Collaborative Filtering, Missing Data, and Ranking

CSC2535

Richard Zemel

March 13, 2013

Contents:

Introduction

Notation

Theory Of Missing Data
- Factorizations
- MCAR
- MAR
- NMAR
- Inference and Learning

Multinomial Models
- Multinomial Mixture
- Multinomial Mixture/CPT-v

Collaborative Filtering Expts.
- Yahoo! Data
- Jester Data
- Results

Collaborative Ranking

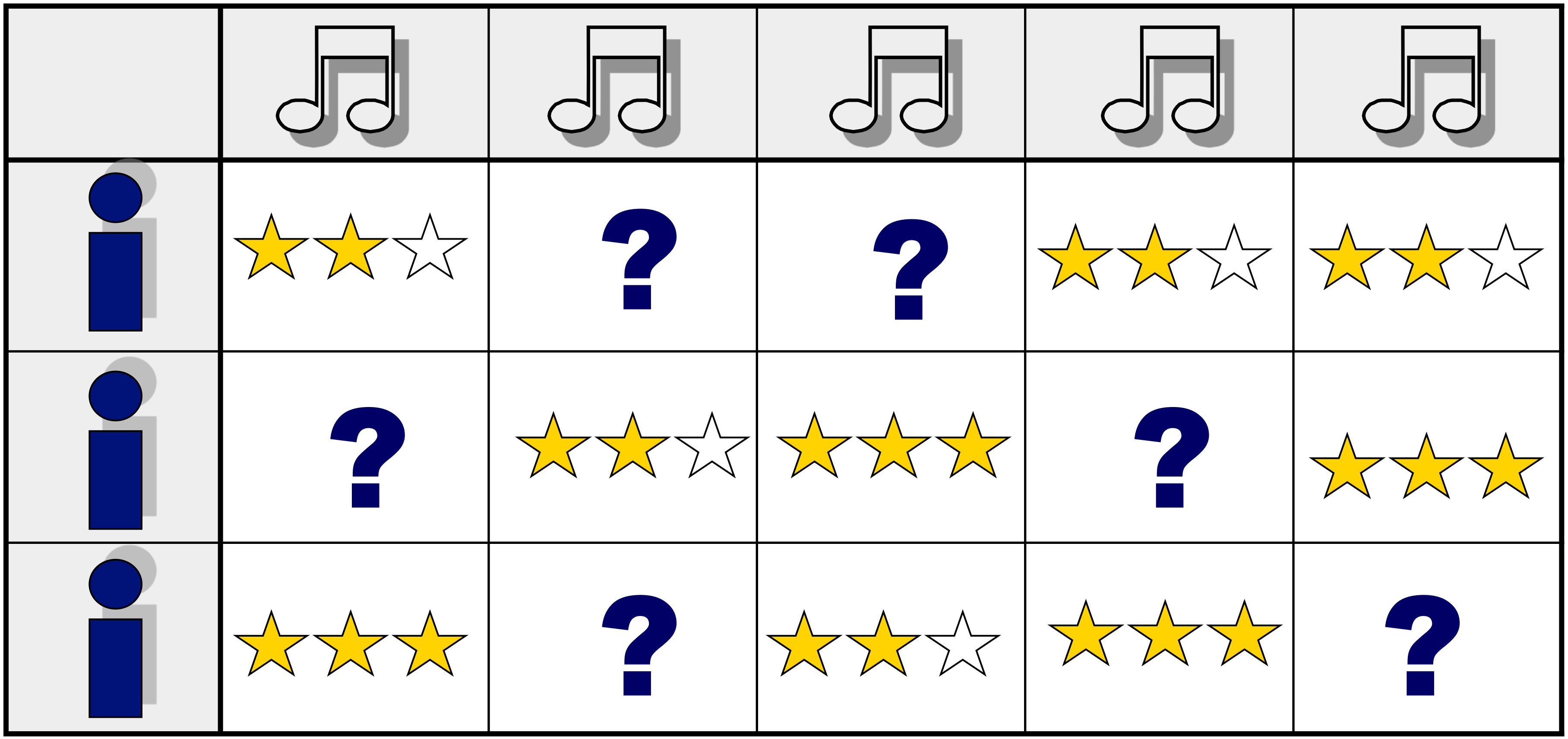
Introduction

Introduction: Collaborative Filtering

Collaborative filtering – users assign ratings to items → system uses information from all users to recommend previously unseen items that a user might like

One approach to recommendation: predict ratings for all unrated items, recommend highest predicted ratings

Collaborative Filtering:
Collaborative Prediction Problem



Introduction: Missing Data

Critical assumption: missing ratings are missing at random

One way to violate: value of variable affects probability value will be missing – bias in observed ratings, and hence learned parameters

Also complementary bias in standard testing procedure – distribution of observed data different from distribution of complete data, so estimated error on observed test data poor estimate of complete data error

Introduction: Survey Sampling Example

Data Variables	X	Answers to questions.
Response Variables	R	Did the respondent answer the question?

Introduction: Medical Diagnosis Example

Latent Variables	Z	Diseases
Data Variables	X	Symptoms
Response Variables	R	Was a test performed to check for the symptom?

Introduction: Recommender Systems Example

Latent Variables — Z — Preferences and Tastes

Data Variables — X — Ratings or Purchase History

Response Variables — R — Did the user rate or buy the item?

Introduction: Basic Notation

N	Number of data cases.
D	Number of data dimensions.
C	Number of classes.
V	Number of multinomial values.
K	Number of clusters or hidden units.

Introduction: Notation for Missing Data

$\mathbf{x}_n$	0.1 | 0.9 | 0.2 | 0.7 | 0.3	Data Vector
$\mathbf{r}_n$	1 | 0 | 0 | 1 | 1	Response Vector
$\mathbf{o}_n$	1 | 4 | 5	Observed Dimensions
$\mathbf{m}_n$	2 | 3	Missing Dimensions
$\mathbf{x}_n^{\mathbf{o}_n}, \mathbf{x}_n^{o}$	0.1 | 0.7 | 0.3	Observed Data
$\mathbf{x}_n^{\mathbf{m}_n}, \mathbf{x}_n^{m}$	0.9 | 0.2	Missing Data

Contents:

Introduction

Notation

Theory Of Missing Data

- Factorizations
- MCAR
- MAR
- NMAR
- Inference and Learning

Multinomial Models

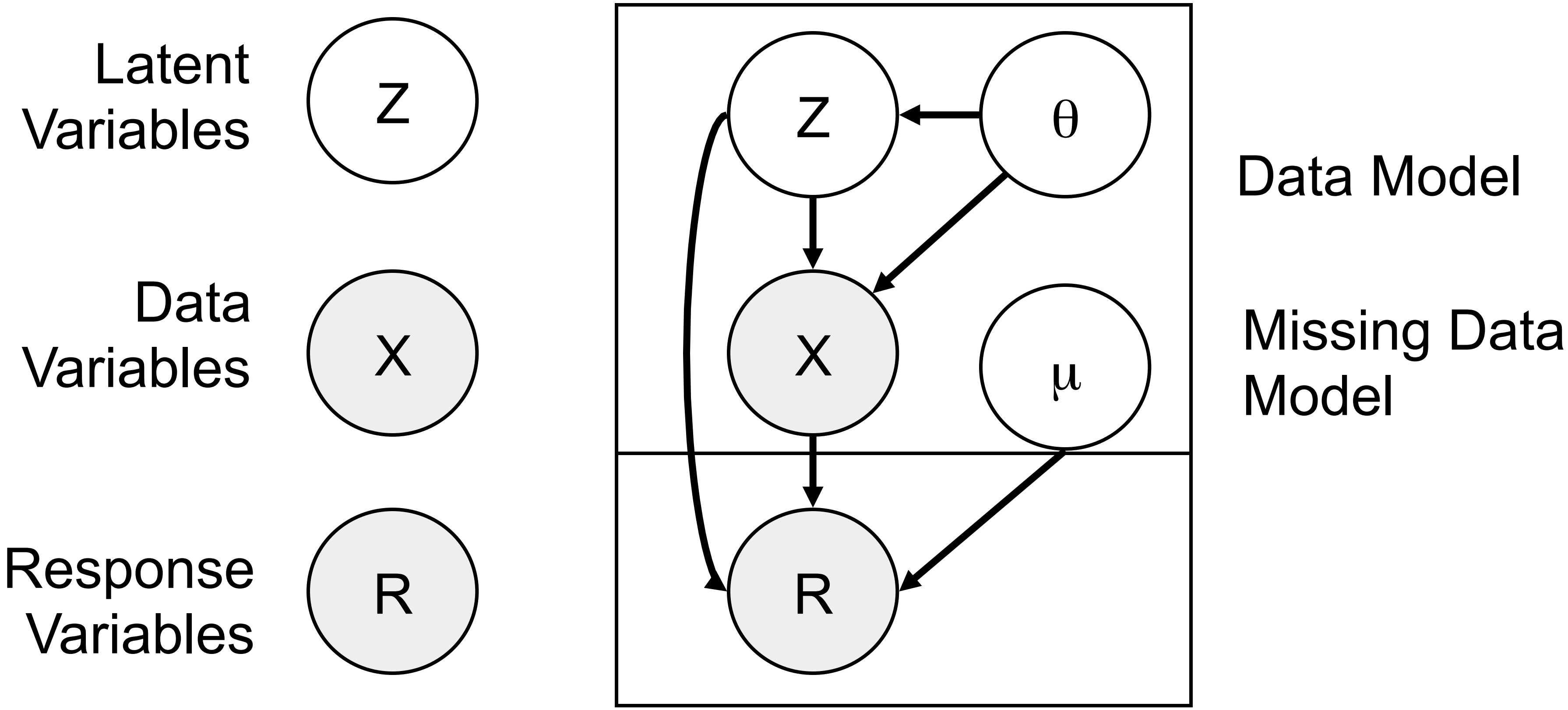
- Multinomial Mixture
- Multinomial Mixture/CPT-v

