Bayesian Co-clustering for Dyadic Data Analysis

Arindam Banerjee
banerjee@cs.umn.edu

Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

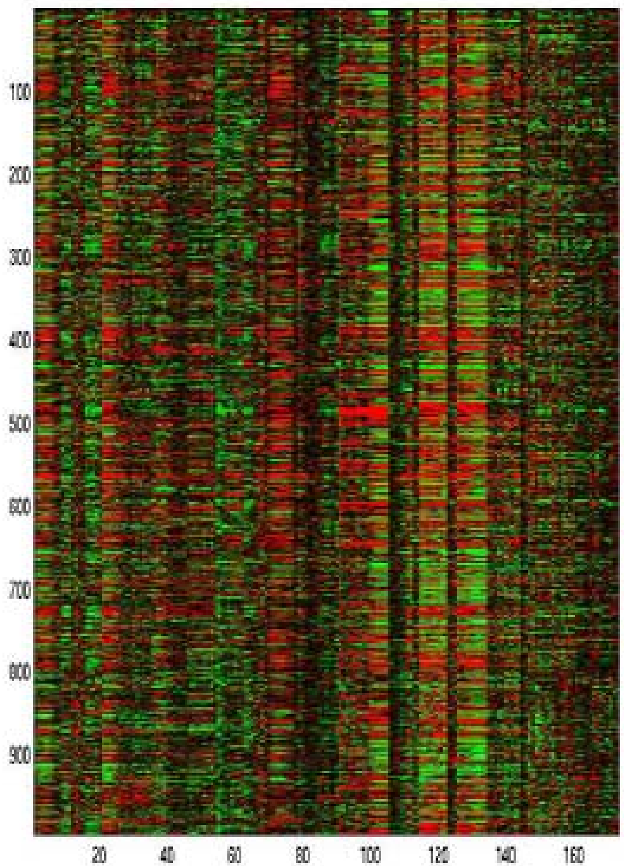
Workshop on Algorithms for Modern Massive Datasets (MMDS 2008)

Joint work with Hanhuai Shan

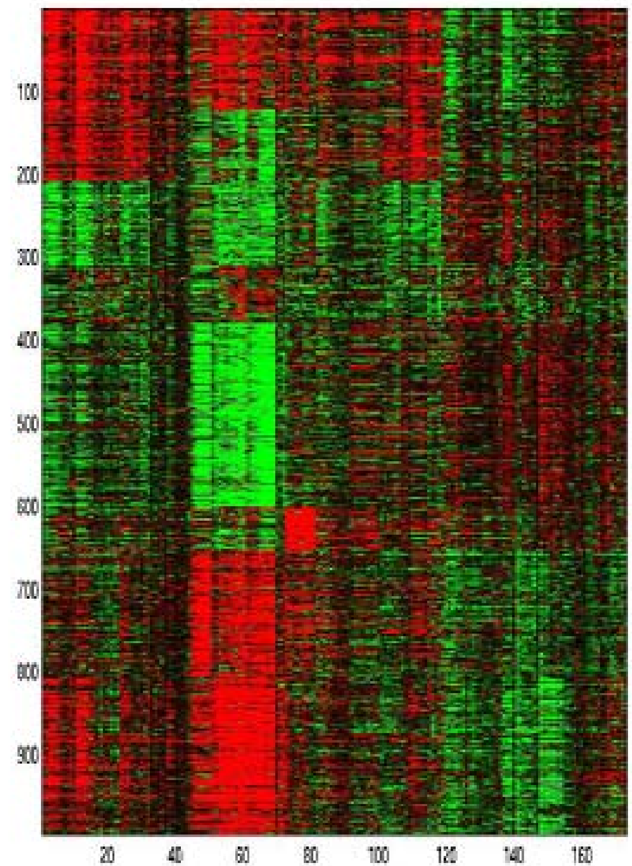
Introduction

- Dyadic Data
  - Relationship between two entities
- Examples
  - (Users, Movies): Ratings, Tags, Reviews
  - (Genes, Experiments): Expression
  - (Buyers, Products): Purchase, Ratings, Reviews
  - (Webpages, Advertisements): Click-through rate
- Co-clustering
  - Simultaneous clustering of rows and columns
  - Matrix approximation based on co-clusters
- Mixed membership co-clustering
  - Row/column has memberships in multiple row/column clusters
  - Flexible model, naturally handles sparsity

