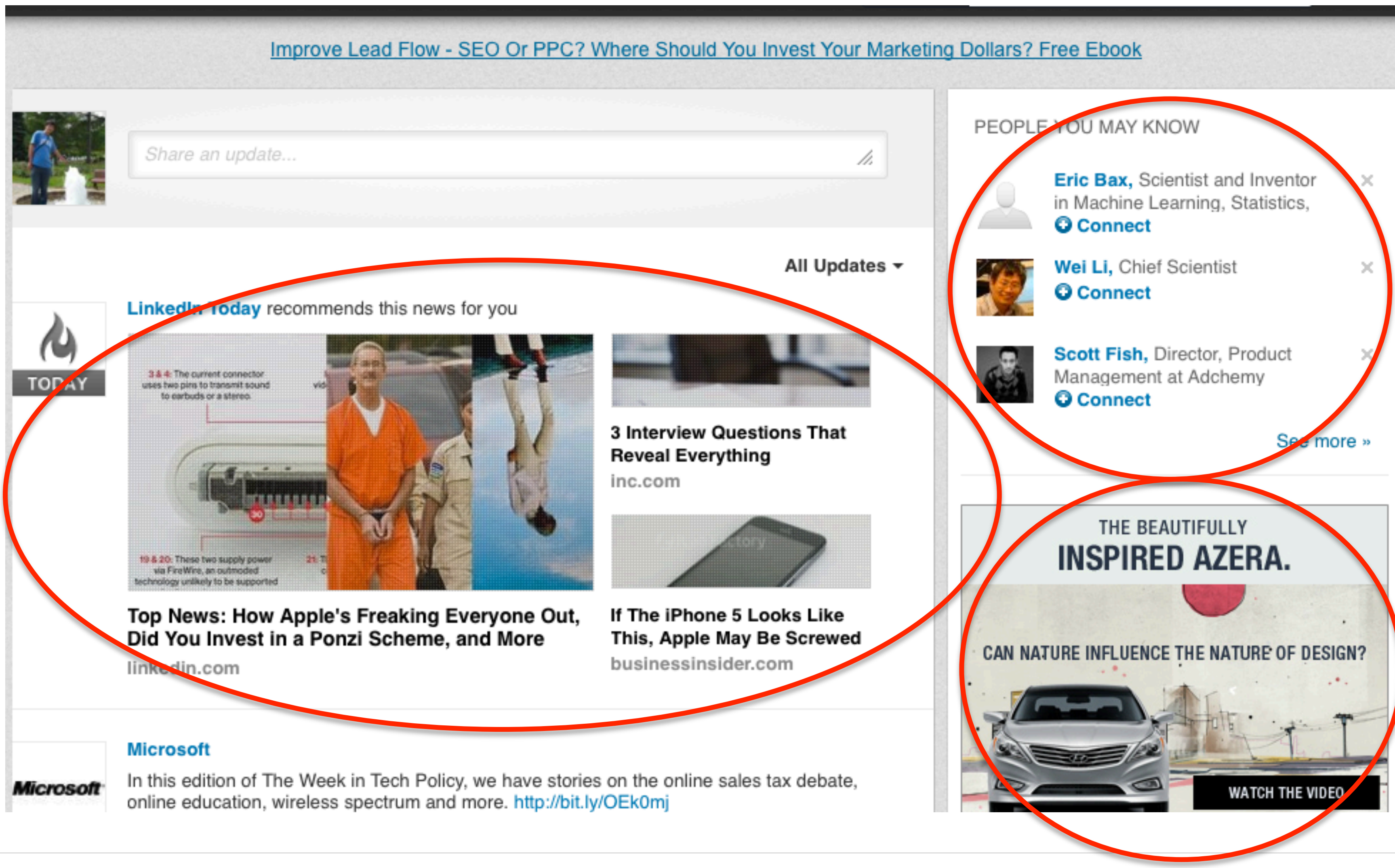
Webpage Personalization and User Profiling

Liang Zhang

Computational Advertising Workshop at SAMSI

Aug 8, 2012

Personalized Webpage Is Everywhere



LinkedIn

Personalized Webpage Is Everywhere

Common Properties of Web Personalization Problem

- One or multiple metrics to optimize
  - Click Through Rate (CTR) (focus of this talk)
  - Revenue per impression
  - Time spent on the landing page
  - Ad conversion rate
  - …
- Large scale data
  - Map-Reduce to solve the problem!
- Sparsity
- Cold-start
  - User features: Age, gender, position, industry, …
  - Item features: Category, key words, creator features, …

Scope of This Talk

- CTR prediction for a user on an item
- Assumptions:
  - There are sufficient data per item to estimate per-item model
  - Serving bias and positional bias are removed by randomly serving scheme
  - Item popularities are quite dynamic and have to be estimated in real-time fashion
- Examples:
  - Yahoo! Front page Today module
  - Linkedin Today module

Online Logistic Regression (OLR)