Collaborative Filtering Expts.

- Yahoo! Data
- Jester Data
- Results

Collaborative Ranking

Theory of Missing Data: Generative Process



Theory of Missing Data: Factorizations

Data/Selection Model Factorization:

$$P(\mathbf{x},\mathbf{r},\mathbf{z}|\theta,\mu) = P(\mathbf{r}|\mathbf{x},\mathbf{z},\mu)P(\mathbf{x},\mathbf{z}|\theta)$$

- The probability of selection depends on the true values of the data variables and latent variables.

Pattern Mixture Model Factorization:

$$P(\mathbf{x},\mathbf{r},\mathbf{z}|\vartheta,\nu) = P(\mathbf{x},\mathbf{z}|\mathbf{r},\vartheta)P(\mathbf{r}|\nu)$$

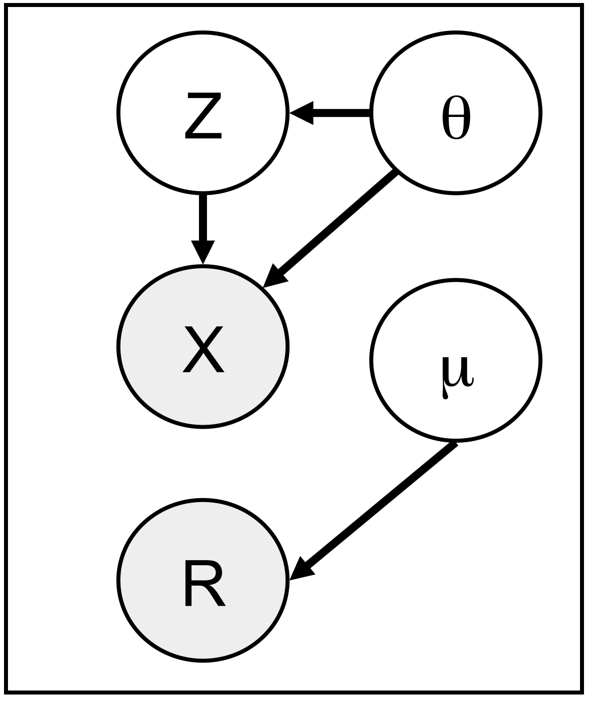
- Each response vector defines a different pattern, and each pattern has a different distribution over the data.

Theory of Missing Data: Classification

Missing Completely at Random:

• Response probability is independent of data variables and latent variables.

$$P(\mathbf{r}|\mathbf{x}, \mathbf{z}, \mu) = P(\mathbf{r}|\mu)$$



MCAR Examples:

Send questionnaires to a random subset of the population or use random digit dialing.

Theory of Missing Data: Classification

Missing at Random:

• Typically written in a short-hand form that looks like a statement of probabilistic independence:

$$P(\mathbf{r}|\mathbf{x}, \mathbf{z}, \mu) = P(\mathbf{r}|\mathbf{x}^o, \mu)$$

• MAR is actually a different type of condition that requires a particular set of symmetries hold in $P(r|x,z,\mu)$:

$$P(\mathbf{r}|\mathbf{x}^{o(\mathbf{r})}, \mathbf{x}^{m(\mathbf{r})}, \mathbf{Z}, \mu) = f(\mathbf{r}, \mathbf{x}^{o(\mathbf{r})}, \mu) \text{ ... for all } \mathbf{x}^m$$

Theory of Missing Data: Classification

Missing at Random Examples:

Respondents are not required to provide information about their employer if they are not currently employed.

Doctor only orders test B if the result of test A was negative. If result of test A is positive, result for test B is missing.

Theory of Missing Data: Classification

What Does it mean to be Missing at Random?

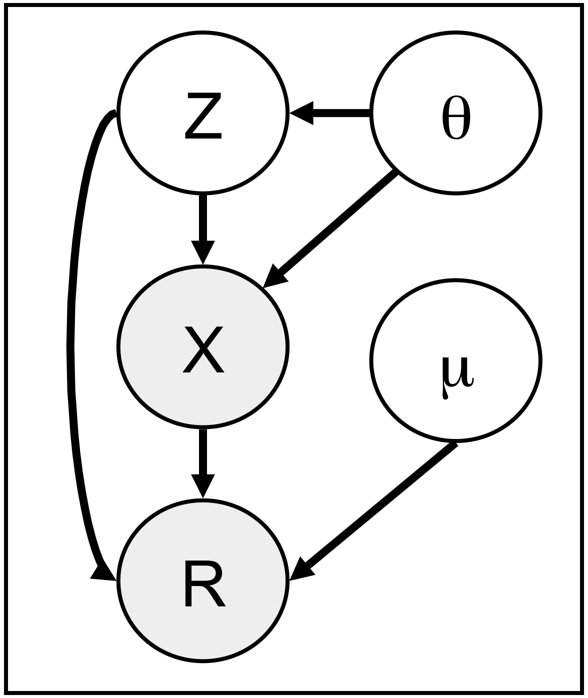
- MAR is ***not*** a statement of independence between random variables. MAR requires that particular symmetries hold so that P(R=r|X=x) can be determined from observed data only.

$X \backslash R$	$0\ 0$	$0\ 1$	$1\ 0$	$1\ 1$
$0\ 0$	α	β	γ	$1-\alpha-\beta-\gamma$
$0\ 1$	α	δ	γ	$1-\alpha-\delta-\gamma$
$1\ 0$	α	β	λ	$1-\alpha-\beta-\lambda$
$1\ 1$	α	δ	λ	$1-\alpha-\delta-\lambda$

Theory of Missing Data: Classification

Not Missing at Random:

• Allows for arbitrary dependence of response probabilities on missing data values and latent variables:



$$P(\mathbf{r}|\mathbf{x}, \mathbf{z}, \mu)$$ **No Simplifications**

An Easy Way to Violate MAR:

• Let the probability that a data variable is observed depend on the value of that data variable.

Theory of Missing Data: Classification

Not Missing at Random Examples:

Snowfall reading is likely to be missing if weather station is covered with snow.

Participants in a longitudinal health study for a heart medication may die of a heart attack during the study.

Users are more likely to rate or buy items they like than items they don’t like.

Theory of Missing Data: Inference

MCAR/MAR Posterior:

$$
\begin{aligned}
P(\theta|\mathbf{x}^o,\mathbf{r}) &\propto \int\int\int P(\mathbf{x},\mathbf{z}|\theta)P(\mathbf{r}|\mathbf{x},\mathbf{z},\mu)P(\theta|\omega)P(\mu|\eta)d\mu dZ d\mathbf{x}^m \\
&\propto \int f(\mathbf{r},\mathbf{x},\mu)P(\mu|\eta)d\mu \cdot \int\int P(\mathbf{x},\mathbf{z}|\theta)P(\theta|\omega)dZ d\mathbf{x}^m \\
&\propto P(\mathbf{x}^o|\theta)P(\theta|\omega)
\end{aligned}
$$

- When MCAR or MAR holds, the posterior can be greatly simplified. Inference for θ does not depend on r, μ, or η. The missing data can be *ignored*.

