Mixed Membership Matrix Factorization

Lester Mackey
University of California, Berkeley

Collaborators:
David Weiss, University of Pennsylvania
Michael I. Jordan, University of California, Berkeley

2011 Joint Statistical Meetings

A Problem

			
	5	3	?
	?	2	?
	1	?	4

Dyadic Data Prediction (DDP)

Learning from Pairs

- Given two sets of objects
  - Set of users and set of items
- Observe labeled object pairs
  - User u gave item j a rating r_{uj} of 5
- Predict labels of unobserved pairs
  - How will user u rate item k?

Examples

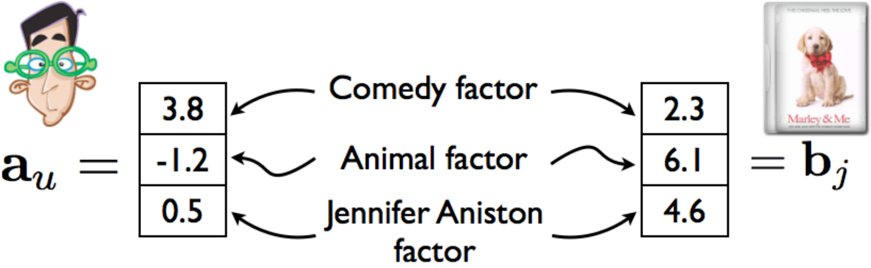
- Movie rating prediction in collaborative filtering
  - How will user u rate movie j?
- Click prediction in web search
  - Will user u click on URL j?
- Link prediction in a social network
  - Is user u friends with user j?

Prior Models for Dyadic Data

Latent Factor Modeling / Matrix Factorization

Rennie & Srebro (2005); DeCoste (2006); Salakhutdinov & Mnih (2008); Takács et al. (2009); Lawrence & Urtasun (2009)

- Associate latent factor vector, $\mathbf{a}_u \in \mathbb{R}^D$, with each user u
- Associate latent factor vector, $\mathbf{b}_j \in \mathbb{R}^D$, with each item j
- Generate expected rating via inner product



$$\mathbb{E}(r_{uj}) = \mathbf{a}_u \cdot \mathbf{b}_j = 3$$

Prior Models for Dyadic Data

Latent Factor Modeling / Matrix Factorization

Rennie & Srebro (2005); DeCoste (2006); Salakhutdinov & Mnih (2008); Takács et al. (2009); Lawrence & Urtasun (2009)

- Associate latent factor vector, $\mathbf{a}_u \in \mathbb{R}^D$, with each user u
- Associate latent factor vector, $\mathbf{b}_j \in \mathbb{R}^D$, with each item j
- Generate expected rating via inner product: $\mathbb{E}(r_{uj}) = \mathbf{a}_u \cdot \mathbf{b}_j$

Pro: State-of-the-art predictive performance

Con: Fundamentally static rating mechanism

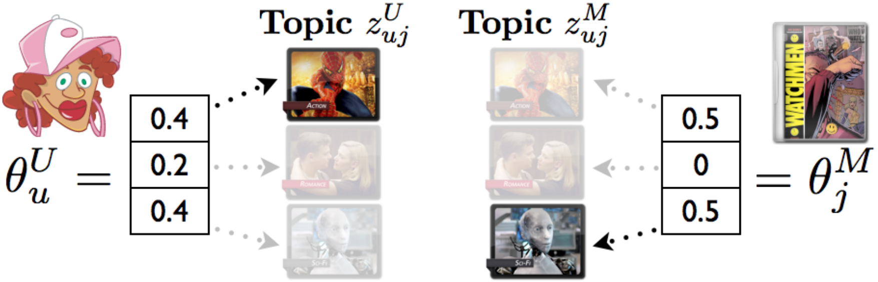
- Assumes user u rates according to $\mathbf{a}_u$, regardless of context
- In reality, dyadic interactions are heterogeneous
  - User's ratings may be influenced by instantaneous mood
  - Distinct users may share single account or web browser

Prior Models for Dyadic Data

Mixed Membership Topic Modeling

Airoldi, Blei, Fienberg, and Xing (2008); Porteous, Bart, and Welling (2008)

- Each user u maintains distribution over topics, $\theta_u^U \in \mathbb{R}^{K^U}$
- Each item j maintains distribution over topics, $\theta_j^M \in \mathbb{R}^{K^M}$
- Expected rating $\mathbb{E}(r_{uj})$ determined by *interaction-specific* topics sampled from user and item topic distributions



$$\mathbb{E}(r_{uj}) = f(z_{uj}^U, z_{uj}^M)$$

Prior Models for Dyadic Data

Mixed Membership Topic Modeling

Airoldi, Blei, Fienberg, and Xing (2008); Porteous, Bart, and Welling (2008)