Example: Gene Expression Analysis



Original



Co-clustered

Co-clustering and Matrix Approximation

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix $\tilde{Z}$

$U, \hat{U}$	1	2
1	0	1
2	1	0
3	0	1
4	1	0
5	0	1

Row Clustering

$\times$

$\hat{U}, \hat{V}$	1	2	3
1	33.5	83.5	-23.3
2	-64.0	53.5	93.7

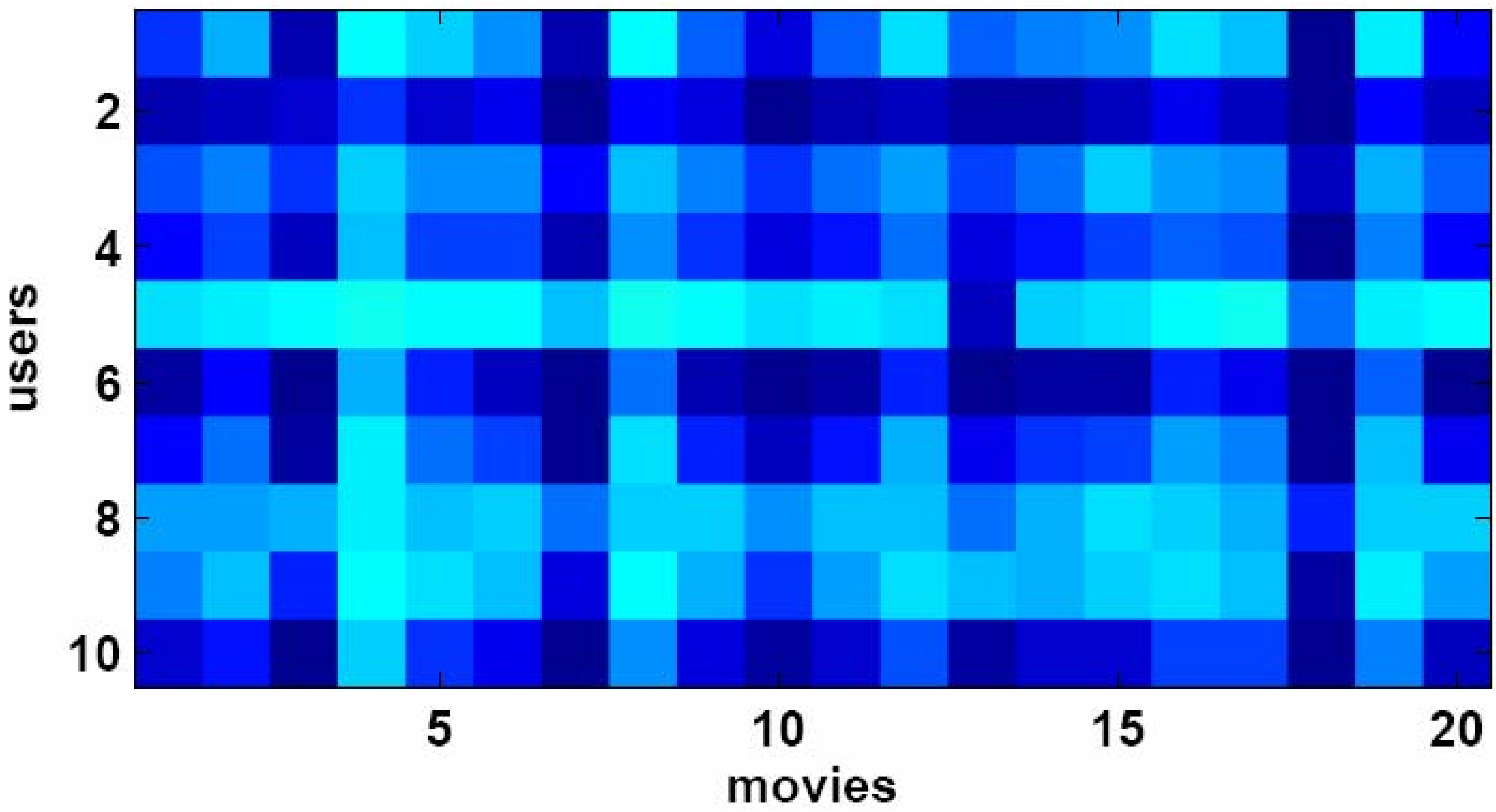
Low Parameter Matrix

$\times$

$\hat{V}, V$	1	2	3	4	5	6
1	1	0	1	0	0	0
2	0	1	0	0	1	0
3	0	0	0	1	0	1

Column Clustering

Example: Collaborative Filtering

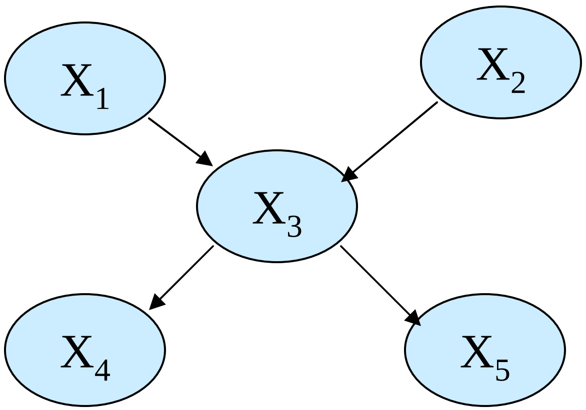


Related Work

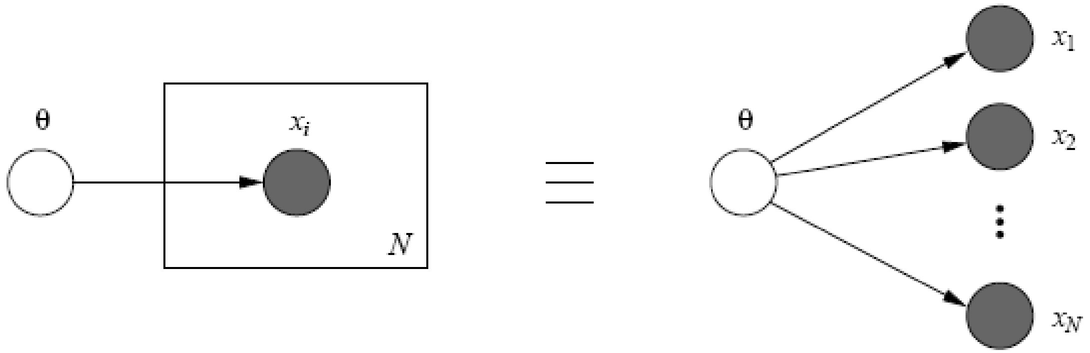
- Partitional co-clustering
  - Bi-clustering (Hartigan '72)
  - Bi-clustering of expression data (Cheng et al., '00)
  - Information theoretic co-clustering (Dhillon et al., '03)
  - Bregman co-clustering and matrix approximation (Banerjee et al., '07)
- Mixed membership models
  - Probabilistic latent semantic indexing (Hoffman, '99)
  - Latent Dirichlet allocation (Blei et al., '03)
- Bayesian relational models
  - Stochastic block structure (Nowicki et al, '01)
  - Infinite relational model (Kemp et al, '06)
  - Mixed membership stochastic block model (Airoldi et al, '07)