Theory of Missing Data: Inference

NMAR Posterior:

$$P(\theta|\mathbf{x}^o, \mathbf{r}) \propto \int\int\int P(\mathbf{x}, \mathbf{z}|\theta)P(\mathbf{r}|\mathbf{x}, \mathbf{z}, \mu)P(\theta|\omega)P(\mu|\eta)d\mu dZ d\mathbf{x}^m$$

- When MAR fails to hold, the posterior does not simplify.

- Basing inference on the observed data posterior and ignoring the missing data model leads to provably biased inference for data model parameters.

Contents:

Introduction

Notation

Theory Of Missing Data

- Factorizations
- MCAR
- MAR
- NMAR
- Inference and Learning

Multinomial Models

- Multinomial Mixture
- Multinomial Mixture/CPT-v

Collaborative Filtering Expts.

- Yahoo! Data
- Jester Data
- Results

Collaborative Ranking

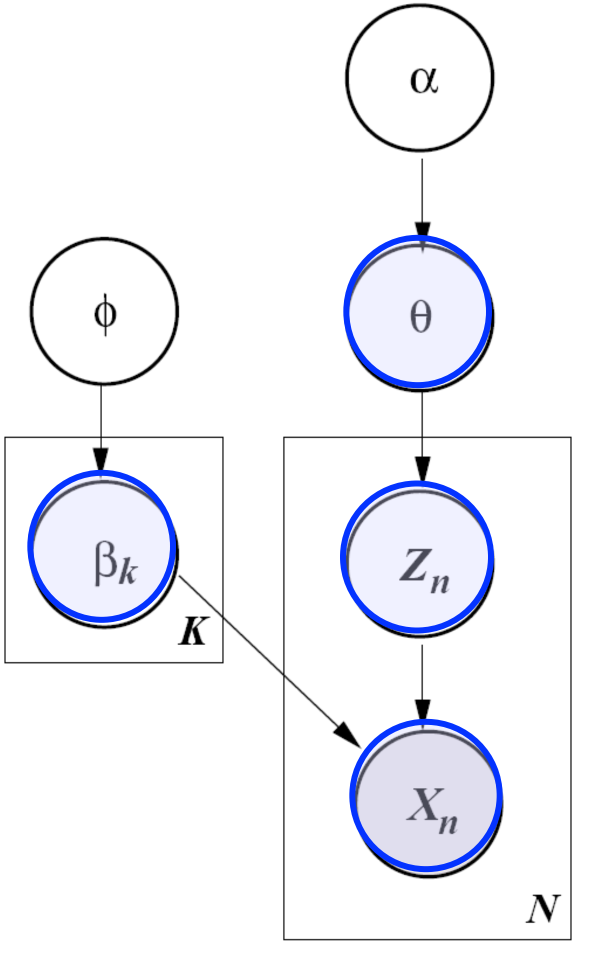
Multinomial Models: Mixture

Probability Model:

$$P(Z_n = k|\theta) = \theta_k$$

$$P(\mathbf{X}_n = \mathbf{x}_n|Z_n = k, \beta) = P(\mathbf{x}_n|\beta_k)$$

$$P(\theta, \beta|\alpha, \phi) = P(\theta|\alpha) \prod_k P(\beta_k|\phi)$$



Properties:

- Allows for a fixed, finite number of clusters.
- In the multinomial mixture, $P(x_n|\beta_k)$ is a product of discrete distributions. The prior on β and θ is Dirichlet.

Multinomial Models: Mixture

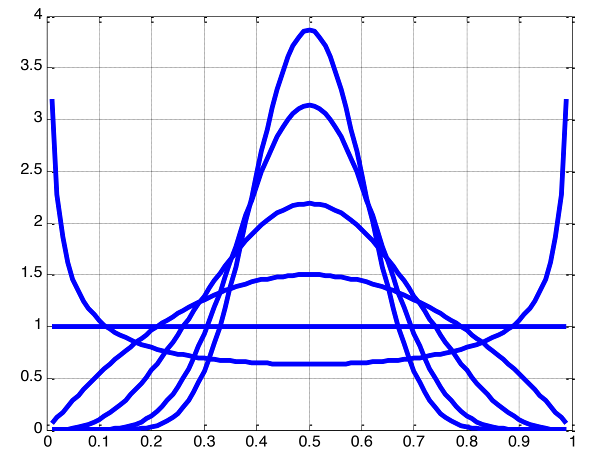
Dirichlet Distribution:

Bayesian mixture modeling becomes much easier when conjugate priors are used for the model parameters. The conjugate prior for the mixture proportions θ is the Dirichlet distribution.

$$P(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$

$$E[\theta_k|\alpha] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}$$

$$P(\theta|\alpha, \mathbf{z}) = \frac{\Gamma(N + \sum_k \alpha_k)}{\prod_k \Gamma(C_k + \alpha_k)} \prod_k \theta_k^{C_k + \alpha_k - 1}$$



Multinomial Models: Mixture

MAP EM Algorithm:

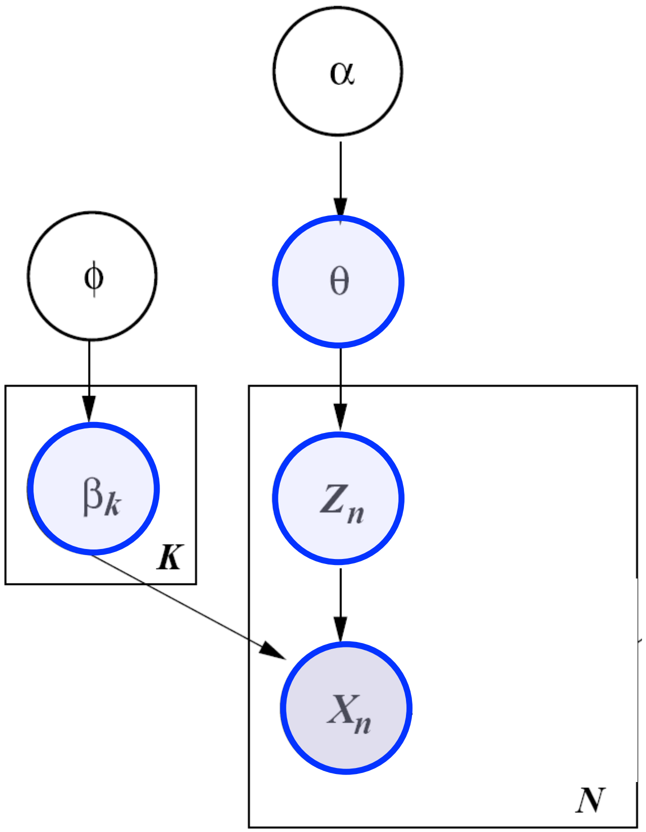
$$\text{E-Step: } q_n(k) \leftarrow \frac{\theta_k \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk}^{[r_{dn}=1][x_{dn}=v]}}{\sum_{k'=1}^{K} \theta_{k'} \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk'}^{[r_{dn}=1][x_{dn}=v]}}$$

$$\text{M-Step: } \theta_k \leftarrow \frac{\alpha_k - 1 + \sum_{n=1}^{N} q_n(k)}{N - K + \sum_{k=1}^{K} \alpha_k}$$

$$\beta_{vdk} \leftarrow \frac{\phi_{vdk} - 1 + \sum_{n=1}^{N} q_n(k)[r_{dn} = 1][x_{dn} = v]}{\sum_{n=1}^{N} q_n(k)[r_{dn} = 1] - V + \sum_{v=1}^{V} \phi_{vdk}}$$

Multinomial Models: Mixture/CPT-v

Probability Model:



Multinomial Models: Mixture/CPT-v

Probability Model:

$$P(\theta|\alpha) = \mathcal{D}(\theta|\alpha)$$

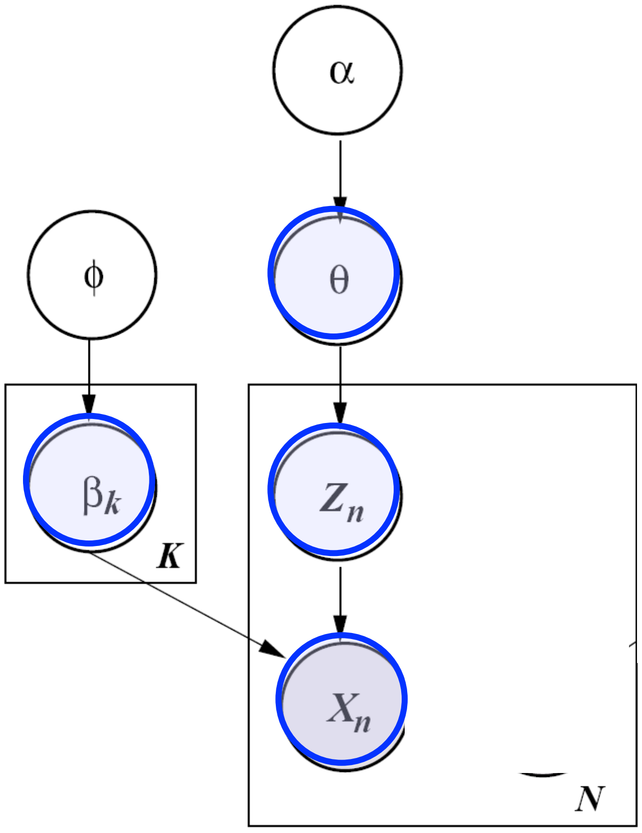
$$P(\beta|\phi) = \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{D}(\beta_{dk}|\phi_{dk})$$

$$P(Z_n = k|\theta) = \theta_k$$

$$P(\mathbf{X} = \mathbf{x}_n|Z_n = k, \beta) = \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk}^{[x_{dn}=v]}$$

$$P(\mu|\xi) = \prod_{v} \mathcal{B}(\mu_v|\xi_v)$$

$$P(\mathbf{R} = \mathbf{r}_n|\mathbf{X} = \mathbf{x}_n, \mu) = \prod_{d=1}^{D} \prod_{v=1}^{V} \mu_v^{[r_{dn}=1][x_{dn}=v]}(1 - \mu_v)^{[r_{dn}=0][x_{dn}=v]}$$



Multinomial Models: Mixture/CPT-v

MAP EM Algorithm (E-Step):

$$q_n(k) = P(z_n = k | \mathbf{x}_n^o, \mathbf{r}_n, \theta, \beta, \mu)$$

$$= \frac{\theta_k \prod_{d=1}^{D} \left(\prod_{v=1}^{V} (\beta_{vdk}\mu_v)^{[x_{dn}=v]} \right)^{[r_{dn}=1]} \left(\sum_{v=1}^{V} \beta_{vdk}(1-\mu_v) \right)^{[r_{dn}=0]}}{\sum_{k=1}^{K} \theta_k \prod_{d=1}^{D} \left(\prod_{v=1}^{V} (\beta_{vdk}\mu_v)^{[x_{dn}=v]} \right)^{[r_{dn}=1]} \left(\sum_{v=1}^{V} \beta_{vdk}(1-\mu_v) \right)^{[r_{dn}=0]}}$$

$$q_n(k, v, d) = P(z_n = k, x_{dn} = v | \mathbf{x}_n^o, \mathbf{r}_n, \theta, \beta, \mu)$$

$$= q_n(k) \left(\frac{\mu_v \beta_{vdk}}{\sum_{v'=1}^{V} \mu_{v'} \beta_{v'dk}} \right)^{[r_{dn}=1]} \left(\frac{(1-\mu_v)\beta_{vdk}}{\sum_{v'=1}^{V} (1-\mu_{v'})\beta_{v'dk}} \right)^{[r_{dn}=0]}$$

Multinomial Models: Mixture/CPT-v

MAP EM Algorithm (M-Step):

$$\theta_k = \frac{\alpha_k - 1 + \sum_{n=1}^N q_n(k)}{N - K + \sum_{k=1}^K \alpha_k}$$

$$\beta_{vdk} = \frac{\phi_{vdk} - 1 + \sum_{n=1}^N q_n(k)[r_{dn} = 1][x_{dn} = v] + q_n(k,v,d)[r_{dn} = 0]}{\sum_{n=1}^N q_n(k) - V + \sum_{v=1}^V \phi_{vdk}}$$

$$\mu_v = \frac{\xi_{1v} - 1 + \sum_{n=1}^N \sum_{d=1}^D [r_{dn} = 1][x_{dn} = v]}{\xi_{1v} + \xi_{0v} - 2 + \sum_{n=1}^N \sum_{d=1}^D [r_{dn} = 1][x_{dn} = v] + q_n(v,d)[r_{dn} = 0]}$$

DEMO

Multinomial Mixture Learning With Random and Non-Random Missing Data

Other Models for Missing Data:

- K-Nearest Neighbors
- Probabilistic Principal Components Analysis
- Factor Analysis
- Mixtures of Gaussians
- Mixtures of PPCA/FA
- Probabilistic Matrix Factorization
- Maximum Margin Matrix Factorization
- Conditional Restricted Boltzmann Machines

Contents:

Introduction

Notation

Theory Of Missing Data

- Factorizations
- MCAR
- MAR
- NMAR
- Inference and Learning

Multinomial Models

- Multinomial Mixture
- Multinomial Mixture/CPT-v

Collaborative Filtering Expts.

- Yahoo! Data
- Jester Data
- Results

Collaborative Ranking

Collaborative Filtering:
Collaborative Prediction Problem

	♫	♫	♫	♫	♫
	★★☆	?	?	★★☆	★★☆
	?	★★☆	★★★	?	★★★
	★★★	?	★★☆	★★★	?

Collaborative Filtering : Yahoo!

Data was collected through an online survey of Yahoo! Music LaunchCast radio users.

- 1000 songs selected at random.
- Users rate 10 songs selected at random from 1000 songs.
- Answer 16 questions.
- Collected data from 35,000+ users.

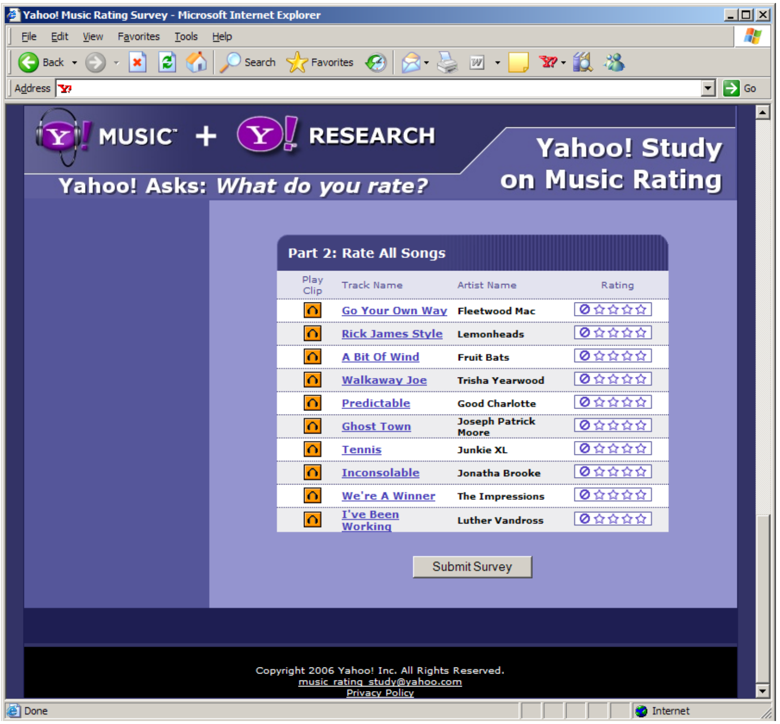
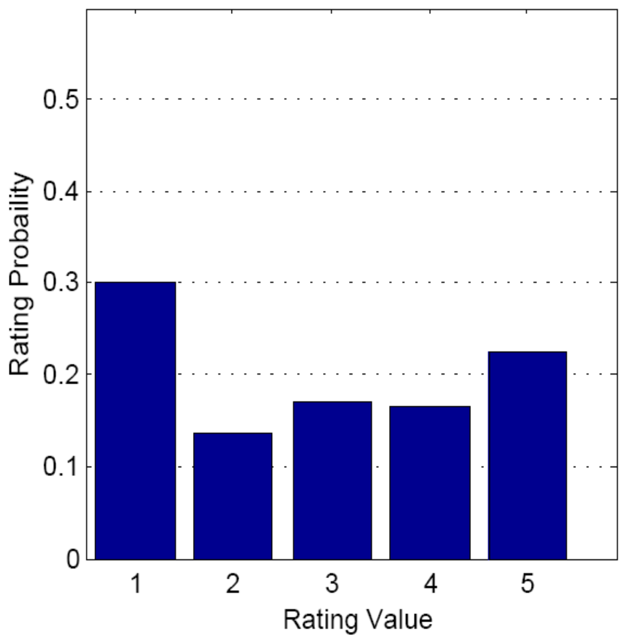