- Each user u maintains distribution over topics, $\theta_u^U \in \mathbb{R}^{K^U}$
- Each item j maintains distribution over topics, $\theta_j^M \in \mathbb{R}^{K^M}$
- Expected rating $\mathbb{E}(r_{uj})$ determined by *interaction-specific* topics sampled from user and item topic distributions

Pro: Context-sensitive clustering

- User moods: in the mood for comedy vs. romance
- Item contexts: opening night vs. in high school classroom
- Multiple raters per account: parent vs. child

Con: Purely groupwise interactions

- Assumes user and item interact only through their topics
- Relatively poor predictive performance

Mixed Membership Matrix Factorization (M^{3}F)

Goal: Leverage the complementary strengths of latent factor models and mixed membership models for improved dyadic data prediction

General M^{3}F Framework (Mackey, Weiss, and Jordan, 2010):

- Users and items endowed both with latent factor vectors ($\mathbf{a}_u$ and $\mathbf{b}_j$) and with topic distribution parameters (θ_u^U and θ_j^M)
- To rate an item
  - User u draws topic i from θ_u^U
  - Item j draws topic k from θ_j^M
  - Expected rating

$$\mathbb{E}(r_{uj}) = \underbrace{\mathbf{a}_u \cdot \mathbf{b}_j}_{\text{static base rating}} + \underbrace{\beta_{uj}^{ik}}_{\text{context-sensitive bias}}$$

- M^{3}F models differ in specification of β_{uj}^{ik}
- Fully Bayesian framework

Mixed Membership Matrix Factorization (M^{3}F)

Goal: Leverage the complementary strengths of latent factor models and mixed membership models for improved dyadic data prediction

General M^{3}F Framework (Mackey, Weiss, and Jordan, 2010):

- M^{3}F models differ in specification of β_{uj}^{ik}

Specific M^{3}F Models:

- M^{3}F Topic-Indexed Bias Model
- M^{3}F Topic-Indexed Factor Model

M^{3}F Models

M^{3}F Topic-Indexed Bias Model (M^{3}F-TIB)

- Contextual bias decomposes into latent user and latent item bias

$$\beta_{uj}^{ik} = c_u^k + d_j^i$$

- Item bias d_j^i influenced by user topic i
  - Group predisposition toward liking/disliking item j
  - Captures polarizing *Napoleon Dynamite* effect
    - Certain movies provoke strongly differing reactions from otherwise similar users
- User bias c_u^k influenced by item topic k
  - Predisposition of u toward liking/disliking item group

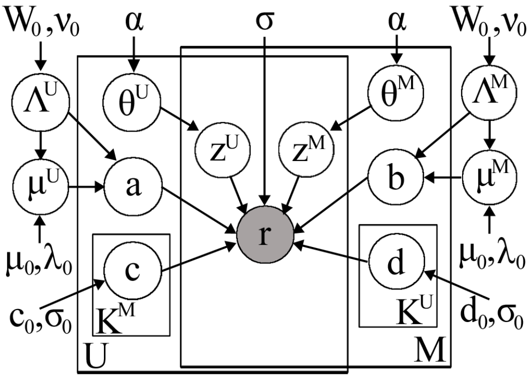
M^{3}F Models

M^{3}F Topic-Indexed Factor Model (M^{3}F-TIF)

- Contextual bias is an inner product of topic-indexed factor vectors
$$\beta_{uj}^{ik} = \mathbf{c}_u^k \cdot \mathbf{d}_j^i$$
- User u maintains latent vector $\mathbf{c}_u^k \in \mathbb{R}^{\tilde{D}}$ for each item topic k
- Item j maintains latent vector $\mathbf{d}_j^i \in \mathbb{R}^{\tilde{D}}$ for each user topic i
- Extends globally predictive factor vectors $(\mathbf{a}_u, \mathbf{b}_j)$ with context-specific factors

M^{3}F Inference and Prediction

Goal: Predict unobserved labels given labeled pairs



- Posterior inference over latent topics and parameters intractable
- Use block Gibbs sampling with closed form conditionals
  - User parameters sampled in parallel (same for items)
  - Interaction-specific topics sampled in parallel