Background

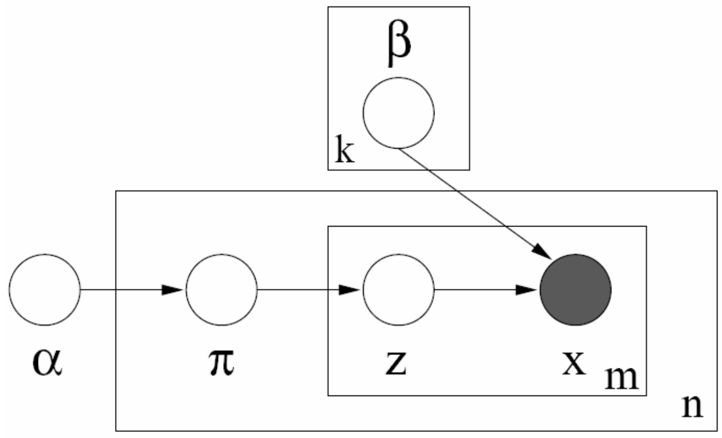
- Bayesian Networks

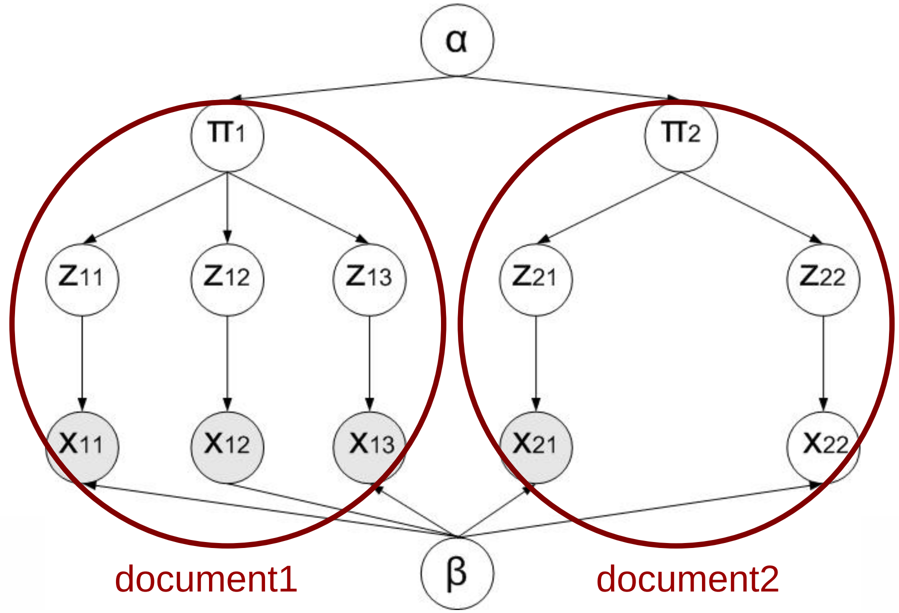


- Plates



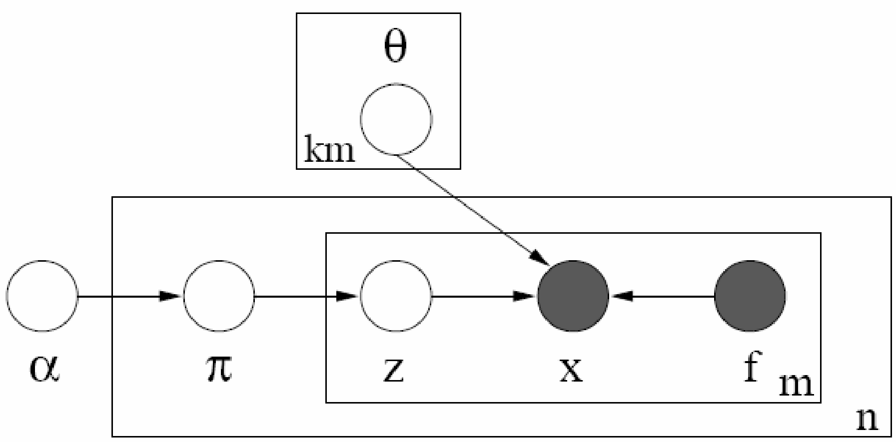
Latent Dirichlet Allocation (LDA) [BNJ’03]

