Probabilistic Models for Data Combination in Recommender Systems

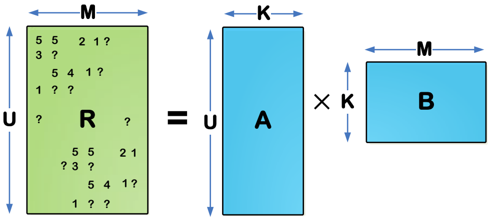
Sinead Williamson and Zoubin Ghahramani

UNIVERSITY OF CAMBRIDGE

December 13, 2008

The recommendation problem

- **Goal**: Given a set of users and a set of items, predict which items a given user will like.
- **Data**:
  - Ratings data: Incomplete set of ratings or preference indicators $r_{u,m}$ - may be continuous, discrete or binary
  - Item-specific information: Information about the items (technical specifications, descriptions, tag data...)
  - User-specific information: Demographic information, social network information...



Collaborative filtering vs content-based filtering

Collaborative filtering

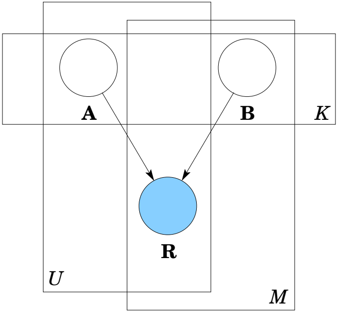
- **Idea**: Use information about which users liked which items to recommend further items to users
- **Advantages**:
  - Works really well! (Not *$1,000,000* well yet, but close...)
  - Captures both quality and genre/type of an item
- **Problems**:
  - What happens when we see a movie that has not been rated yet?
  - What do we recommend to a new user of the system

Content-based filtering

- **Idea**: Recommend items that are similar in content to those a user has liked before
- **Advantages**:
  - Extends easily to new items - it doesn't matter if no-one has rated a movie, provided we have content information
  - Captures genre data pretty well...
- **Problems**:
  - Often poor at identifying quality

Factorizing a single matrix

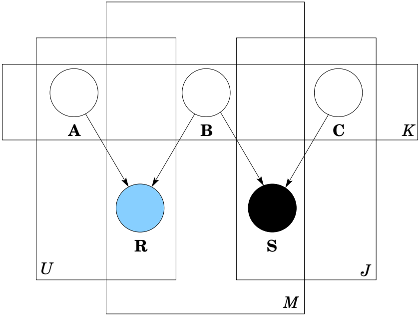
- A common approach in matrix factorization is to try to factorize the ratings matrix as the product of two lower-rank matrices
- **R** is a $U \times M$ matrix; **A** is a $U \times K$ matrix; **B** is an $M \times K$ matrix; $K < U, M$



$$\mathbf{R} \sim f(\mathbf{AB}^T)$$

Factorizing multiple matrices

- We can extend this idea to factorize multiple matrices, holding one of the latent matrices common to both factorizations, allowing us to use information from multiple sources (eg content, demographic, social data) to predict the unobserved ratings.



$$\mathbf{R} \sim f(\mathbf{AB}^T)$$
$$\mathbf{S} \sim f(\mathbf{CB}^T)$$

The linear Gaussian case: Variational Bayesian Inference

$$
\begin{aligned}
r_{um} | \mathbf{a}_u, \mathbf{b}_m &\sim \mathcal{N}(\mathbf{a}_u \mathbf{b}_m^T, \theta^2) \\
s_{jm} \mid \mathbf{c}_j, \mathbf{b}_m &\sim \mathcal{N}(\mathbf{c}_j \mathbf{b}_m^T, \phi^2) \\
a_{uk} &\sim \mathcal{N}(0, \alpha^2) \\
b_{mk} &\sim \mathcal{N}(0, \beta^2) \\
c_{jk} &\sim \mathcal{N}(0, \gamma^2)
\end{aligned}
$$

- Use Jensen's inequality to give a lower bound on the log marginal likelihood:

$$\log p(\mathbf{R}, \mathbf{S}) \geq \int d\mathbf{A} d\mathbf{B} d\mathbf{C} q(\mathbf{A}, \mathbf{B}, \mathbf{C}) \log \frac{p(\mathbf{R}, \mathbf{S}, \mathbf{A}, \mathbf{B}, \mathbf{C})}{q(\mathbf{A}, \mathbf{B}, \mathbf{C})}$$

- Make the variational approximation $q(A, B, C) = q(A)q(B)q(C)$
- $q(a_u)$ is Gaussian with mean μ_u and covariance Σ_u
- Take functional derivitive wrt $q(A)$ to optimize μ_u and Σ_u
- Similarly, optimize with respect to $q(B)$ and $q(C)$, and iterate to convergence