- User i with feature $\mathbf{x}_i$, article j
- Binary response y (click/non-click)
- $y_{ij} = Bernoulli(p_{ij})$
- $s_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \boldsymbol{x}_i' \boldsymbol{\beta}_j$
- Prior $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$
- Using Laplace approximation or variational Bayesian methods to obtain posterior

$$\boldsymbol{\beta}_j | y_{ij} \sim N(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$$

- New prior $\boldsymbol{\beta}_j \sim N(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$
- Can approximate $\boldsymbol{\Sigma}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ as diagonal for high dim $\mathbf{x}_i$

User Features for OLR

- Age, gender, industry, job position for login users
- General behavior targeting (BT) features
  - Music? Finance? Politics?
- User profiles from historical view/click behavior on previous items in the data, e.g.
  - Item-profile: use previously clicked item ids as the user profile
  - Category-profile: use item category affinity score as profile. The score can be simply user's historical CTR on each category.
  - Are there better ways to generate user profiles?
  - Yes! By matrix factorization!

Generalized Matrix Factorization (GMF) Framework

- $$y_{ij} \sim Bernoulli(p_{ij}),$$

$$s_{ij} = \log \frac{p_{ij}}{1-p_{ij}}$$

$$s_{ij} = \underbrace{f(x_{ij})}_{\text{Global Features}} + \underbrace{\alpha_i}_{\text{User effect}} + \underbrace{\beta_j}_{\text{Item effect}} + \underbrace{\boldsymbol{u}_i'}_{\text{User factors}}\underbrace{\boldsymbol{v}_j}_{\text{Item factors}}.$$

Bell et al. (2007)

Regression Priors

- User covariates

$$\alpha_i \sim N(g(x_i), \sigma_\alpha^2), \qquad \boldsymbol{u}_i \sim N(G(x_i), \sigma_u^2 I),$$
$$\beta_j \sim N(h(x_j), \sigma_\beta^2), \qquad \boldsymbol{v}_j \sim N(H(x_j), \sigma_v^2 I),$$

Item covariates

- g(·), h(·), G(·), H(·) can be any regression functions

- Agarwal and Chen (KDD 2009); Zhang et al. (RecSys 2011)

-

Different Types of Prior Regression Models

- Zero prior mean
  - Bilinear random effects (BIRE)
- Linear regression
  - Simple regression (RLFM)
  - Lasso penalty (LASSO)
- Tree Models
  - Recursive partitioning (RP)
  - Random forests (RF)
  - Gradient boosting machines (GB)
  - Bayesian additive regression trees (BART)

Model Fitting Using MCEM

- Monte Carlo EM (Booth and Hobert 1999)
- Let $\boldsymbol{\Theta} = (f, g, h, G, H, \sigma_\alpha^2, \sigma_u^2, \sigma_\beta^2, \sigma_v^2)$
- Let $\boldsymbol{\Delta} = \{\alpha_i, \beta_j, \boldsymbol{u}_i, \boldsymbol{v}_j\}\forall i,j$
- E Step: $q_t(\boldsymbol{\Theta}) = E_{\boldsymbol{\Delta}}[\log L(\boldsymbol{\Theta}; \boldsymbol{\Delta}, \boldsymbol{y}) \,|\, \hat{\boldsymbol{\Theta}}^{(t)}]$
  - Obtain N samples of conditional posterior

$$p(\alpha_i | \sim), p(\beta_j | \sim), p(\boldsymbol{u}_i | \sim), p(\boldsymbol{v}_j | \sim)$$

- M Step: $\hat{\boldsymbol{\Theta}}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} \; q_t(\boldsymbol{\Theta})$.

Handling Binary Responses

- Gaussian responses:

  $p(\alpha_i | \sim), p(\beta_j | \sim), p(\boldsymbol{u}_i | \sim), p(\boldsymbol{v}_j | \sim)$ have closed form

- Binary responses + Logistic: no longer closed form

- Variational approximation (VAR)

- Adaptive rejection sampling (ARS)

Simulation Study

- 10 simulated data sets, 100K samples for both training and test
- 1000 users and 1000 items in training
- Extra 500 new users and 500 new items in test + old users/items
- For each user/item, 200 covariates, only 10 useful
- Construct non-linear regression model from 20 Gaussian functions for simulating α, β, u and v following Friedman (2001)