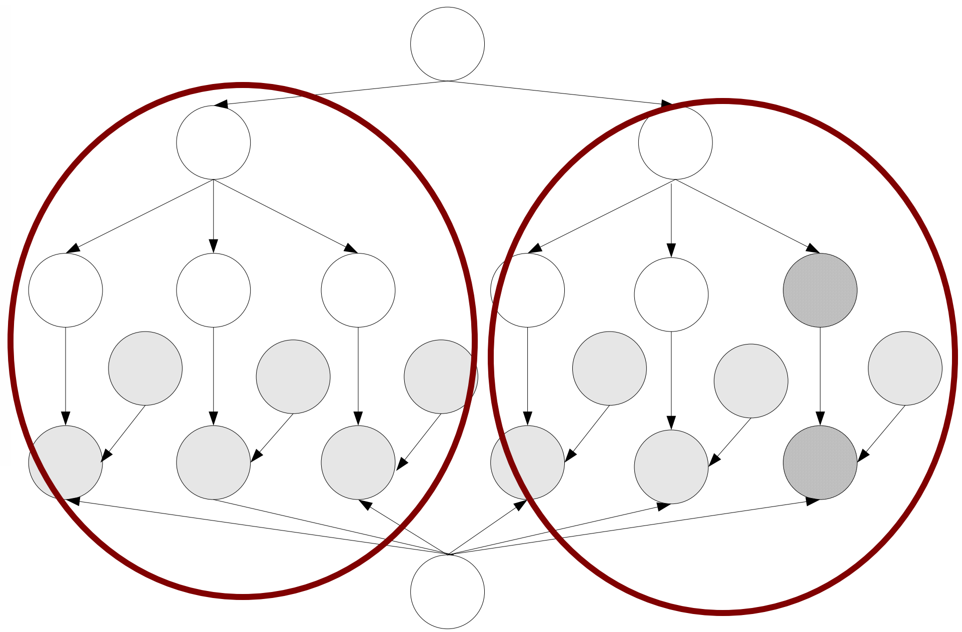




1. Choose $\pi \sim \mathrm{Dir}(\alpha)$
2. For each of d tokens $(x_j, [j]_1^m)$ in $\mathbf{x}$:
   (a) Choose a component $z_j \sim \mathrm{Discrete}(\pi)$.
   (b) Choose x_j from $p(x_j|\beta_{z_j})$, a Discrete distribution conditioned on the topic z_j.

Bayesian Naïve Bayes (BNB) [BS'07]