M^{3}F Inference and Prediction

Goal: Predict unobserved labels given labeled pairs

- Bayes optimal prediction under root mean squared error (RMSE)

$$\textbf{M}^3\textbf{F-TIB:}\ \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)}+\sum_{k=1}^{K^M}c_u^{k(t)}\theta_{jk}^{M(t)}+\sum_{i=1}^{K^U}d_j^{i(t)}\theta_{ui}^{U(t)}\right)$$

$$\textbf{M}^3\textbf{F-TIF:}\ \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)}+\sum_{i=1}^{K^U}\sum_{k=1}^{K^M}\theta_{ui}^{U(t)}\theta_{jk}^{M(t)}\mathbf{c}_u^{k(t)}\cdot\mathbf{d}_j^{i(t)}\right)$$

Experimental Evaluation

The Data

- Real-world movie rating collaborative filtering datasets
- 1M MovieLens Dataset[1]
  - 1 million ratings in $\{1, \ldots, 5\}$
  - 6,040 users, 3,952 movies
- EachMovie Dataset
  - 2.8 million ratings in $\{1, \ldots, 6\}$
  - 1,648 movies, 74,424 users
- Netflix Prize Dataset[2]
  - 100 million ratings in $\{1, \ldots, 5\}$
  - 17,770 movies, 480,189 users

[1] http://www.grouplens.org/
[2] http://www.netflixprize.com/

Experimental Evaluation

The Setup

- Evaluate movie rating prediction performance on each dataset
  - RMSE as primary evaluation metric
  - Performance averaged over standard train-test splits
- Compare to state-of-the-art latent factor models
  - Bayesian Probabilistic Matrix Factorization[3] (BPMF)
    - M^3F reduces to BPMF when no topics are sampled
  - Gaussian process matrix factorization model[4] (L&U)
- Matlab/MEX implementation on dual quad-core CPUs

---

[3] Salakhutdinov and Mnih (2008)

[4] Lawrence and Urtasun (2009)

1M MovieLens Data

Question: How does M^3F performance vary with number of topics and static factor dimensionality?

- 3,000 Gibbs samples for M^3F-TIB and BPMF
- 512 Gibbs samples for M^3F-TIF ($\tilde{D} = 2$)

Method	**D=10**	**D=20**	**D=30**	**D=40**
BPMF	0.8695	0.8622	0.8621	0.8609
M^3F-TIB (1,1)	0.8671	0.8614	0.8616	0.8605
M^3F-TIF (1,2)	0.8664	0.8629	0.8622	0.8616
M^3F-TIF (2,1)	0.8674	0.8605	0.8605	0.8595
M^3F-TIF (2,2)	**0.8642**	**0.8584***	0.8584	0.8592
M^3F-TIB (1,2)	0.8669	0.8611	0.8604	0.8603
M^3F-TIB (2,1)	0.8649	0.8593	**0.8581***	**0.8577***
M^3F-TIB (2,2)	0.8658	0.8609	0.8605	0.8599
L&U (2009)	0.8801 (RBF)		0.8791 (Linear)	

EachMovie Data

Question: How does M^3F performance vary with number of topics and static factor dimensionality?

- 3,000 Gibbs samples for M^3F-TIB and BPMF
- 512 Gibbs samples for M^3F-TIF ($\tilde{D} = 2$)

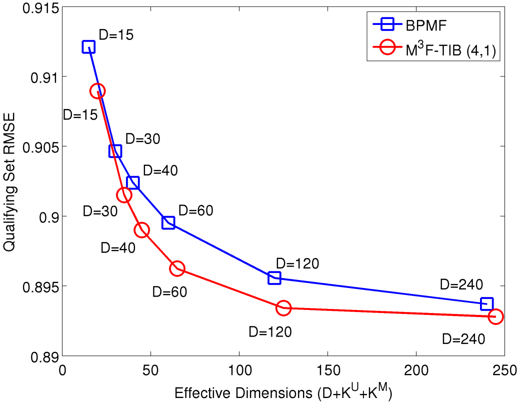
Method	**D=10**	**D=20**	**D=30**	**D=40**
BPMF	1.1229	1.1212	1.1203	1.1163
M^3F-TIB (1,1)	1.1205	1.1188	1.1183	1.1168
M^3F-TIF (1,2)	1.1351	1.1179	1.1095	1.1072
M^3F-TIF (2,1)	1.1366	1.1161	1.1088	1.1058
M^3F-TIF (2,2)	1.1211	1.1043	1.1035	1.1020
M^3F-TIB (1,2)	1.1217	1.1081	1.1016	1.0978
M^3F-TIB (2,1)	1.1186	1.1004	1.0952	1.0936
M^3F-TIB (2,2)	**1.1101***	**1.0961***	**1.0918***	**1.0905***
L&U (2009)	1.1111 (RBF)		1.0981 (Linear)	

Netflix Prize Data

Question: How does performance vary with latent dimensionality?