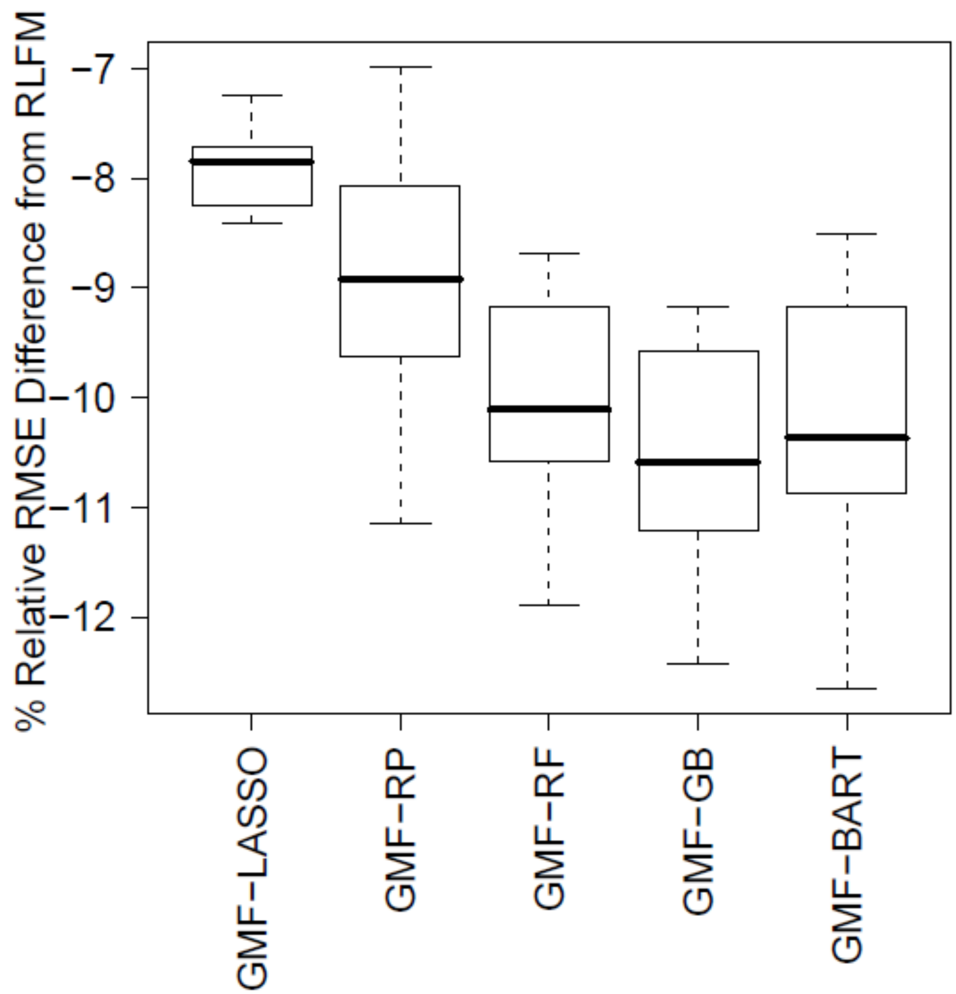


Figure 1: Boxplot of percentage RMSE difference relative to RLFM for 10 simulated datasets

MovieLens 1M Data Set

- 1M ratings
- 6040 users
- 3706 movies
- Sort by time, first 75% training, last 25% test
- A lot of new users in the test data set
- User features: Age, gender, occupation, zip code
- Item features: Movie genre

Performance Comparison

Model	Test RMSE	Warm-start RMSE	Cold-start RMSE
Constant	1.1190	---	---
Feature-only	1.0906	---	---
Most Popular	0.9726	---	---
BIRE	0.9435	---	---
RLFM	0.9363	0.8814	0.9766
GMF-RP	0.9359	0.8784	0.9783
GMF-GB	0.9344	0.8791	0.9753
GMF-RF	0.9343	0.8777	0.9760
GMF-LASSO	0.9341	0.8779	0.9755
GMF-BART	0.9340	0.8780	0.9753