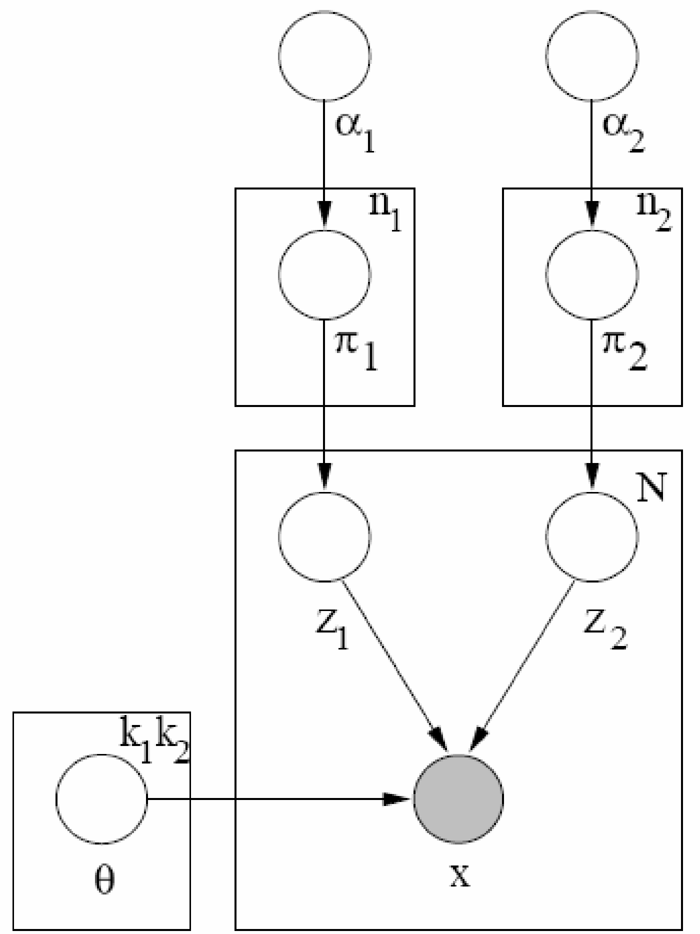




1. Choose $\pi \sim \text{Dir}(\alpha)$.

2. For each of the observed features f_j, $[j]_1^m$:
   (a) Choose a class $z_j \sim \text{Discrete}(\pi)$,
   (b) Choose a feature value $x_j \sim p_\psi(x_j | z_j, f_j, \Theta)$.

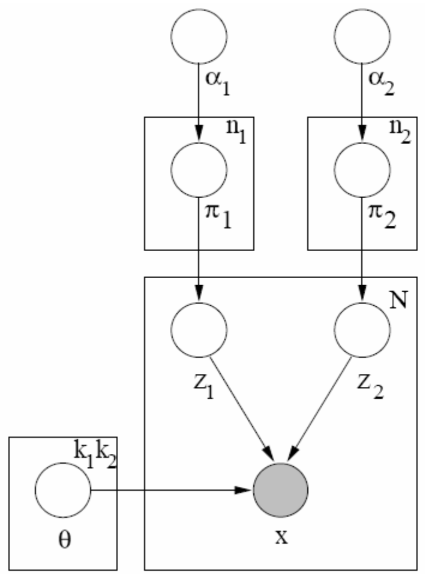
Bayesian Co-clustering (BCC)



1. For each row $u, [u]_1^{n_1}$, choose $\pi_{1u} \sim \mathrm{Dir}(\alpha_1)$.
2. For each column $v, [v]_1^{n2}$, choose $\pi_{2v} \sim \mathrm{Dir}(\alpha_2)$.
3. For each non-missing entry in row u and column v:
   (a) Choose $z_1 \sim \mathrm{Discrete}(\pi_{1u})$.
   (b) Choose $z_2 \sim \mathrm{Discrete}(\pi_{2v})$.
   (c) Choose $x_{uv} \sim p_(x|\theta_{z_1 z_2})$.

$$\log p(X|\alpha_1, \alpha_2, \Theta) \neq \sum_{n=1}^{N} \log p(x_n|\alpha_1, \alpha_2, \Theta)$$

Bayesian Co-clustering (BCC)



$$p(X, \pi_{1u}, \pi_{2v}, z_{1uv}, z_{2uv}, [u]_1^{n_1}, [v]_1^{n_2} | \alpha_1, \alpha_2, \Theta)$$
$$= \left(\prod_{u=1}^{n_1} p(\pi_{1u}|\alpha_1)\right) \times \left(\prod_{v=1}^{n_2} p(\pi_{2v}|\alpha_2)\right)$$
$$\times \left(\prod_{u,v} p(z_{1uv}|\pi_{1u}) p(z_{2uv}|\pi_{2v}) p(x_{uv}|\theta_{z_{1uv}, z_{2uv}})^{\delta_{uv}}\right)$$