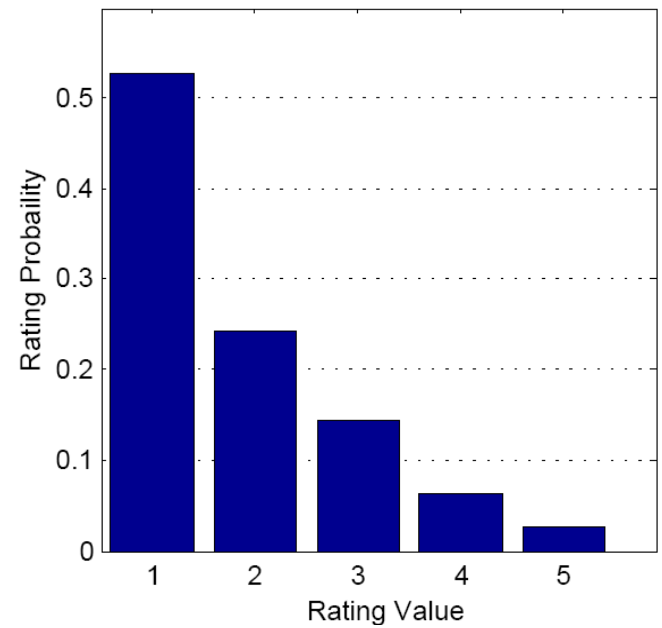
Image copyright Yahoo! Inc. 2006. Used with permission.

Collaborative Filtering: Yahoo!

User Selected

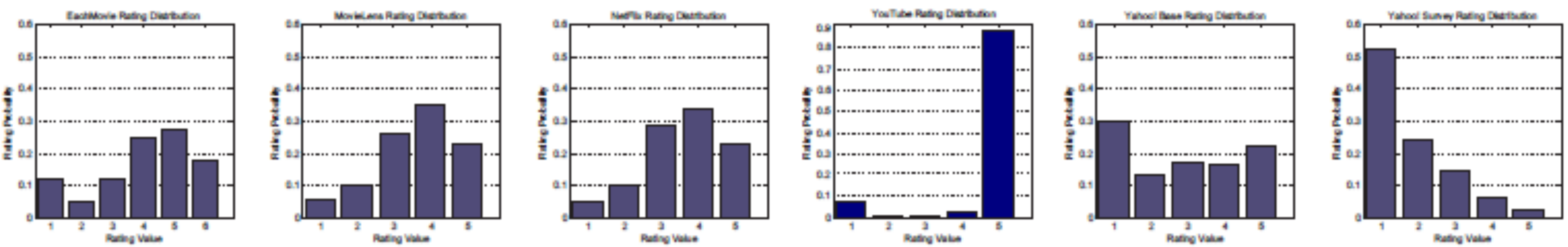


Randomly Selected



$$P_E^U(x=v|r=1) \quad \frac{\sum_i \sum_m [X_{im} = v][R_{im} = 1]}{\sum_i \sum_m [R_{im} = 1]} \quad P_E^R(x=v|r=1)$$

More Empirical Distributions

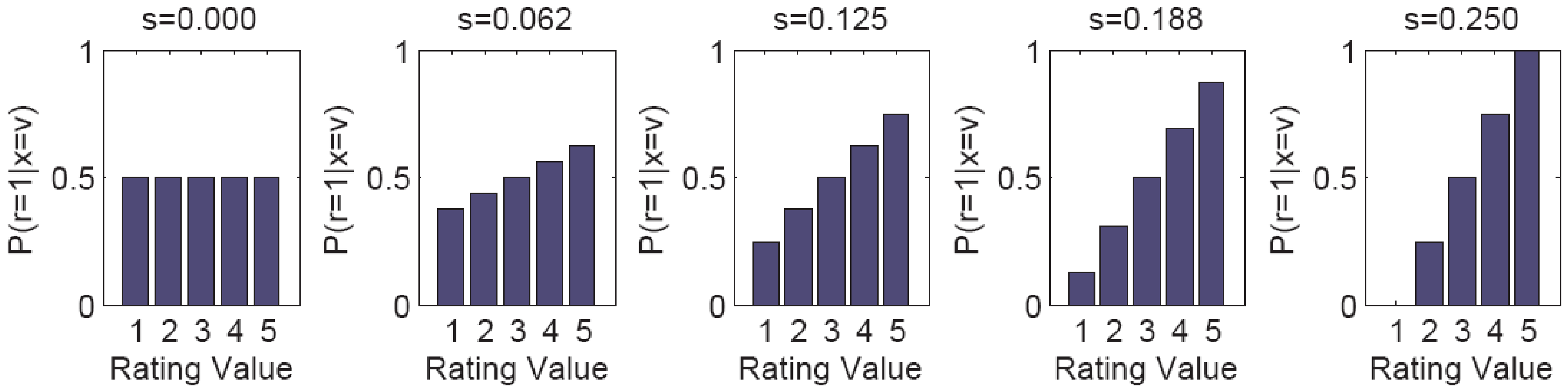


(a) EachMovie (b) MovieLens (c) NetFlix (d) YouTube (e) Yahoo! User (f) Yahoo! Random

Collaborative Filtering: Jester

Jester gauge set of 10 jokes used as complete data. Synthetic missing data was added.

- 15,000 users randomly selected
- Missing data model: $\mu_v(s) = s(v-3)+0.5$



Experimental Protocol

Randomly partition users into 5 blocks of 1080 users

Three sets of ratings:

1. Observed ratings – all but one of original ratings
2. Test ratings for user-selected – remaining one
3. Test ratings for randomly-selected – ten survey responses