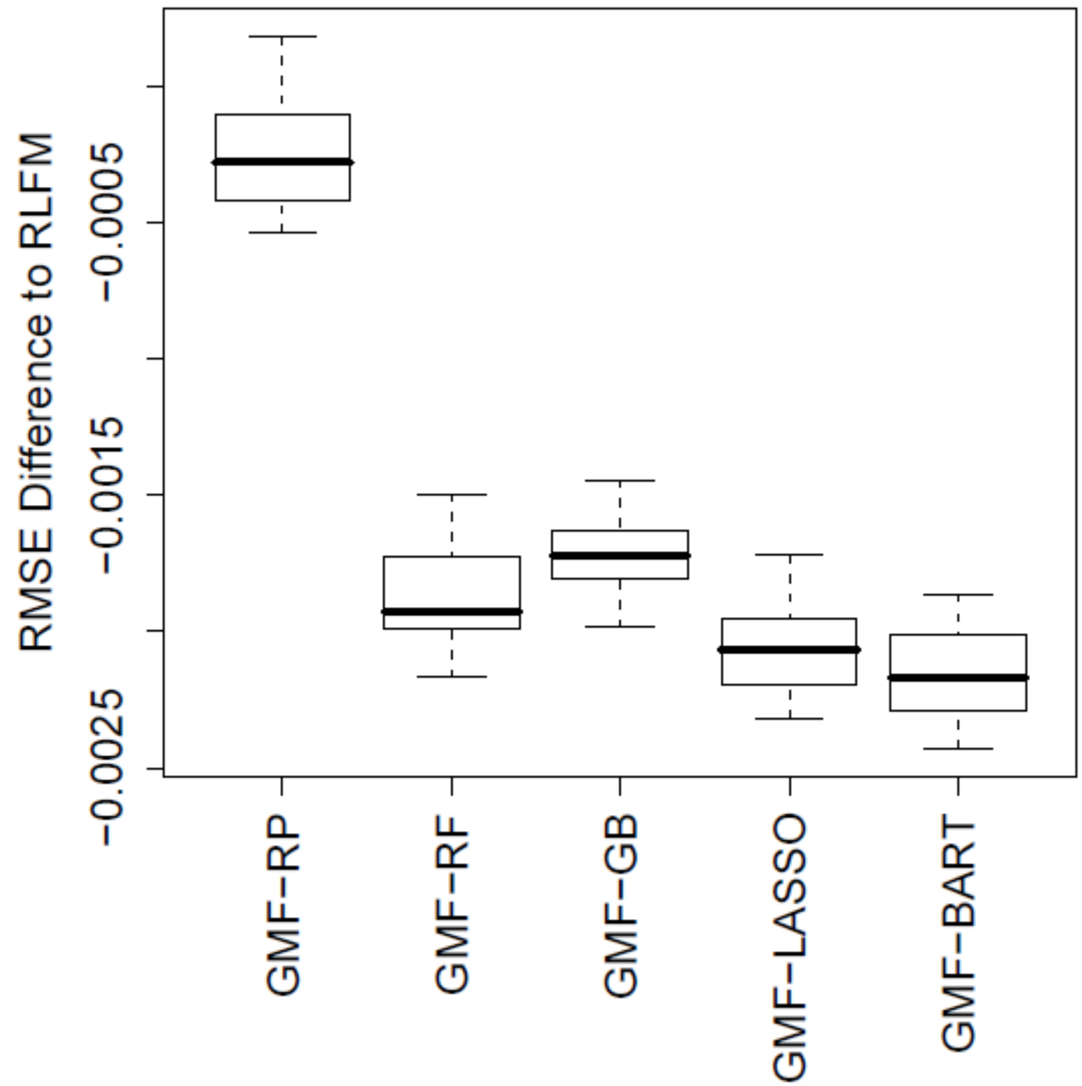


Figure 2: Boxplot of test-set RMSE differences from RLFM on 20 bootstrap samples of MovieLens-1M

However…

- We are working with very large scale data sets!
- Parallel matrix factorization methods using Map-Reduce has to be developed!
- Khanna et al. 2012 Technical report

Model Fitting Using MCEM (Single Machine)

- Monte Carlo EM (Booth and Hobert 1999)
- Let $\boldsymbol{\Theta} = (f, g, h, G, H, \sigma_\alpha^2, \sigma_u^2, \sigma_\beta^2, \sigma_v^2)$
- Let $\boldsymbol{\Delta} = \{\alpha_i, \beta_j, \boldsymbol{u}_i, \boldsymbol{v}_j\}_{\forall i,j}$
- E Step: $q_t(\boldsymbol{\Theta}) = E_{\boldsymbol{\Delta}}[\log L(\boldsymbol{\Theta}; \boldsymbol{\Delta}, \boldsymbol{y}) \,|\, \hat{\boldsymbol{\Theta}}^{(t)}]$
  - Obtain N samples of conditional posterior

$$p(\alpha_i | \sim), p(\beta_j | \sim), p(\boldsymbol{u}_i | \sim), p(\boldsymbol{v}_j | \sim)$$

- M Step: $\hat{\boldsymbol{\Theta}}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} \; q_t(\boldsymbol{\Theta}).$

Parallel Matrix Factorization

- Partition data into m partitions
- For each partition $\ell \in \{1, ..., m\}$ run MCEM algorithm and get $\hat{\Theta}_\ell$.
- Let $\hat{\Theta} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\Theta}_\ell$.
- Ensemble runs: for k = 1, … , n
  - Repartition data into m partitions with a new seed
  - Run E-step only job for each partition given $\hat{\Theta}$
- Average over user/item factors for all partitions and k's to obtain the final estimate

Key Points

- Partitioning is tricky!
  - By events? By items? By users?

- Empirically, “divide and conquer” + average over $\hat{\Theta}_\ell$ to obtain $\hat{\Theta}$ work well!

- Ensemble runs: After obtained $\hat{\Theta}$, we run n E-step-only jobs and take average, for each job using a different user-item mix.

Identifiability Issues

- Same log-likelihood can be achieved by
  - g ( ) = g ( ) + r, h ( ) = h ( ) – r
    - Center **α, β, u** to zero-mean every E-step
  - **u = -u, v = -v**
    - Constrain v to be positive
  - Switching $u_{\cdot 1}$, $v_{\cdot 1}$ with $u_{\cdot 2}$, $v_{\cdot 2}$
    - $\mathbf{u}_i \sim N(G(x_i), \mathbf{I})$, $\mathbf{v}_j \sim N(H(x_j), \boldsymbol{\lambda}\mathbf{I})$
    - Constraint: Diagonal entries $\lambda_1 >= \lambda_2 >= \ldots$

Matrix Factorization For User Profile

- Offline user profile building period, obtain the user factor $\boldsymbol{u}_i$ for user i