- Contrast M^{3}F-TIB $(K^U, K^M) = (4, 1)$ with BPMF
- 500 Gibbs samples for M^{3}F-TIB and BPMF

Method	RMSE	Time
BPMF/15	0.9121	27.8s
TIB/15	**0.9090**	46.3s
BPMF/30	0.9047	38.6s
TIB/30	**0.9015**	56.9s
BPMF/40	0.9027	48.3s
TIB/40	**0.8990**	70.5s
BPMF/60	0.9002	94.3s
TIB/60	**0.8962**	97.0s
BPMF/120	0.8956	273.7s
TIB/120	**0.8934**	285.2s
BPMF/240	0.8938	1152.0s
TIB/240	**0.8929**	1158.2s



Stratification

Question: Where are improvements over BPMF being realized?

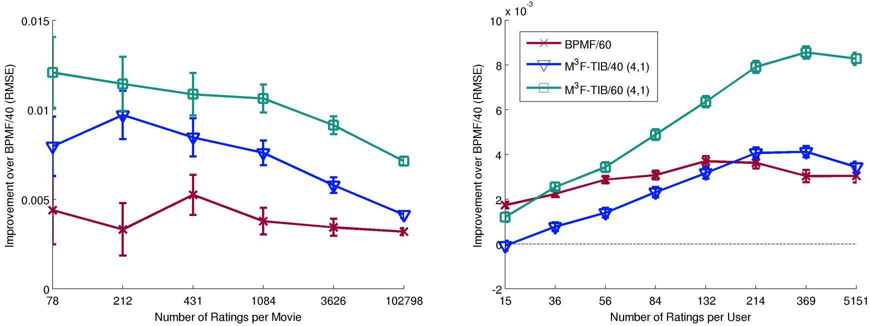


Figure: RMSE improvements over BPMF/40 on the Netflix Prize as a function of movie or user rating count. Left: Each bin represents 1/6 of the movie base. Right: Each bin represents 1/8 of the user base.

The *Napolean Dynamite* Effect

Question: Do M^{3}F models capture polarization effects?

Table: Top 200 Movies from the Netflix Prize dataset with the highest and lowest cross-topic variance in $\mathbb{E}(d^i_j|\mathbf{r}^{(\mathrm{v})})$.

Movie Title	$\mathbb{E}(d^i_j\|\mathbf{r}^{(\mathrm{v})})$
Napoleon Dynamite	-0.11 $\pm$ 0.93
Fahrenheit 9/11	-0.06 $\pm$ 0.90
Chicago	-0.12 $\pm$ 0.78
The Village	-0.14 $\pm$ 0.71
Lost in Translation	-0.02 $\pm$ 0.70
LotR: The Fellowship of the Ring	0.15 $\pm$ 0.00
LotR: The Two Towers	0.18 $\pm$ 0.00
LotR: The Return of the King	0.24 $\pm$ 0.00
Star Wars: Episode V	0.35 $\pm$ 0.00
Raiders of the Lost Ark	0.29 $\pm$ 0.00

Conclusions

New framework for dyadic data prediction

- Strong predictive performance and static specificity of latent factor models
- Clustered context-sensitivity of mixed membership topic models
- Outperforms pure latent factor modeling while fitting fewer parameters
- Greatest improvements for high-variance, sparsely rated items

Future work

- Modeling user choice: missingness is informative
- Nonparametric priors on topic parameters
- Alternative approaches to inference

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic blockmodels. *JMLR*, 9:1981–2014, 2008.

DeCoste, D. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *ICML*, 2006.

Lawrence, N.D. and Urtasun, R. Non-linear matrix factorization with Gaussian processes. In *ICML*, 2009.

Mackey, L., Weiss, D., and Jordan, M. I. Mixed membership matrix factorization. In *ICML*, June 2010.

Porteous, I., Bart, E., and Welling, M. Multi-HDP: A non parametric Bayesian model for tensor factorization. In *AAAI*, 2008.

Rennie, J. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, 2005.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Marlo. In *ICML*, 2008.

Takács, G., Pilászy, I., Németh, B., and Tikk, D. Scalable collaborative filtering approaches for large recommender systems. *JMLR*, 10:623–656, 2009.

The End

Thanks!