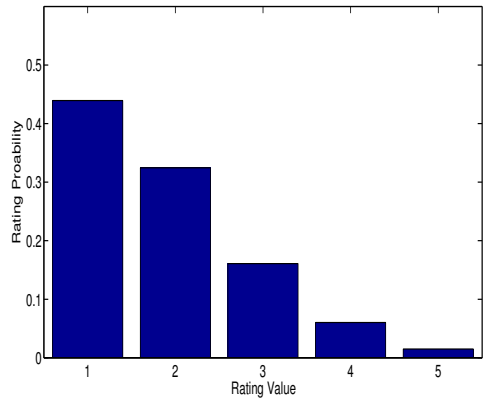






Prediction results on the Yahoo! Random dataset by MM/CPT-V and Bayesian-BM/OR model:





Posterior $q(U_{ij}, M_{ij}, T_{ij})$ explains the reason of presence/missing.

Yahoo! music dataset			
v	User	Item	Value
All	0.266	0.390	0.421
1	0.377	0.416	0.213
2	0.303	0.540	0.161
3	0.177	0.511	0.320
4	0.084	0.424	0.506
5	0.035	0.180	0.832

MovieLens dataset			
v	User	Item	Value
All	0.263	0.330	0.448
1	0.579	0.387	0.044
2	0.509	0.427	0.075
3	0.352	0.377	0.2943
4	0.178	0.291	0.580
5	0.098	0.274	0.701

Question and Discussion