- Online modeling using OLR
  - If a user has a profile (warm-start), use $\boldsymbol{u}_i$ as the user feature
  - If not (cold-start), use $G(x_i)$ as the user feature

Offline Evaluation Metric Related to Clicks

- For model M and J live items (articles) at any time

$$S(M) = J \sum_{visits\ with\ click} 1(\text{item clicked} = \text{item selected by M}).$$

- If M = random (constant) model

  E[S(M)] = #clicks

- Unbiased estimate of expected total clicks (Langford et al. 2008)

Experiments

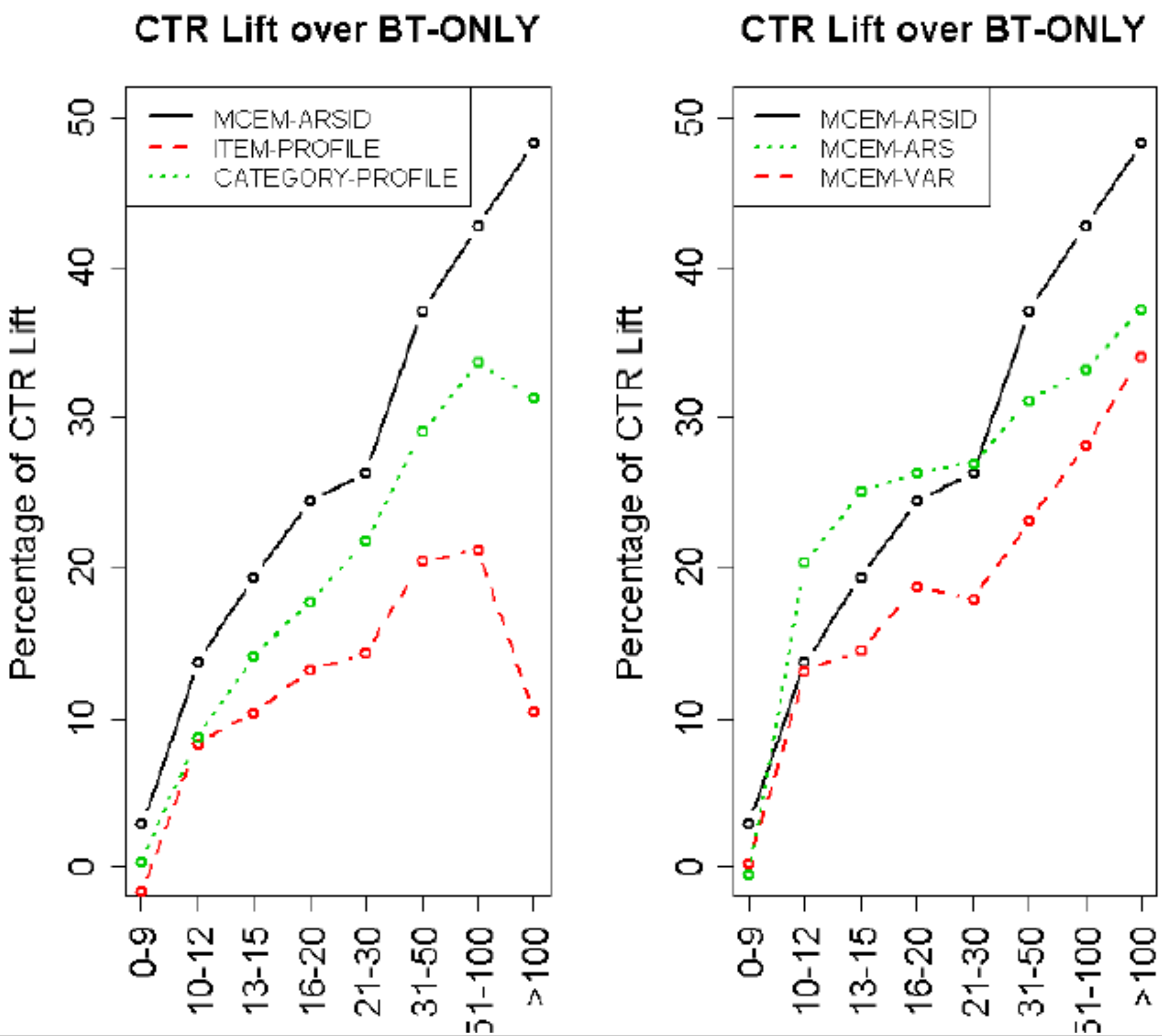
- Yahoo! Front Page Today Module data
- Data for building user profile: 8M users with at least 10 clicks (heavy users) in June 2011, 1B events
- Data for training and testing OLR model: Random served data with 2.4M clicks in July 2011
- Heavy users contributed around 30% of clicks
- User feature for OLR:
  - Intercept-only (MOST POPULAR)
  - 124 Behavior targeting features (BT-ONLY)
  - BT + top 1000 clicked article ids (ITEM-PROFILE)
  - BT + user profile with CTR on 43 binary content categories (CATEGORY-PROFILE)
  - BT + profiles from matrix factorization models

Click Lift Performance For Different User Profiles

Table 3: The overall click lift over the user behavior feature (BT) only model.