Results on MovieLens dataset

- Content data: Genre information provided with MovieLens dataset (100,000 ratings given by 943 users to 1682 movies.)
- Comparison methods: Single matrix factorization (VB inference); JRank [Basilico and Hofmann 2004]; Naive Bayes classifiers [Melville et al. 2002]
- Experiment 1: Randomly hold out one third of the test data; predict held out ratings

Table: MovieLens: RMSE on randomly held-out movies

Method	RMSE
Joint variational matrix factorization	0.9725
Single variational matrix factorization	0.9928
JRank	1.2108
Naive Bayes classifier + single variational matrix factorization	1.2734

Results on MovieLens dataset

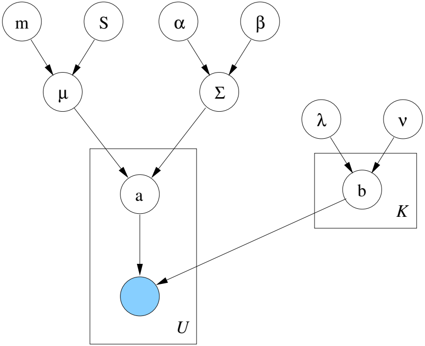
- Experiment 2: Simulated ‘new’ movies.
- Method:
    - Choose 200 movies with at least 200 ratings
    - Remove almost all ratings for these movies, mimicing a newly released movie
    - Predict the values for the held-out ratings

Table: MovieLens: RMSE on ‘new’ movies

Method	RMSE
Joint variational matrix factorization	1.1048
Single variational matrix factorization	1.2135
JRank	2.8722
Naive Bayes classifier + single variational matrix factorization	1.4013

The general Exponential Family case

- Gaussian data is nice, but isn't always appropriate
- Examples:
  - Play counts for different songs or artists on online radio stations - Poisson distribution
  - Binary data recording whether a user has bought an item - Bernoulli distribution
- We extend Bayesian Exponential Family PCA [Mohamed et al. 2008] to allow inference in the general Exponential Family case.
- Bayesian Exponential Family PCA: matrix **R** represents the *natural parameters* of the exponential family distribution of the data



$$\mathbf{r}_u | \mathbf{a}_u, \mathbf{B} \sim Expon(\sum_k a_{uk} \mathbf{b}_k)$$

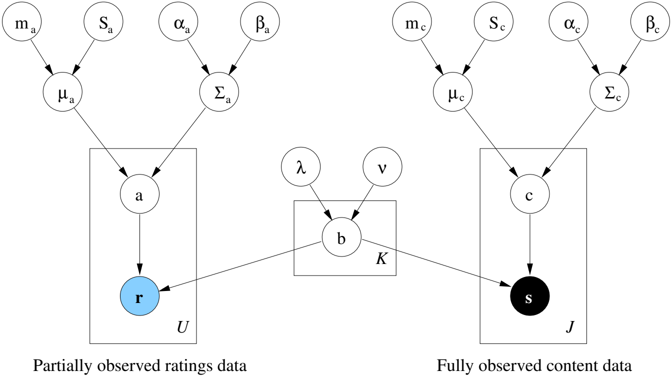
$$\mathbf{b}_k \sim Conj(\lambda, \nu)$$

$$\mathbf{a}_u \sim \mathcal{N}(\mu, \Sigma)$$

$$\mu \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$$

$$\sigma_k^2 \sim \mathcal{G}^{-1}(\alpha, \beta)$$

Extension to multiple matrices



- log joint probability is continuous, so can perform inference using Hybrid Monte Carlo
- Preliminary results on data from Last.fm (collaborative data = user/artist playcounts; content data = artist 'tags'), using the Poisson distribution, show improved log predictive probability in the 'new artist' scenario.

Summary

- Collaborative filtering often performs poorly in the case of new items or users.
- Joint matrix factorization allows us to combine content and ratings data in a reasonable manner.
- Incorporating this content information enables us to improve performance in the new item problem.
- Model can handle any Exponential Family distribution
- Inference via variational Bayes or Hybrid Monte Carlo.
- Model can easily be extended to include further sorts of information, for example information about users.

References

Basilico, J. and Hofmann, T. (2004).
Unifying collaborative and content-based filtering.
In *ICML*.

Melville, P., Mooney, R. J., and Nagarajan, R. (2002).
Content-boosted collaborative filtering.
In *Eighteenth national conference on Artificial Intelligence*.

Mohamed, S., Heller, K., and Ghahramani, Z. (2008).
Bayesian exponential family pca.
In *NIPS*.