User-selected items – same distribution as observed

Randomly selected test items -- MCAR

Experimental Protocol

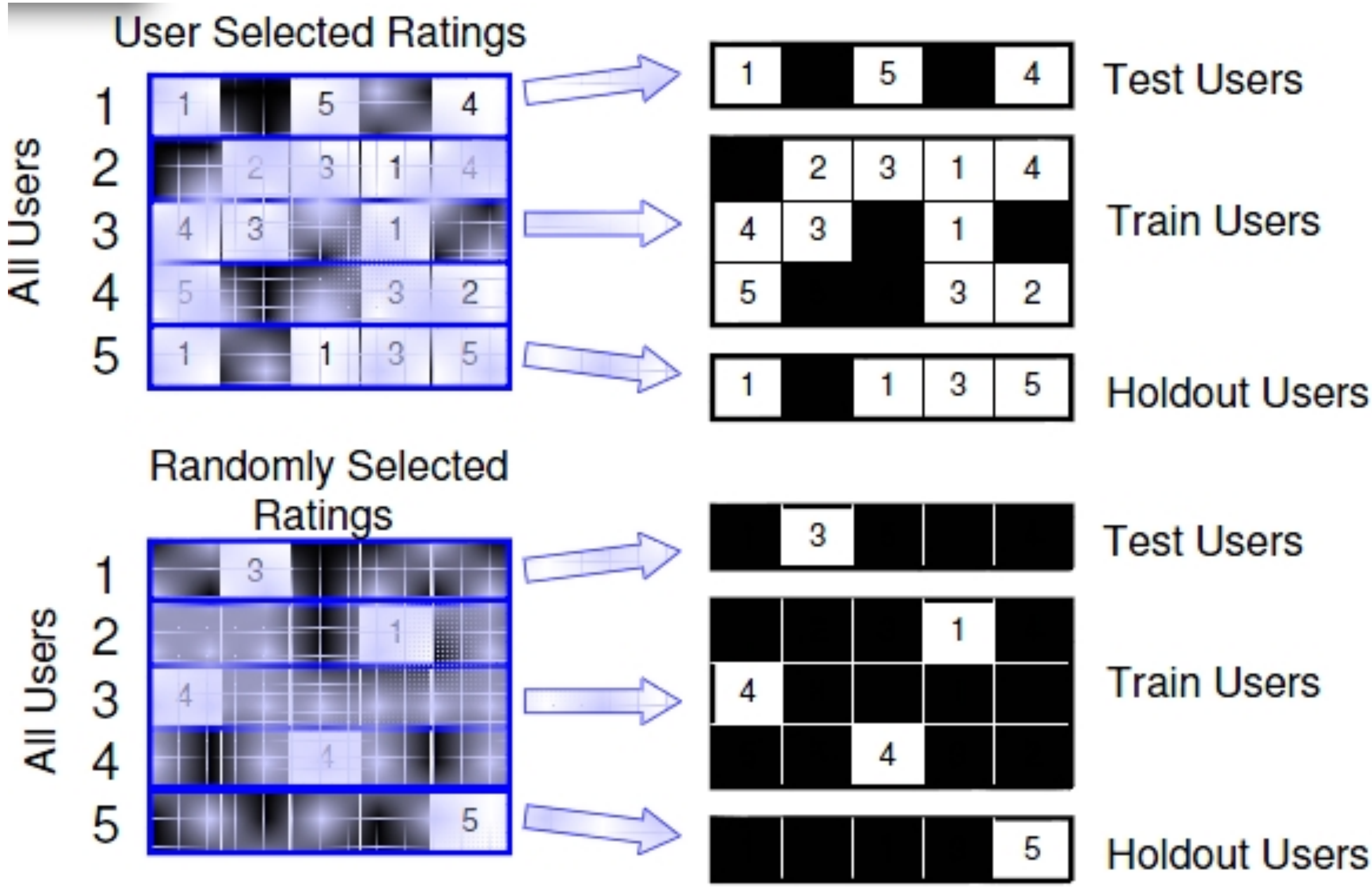
Weak Generalization

- Learn on training user observed ratings
- Evaluate on training user test ratings