Method	#Ensembled Runs	Overall	Warm Start	Cold Start
ITEM-PROFILE	–	3.0%	14.1%	-1.6%
CATEGORY-PROFILE	–	6.0%	20.0%	0.3%
MCEM-VAR	10	5.6%	18.7%	0.2%
MCEM-ARS	10	7.4%	26.8%	-0.5%
MCEM-ARSID	1	9.1%	24.6%	2.8%
MCEM-ARSID	10	9.7%	26.3%	2.9%

Click Lift vs #Clicks in Training Data



User Profile Model with Graphical Lasso (UPG)

- $s_{ij} = f(x_{ij}) + \alpha_i + \beta_j + \phi_{ij}.$

  User-item affinity

- $(\phi_{i1}, \ldots \phi_{ip}) \sim N(0, \boldsymbol{\Sigma})$

- Unknown Σ represents item-item similarity
- Yet another way to model CTR
- Agarwal, Zhang and Mazumder (2011), Annals of Applied Statistics

Covariance Matrix Regularization

- **Σ** need to be regularized, especially for high-dimensional problems (e.g. thousands of items)
- Prior log-likelihood without constant (N_i=#users)

$$\frac{N_i}{2}\log(\det(\Sigma^{-1})) - \frac{1}{2}\sum_i \phi_i \Sigma^{-1} \phi_i - \boxed{N_i \rho \|\Sigma^{-1}\|_k}$$

Regularize the precision matrix Ω

- k=1, Graphical lasso problem (Banerjee et al. 2007, Friedman et al. 2007)

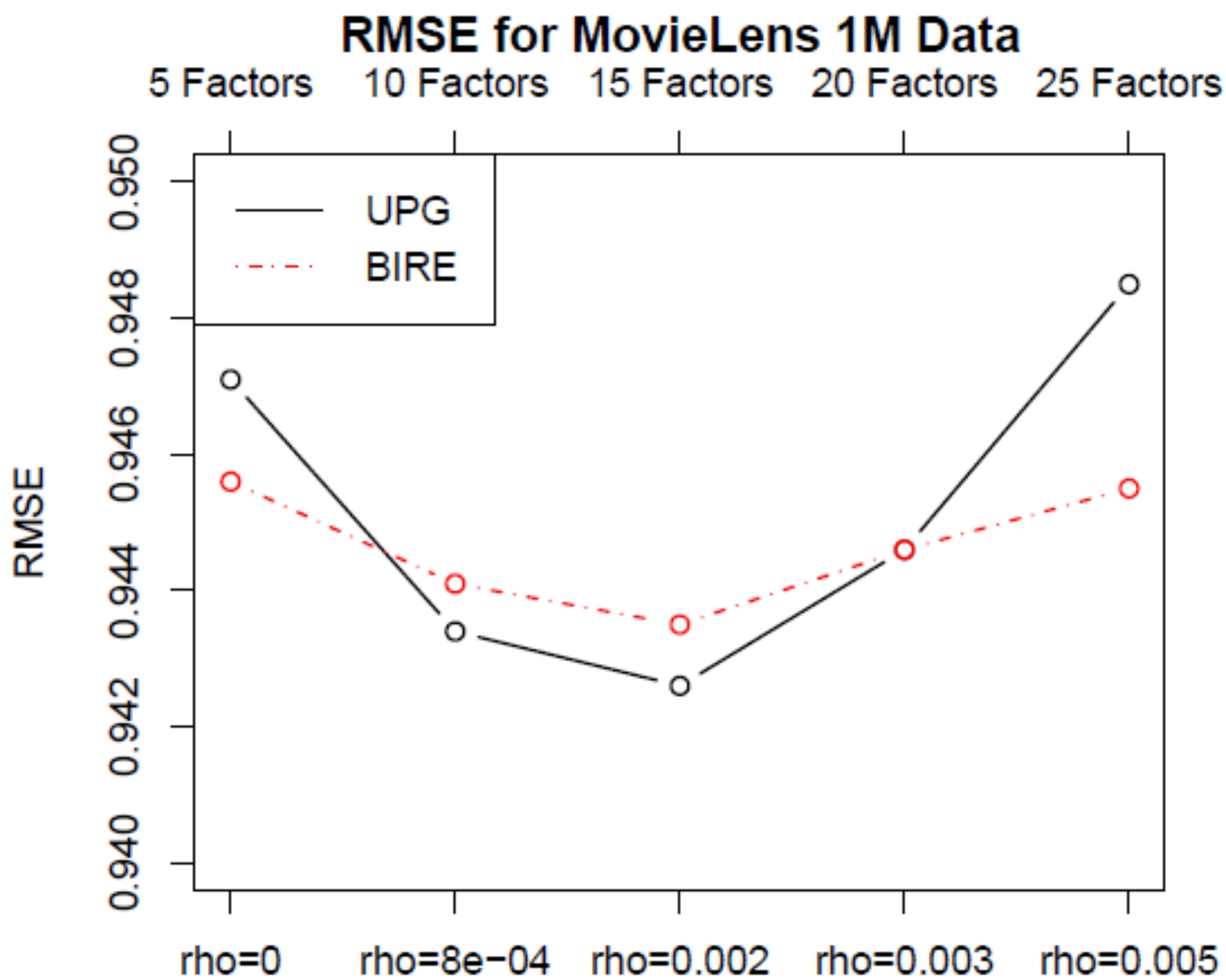
Model Fitting For UPG

- E Step:
  - For each user i, obtain posterior $p(\boldsymbol{\phi}_i | \sim) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
- M Step