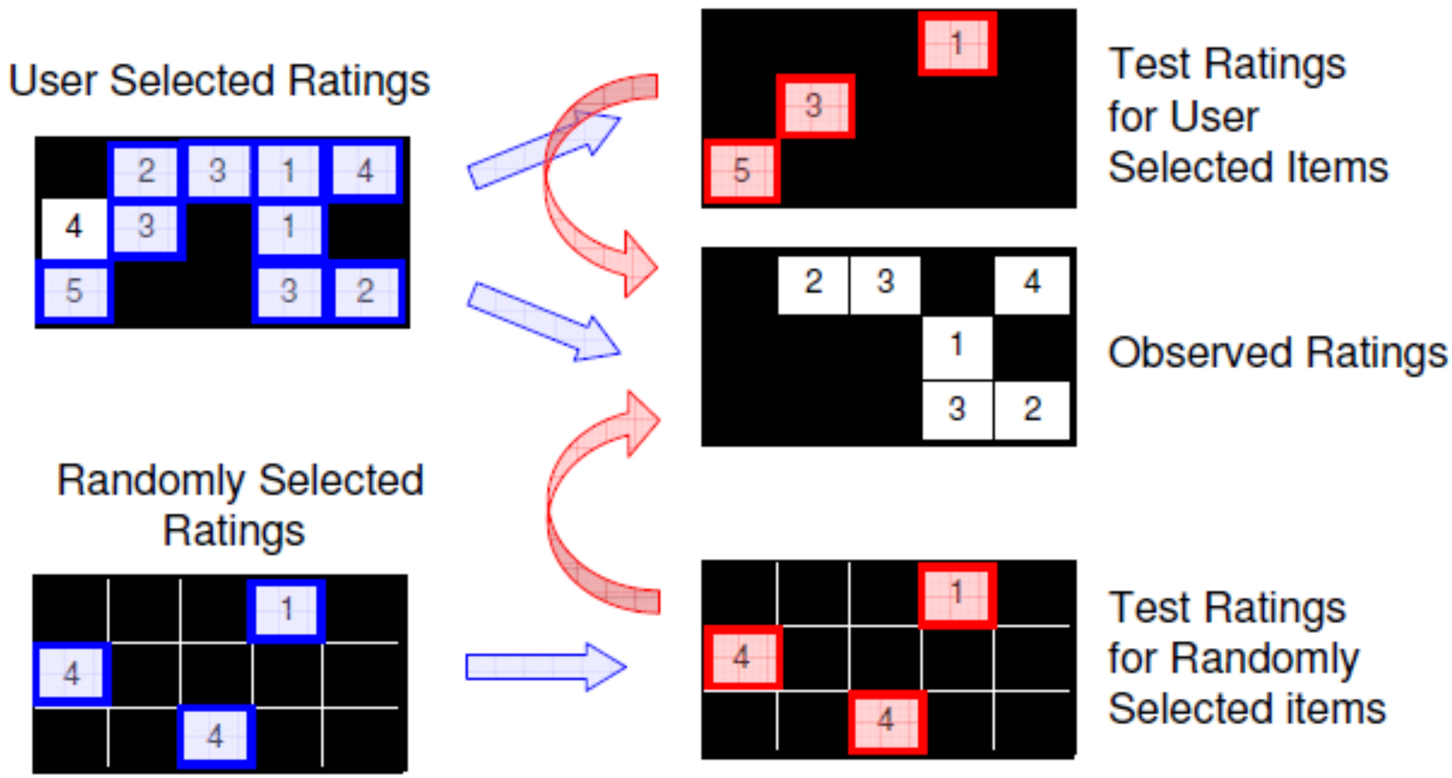
Strong Generalization

- Learn on training user observed ratings
- Evaluate on test user test ratings

Data Sets: User Splits

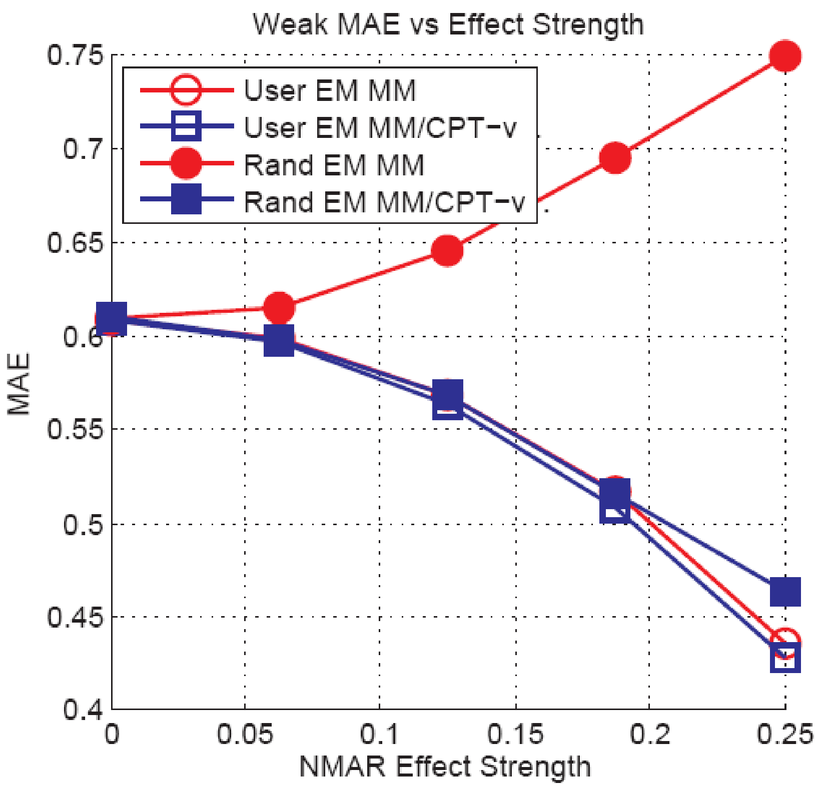


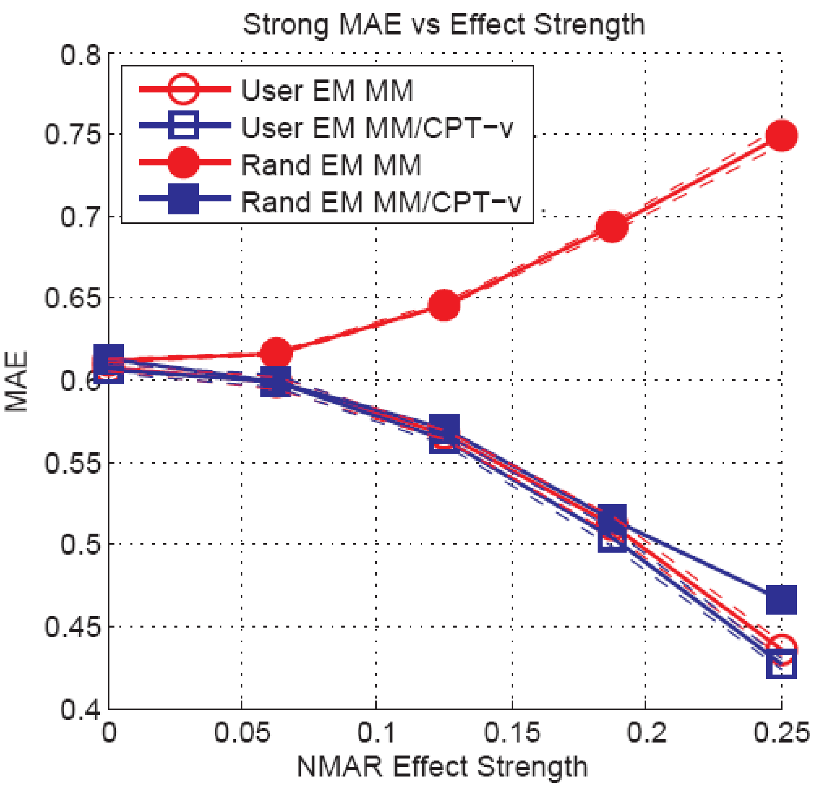
Data Sets: User Splits



Collaborative Filtering: Results

Jester Results: MM vs MM/CPT-v





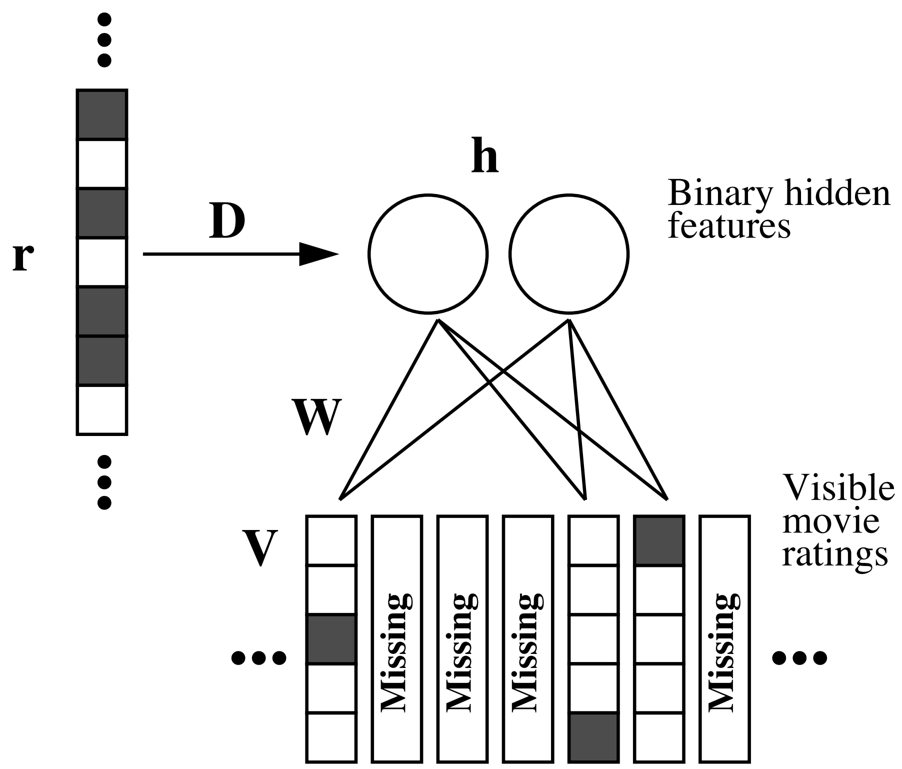
Collaborative Filtering: Baselines

Standard CF methods implicitly assume MAR

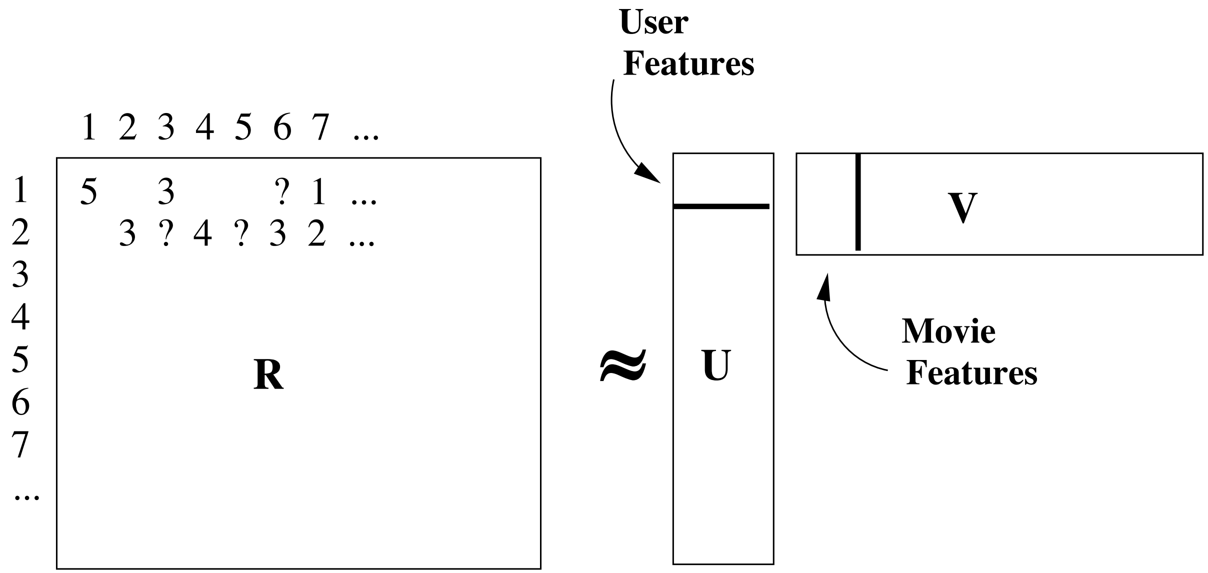
Here we compare to three other CF methods:

1. **Item-based K-nearest neighbor (iKNN)**
2. **cRBM**
3. **Matrix factorization**

Conditional RBM for CF



Probabilistic Matrix Factorization



- Let R_{ij} represent the rating of user i for movie j. The row and column vectors U_i and V_j represent user-specific and movie-specific latent feature vectors respectively.

- The model:

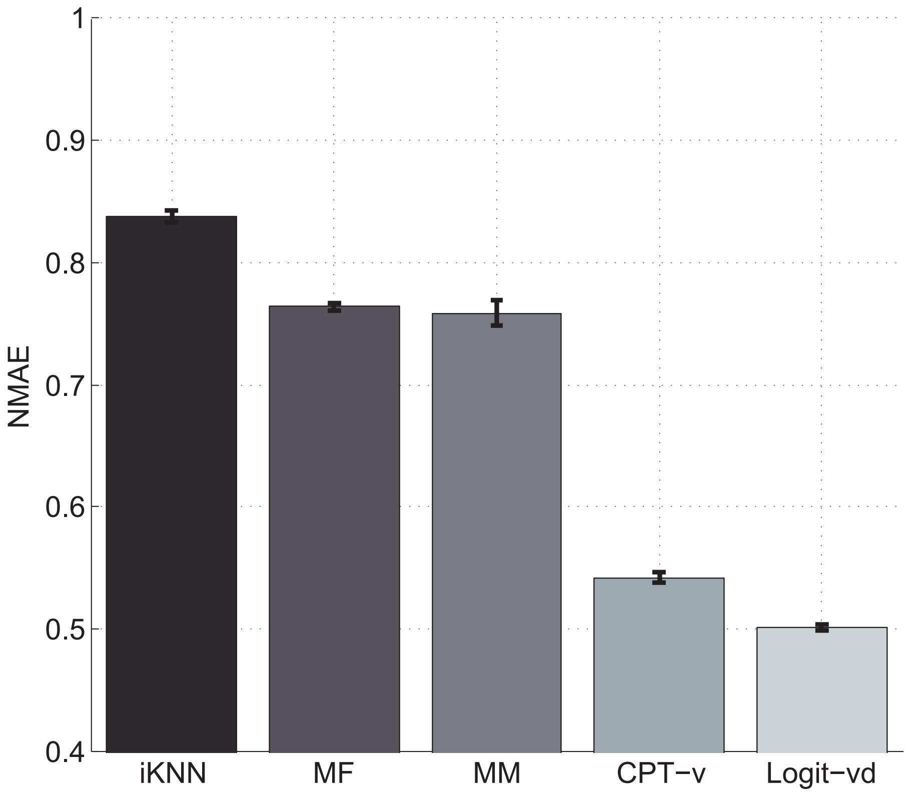
$$p(R_{ij}|U_i, V_j, \sigma^2) = \mathcal{N}(R_{ij}|U_i^T V_j, \sigma^2)$$

Collaborative Filtering: Results

Comparison of Results on Yahoo! Data

Method	Complexity	Rand MAE	User MAE
EM MM	1	0.7725 ± 0.0024	0.7626 ± 0.0077
EM MM/CPT-v	20	0.5431 ± 0.0012	0.6631 ± 0.0026
EM MM/Logit	5	0.5038 ± 0.0030	0.7029 ± 0.0186
EM MM/CPT-v+	5	0.4456 ± 0.0033	0.7235 ± 0.0059
MCMC DP	N/A	0.7624 ± 0.0063	0.5767 ± 0.0077
MCMC DP/CPT-v	N/A	0.5549 ± 0.0026	0.6670 ± 0.0071
MCMC DP/CPT-v+	N/A	0.4428 ± 0.0027	0.7537 ± 0.0026
CD cRBM	20	0.7179 ± 0.0025	0.5421 ± 0.0081
CD cRBM-v	1	0.4553 ± 0.0031	0.7501 ± 0.0066

Collaborative Filtering: Results
Comparison of Results on Yahoo! Data

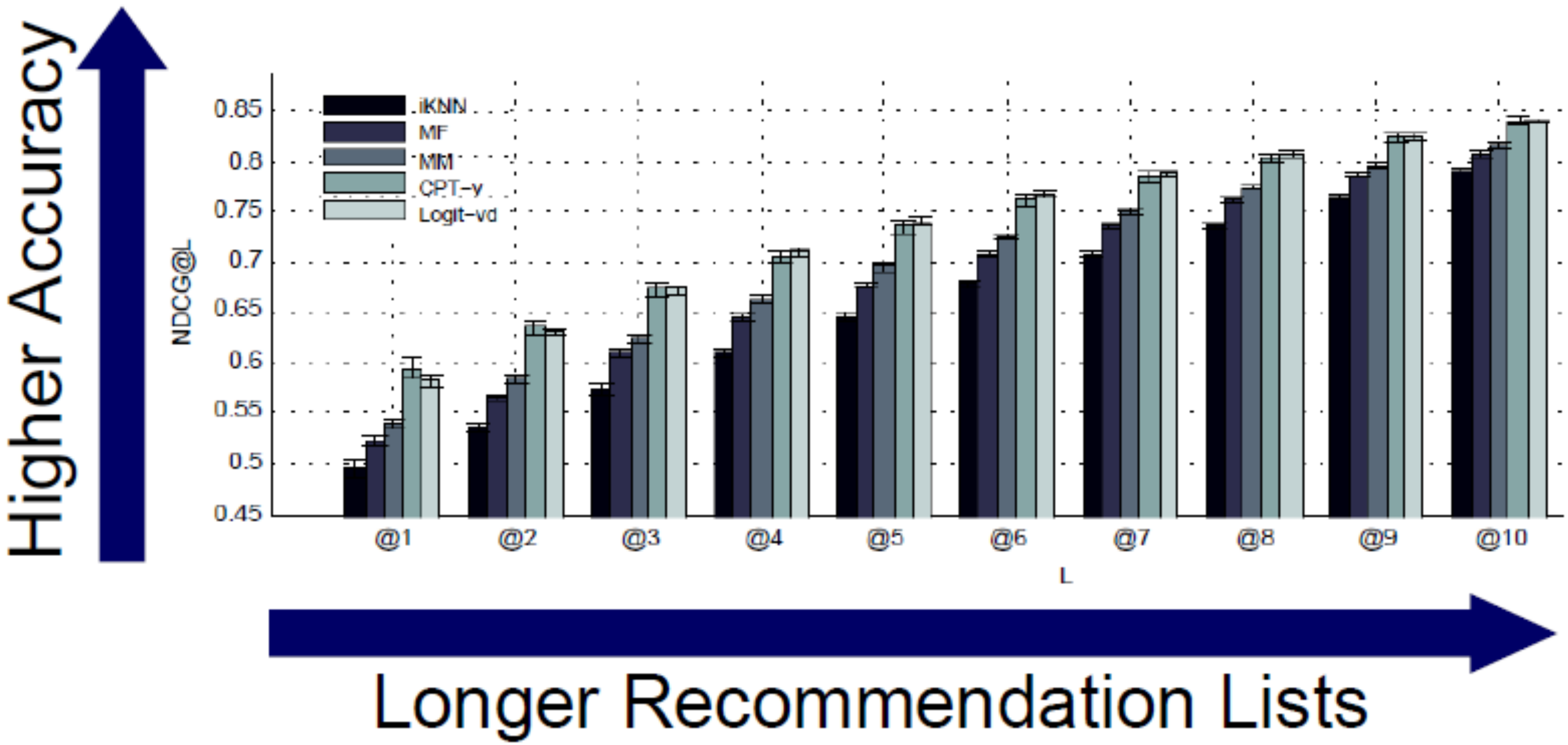


Application to Ranking

$$NDCG@L = \sum_{n=1}^{N} \frac{\sum_{l=1}^{L}(2^{x_{n\hat{\pi}(l,n)}} - 1)/\log(1+l)}{N\sum_{l=1}^{L}(2^{x_{n\pi(l,n)}} - 1)/\log(1+l)}$$

- $\widehat{x}_{ni}^{t}$: mean of posterior predictive distribution for test item i.
- $\widehat{\pi}(i,n)$: rank of test item i according to $\widehat{x}_{ni}^{t}$.
- $\pi(i,n)$: rank of test item i according to x_{ni}^{t} .

Ranking Results



Conclusions

In real recommender system data, the standard missing-at-random assumption is typically violated

Methods that include explicit non-random missing data model out-perform methods that assume MAR

In practice, the important task is collaborative *ranking*, not rating prediction

Our recent results show that combinations of neighbor- and model-based approaches to collaborative ranking permits scaling to large datasets

Collaborative Ranking Results

	10			20			30			40		
	N@1	N@3	N@5	N@1	N@3	N@5	N@1	N@3	N@5	N@1	N@3	N@5
MovieLens-1:												
UB	49.30	54.67	57.36	57.49	61.81	62.88	64.25	65.75	66.58	62.27	64.92	66.14
PMF-R(12K)	69.39	68.33	**68.65**	**72.50**	70.42	69.95	**72.77**	**72.23**	**71.55**	74.02	71.55	70.90
CO(240K)	67.28	66.23	66.59	71.82	**70.80**	**70.30**	71.60	71.15	70.58	71.43	71.64	71.43
WLT(17)	**70.96**	**68.25**	67.98	70.34	69.50	69.21	71.41	71.16	71.02	**74.09**	**71.85**	**71.52**
MovieLens-2:												
UB	67.62	68.23	68.74	71.29	70.78	70.87	72.65	71.98	71.90	73.33	72.63	72.42
PMF-R(500K)	70.12	69.41	69.35	70.65	70.04	70.09	72.22	71.48	71.43	72.18	71.60	71.55
CO(5M)	70.14	68.40	68.46	68.80	68.51	68.76	64.60	65.62	66.38	62.82	63.49	64.25
WLT(17)	**72.78**	**71.70**	**71.49**	**73.93**	**72.63**	**72.37**	**74.67**	**73.37**	**73.04**	**75.19**	**73.73**	**73.30**
Yahoo!:												
UB	57.20	55.29	54.31	64.29	61.48	60.16	66.82	63.83	62.42	68.97	65.89	64.50
PMF-R(1M)	52.86	51.98	51.53	63.93	62.42	**61.65**	66.82	65.41	64.61	69.46	68.05	67.21
CO(10M)	57.42	**56.88**	**56.46**	60.59	59.94	59.48	62.07	61.10	60.54	61.68	60.78	60.24
WLT(17)	**58.76**	55.20	53.53	**66.06**	**62.77**	61.21	**69.74**	**66.58**	**65.02**	**71.50**	**68.52**	**67.00**