$$E_{\boldsymbol{\phi}|\hat{\boldsymbol{\Omega}},\boldsymbol{Z}}[\sum_i \log p(\boldsymbol{\phi}_i|\boldsymbol{\Omega})] =$$

$$-\frac{pN_i}{2}\log(2\pi) + \frac{N_i}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\sum_i \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_i) + \boldsymbol{\mu}_i'\boldsymbol{\Omega}\boldsymbol{\mu}_i - N_i\rho||\boldsymbol{\Omega}||_1$$

- Let $S = \frac{\sum_i(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i')}{N_i}$ be the sample covariance matrix for graphical Lasso


RMSE for MovieLens 1M Data
5 Factors
10 Factors
15 Factors
20 Factors
25 Factors
UPG
BIRE
RMSE
0.940
0.942
0.944
0.946
0.948
0.950
rho=0
rho=8e−04
rho=0.002
rho=0.003
rho=0.005

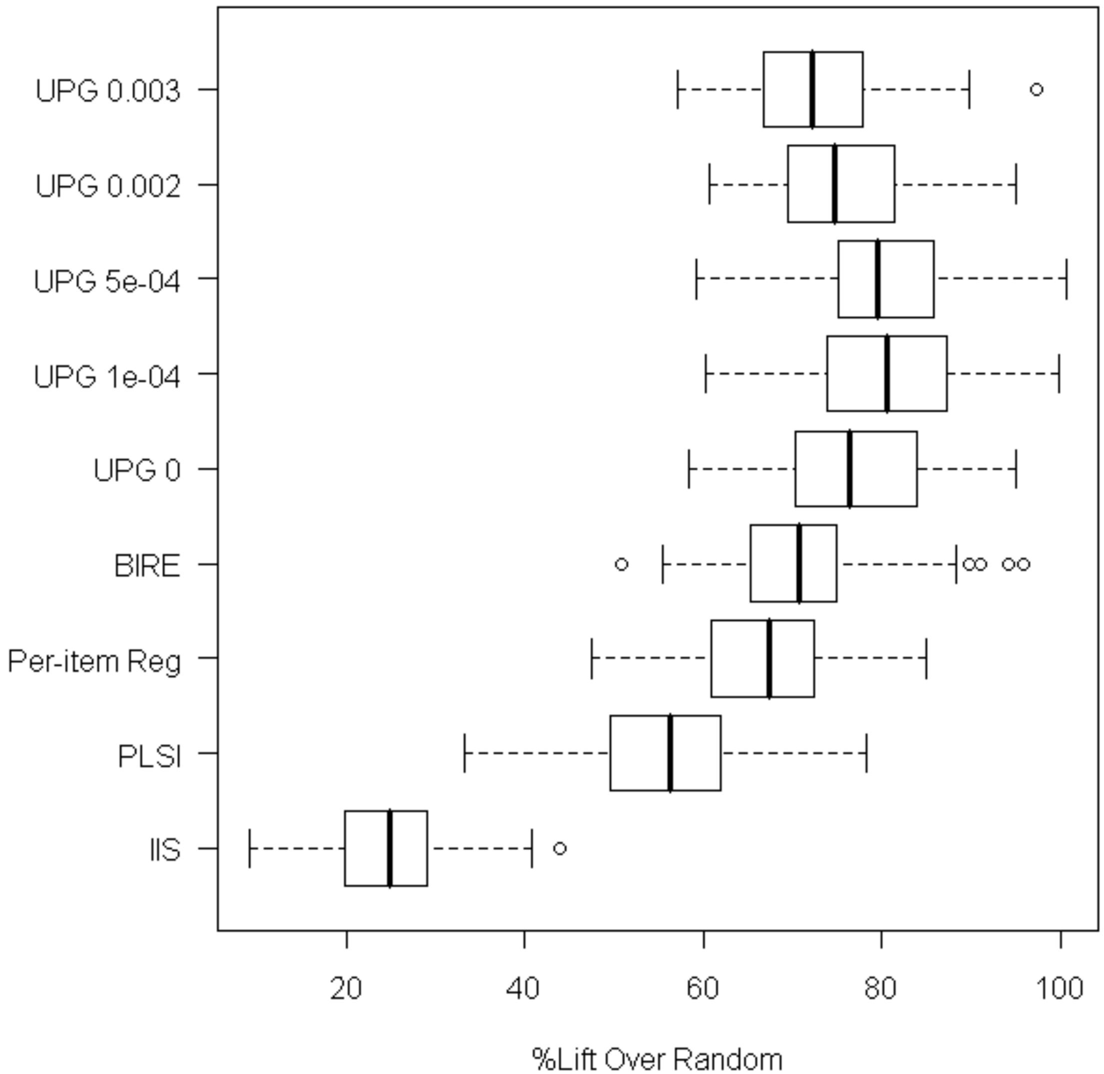
Fitted Precision Matrix

The Pair of Movies	Partial Correlation
The Godfather (1972) The Godfather: Part II (1974)	0.622
Grumpy Old Men (1993) Grumpier Old Men (1995)	0.474
Patriot Games (1992) Clear and Present Danger (1994)	0.448
The Wrong Trousers (1993) A Close Shave (1995)	0.443
Toy Story (1995) Toy Story 2 (1999)	0.428
Austin Powers: International Man of Mystery (1997) Austin Powers: The Spy Who Shagged Me (1999)	0.422
Star Wars: Episode IV - A New Hope (1977) Star Wars: Episode V - The Empire Strikes Back (1980)	0.417
Young Guns (1988) Young Guns II (1990)	0.395
A Hard Day's Night (1964) Help! (1965)	0.378
Lethal Weapon (1987) Lethal Weapon 2 (1989)	0.364

Real World Data from Yahoo! PA

- 51 items
- Training data
  - 5M binary observations (click/non-click)
  - 140K users
- Test data
  - Random bucket
  - 528K binary observations
- User features: Age, gender, behavior targeting

The Click-Lift Measure for PA Data



Fitted Precision Matrix

Table 1: Pairs of applications with top 10 absolute value of partial correlations in the dense precision matrix from user profile model without Glasso.

Application 1	Application 2	Partial Correlation
Fantasy Sports	Fantasy MLB	0.556
Fantasy Sports	Fantasy Football	0.434
AOL Mail	Gmail	0.367
PEOPLE.com	EW.com Featured	0.265
Shopping	Personals	0.237
PEOPLE.com	PopSugar	0.224
Travel	Shopping	0.222
News	Shopping	0.208
EW.com Featured	PopSugar	0.182
News	Personals	0.181

What To Do When Not Enough Data Per Item?

- Example:
  - CTR prediction for ad creatives/campaigns

- User i with feature $\mathbf{x}_i$

- Item j with feature $\mathbf{x}_j$

-

$$y_{ij} = Bernoulli(p_{ij})$$

$$s_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \boxed{\boldsymbol{x}_i' \boldsymbol{A} \boldsymbol{x}_j} + \boxed{\boldsymbol{x}_i' \boldsymbol{\beta}_j}$$

Offline Model Component

Online Model Component

Agarwal et al. (KDD 2010)

Large Scale Logistic Regression

- Naïve:
  - Partition the data and run logistic regression for each partition
  - Take the mean of the learned coefficients
  - Problem: Not guaranteed to converge to the model from single machine!
- Alternating Direction Method of Multipliers (ADMM)
  - Boyd et al. 2011
  - Set up a constraint that each partition's coefficient = global consensus
  - Solve the optimization problem using Lagrange Multipliers
- All-Reduce from Vowpal Wabbit (VW), Langford et al.
  - Reducers talk to each other so that precise gradient can be computed by aggregating all computations from each partition (reducer).

Ongoing Work at LinkedIn and Future Challenges

- Large scale statistical models for ad creative CTR prediction and ad creative ranking
- Explore-exploit for better ad serving strategy
- Incorporating social network signals into user profile (for cold start)

Conclusion

- Generalized Matrix Factorization (GMF) framework to handle cold-start, feature selection and non-linearity simultaneously
- User factors from Parallelized GMF can serve as user profile for OLR, which gives state-of-the-art performance
- A new way to model item-item similarity for CTR prediction

Thank You!

Our Open Source Package for
matrix factorization models:
https://github.com/yahoo/Latent-Factor-Models

Questions or feedback: liang.zhang.stat@gmail.com

Bibliography

Agarwal, D. and Chen, B. (2009). Regression-based latent factor models. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 19–28. ACM.

Agarwal, D., Chen, B., and Elango, P. (2010). Fast online learning through offline initialization for time-sensitive recommendation. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 703–712. ACM.

Agarwal, D., Zhang, L., and Mazumder, R. (2011). Modeling item–item similarities for personalized recommendations on Yahoo! front page. The Annals of Applied Statistics 5, 3, 1839–1875.

Bell, R., Koren, Y., and Volinsky, C. (2007). Modeling relationships at multiple scales to improve accuracy of large recommender systems. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 95–104. ACM.

Zhang, L., Agarwal, D., and Chen, B. (2011). Generalizing matrix factorization through flexible regression priors. In Proceedings of the fifth ACM conference on Recommender systems, 13–20. ACM.

R Khanna, L Zhang, D Agarwal and Chen, B. (2012). Parallel Matrix Factorization for Binary Response. In Arxiv.org.