$$p(X|\alpha_1, \alpha_2, \Theta) = \int_{\substack{\pi_{1u} \\ u=1,\ldots,n_1}} \int_{\substack{\pi_{2v} \\ v=1,\ldots,n_2}} \left(\prod_{u=1}^{n_1} p(\pi_{1u}|\alpha_1)\right) \left(\prod_{v=1}^{n_2} p(\pi_{2v}|\alpha_2)\right)$$
$$\left(\prod_{u,v} \sum_{z_{1uv}=1}^{k_1} \sum_{z_{2uv}=1}^{k_2} p(z_{1uv}|\pi_{1u}) p(z_{2uv}|\pi_{2v}) p(x_{uv}|\theta_{z_{1uv}, z_{2uv}})^{\delta_{uv}}\right) d\pi_{1u[u=1\cdots n_1]} d\pi_{2v[v=1\cdots n_2]}.$$

Variational Inference

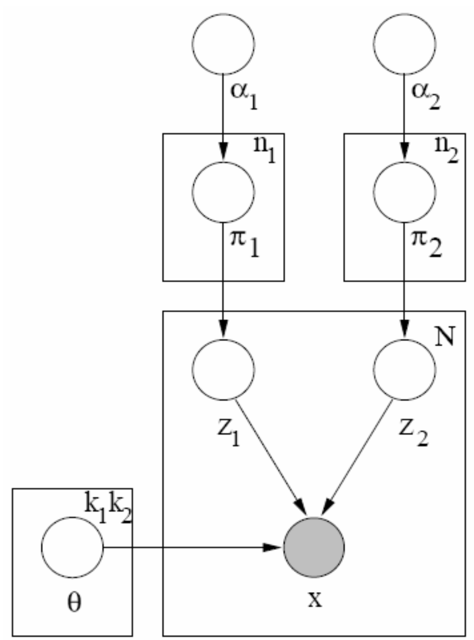
- Expectation Maximization
  - E-step: Calculate posterior probability $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$ to obtain log-likelihood $L(\alpha, \Theta)$.
  - M-step: Maximize $L(\alpha, \Theta)$ w.r.t α, Θ.
- Variational EM
  – Introduce a variational distribution $q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ to approximate $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$.
  – Use Jensen's inequality to get a tractable lower bound for log-likelihood

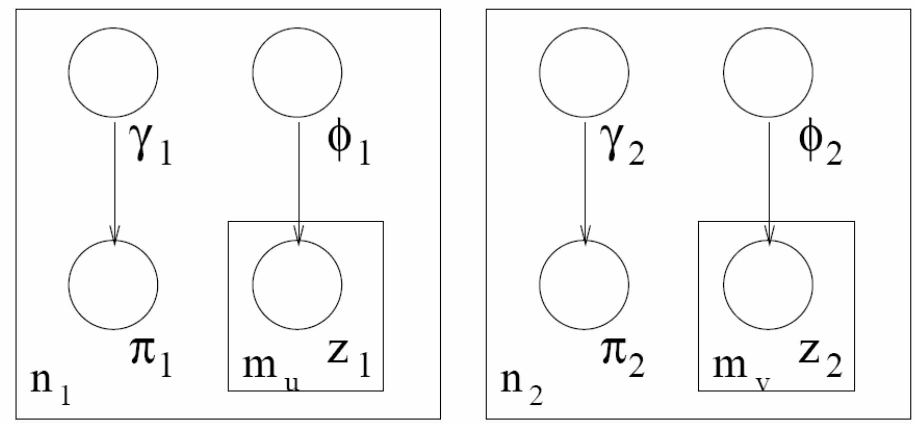
$$\log p(X|\alpha_1, \alpha_2, \Theta) \geq E_q[\log p(X, \mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2 | \alpha_1, \alpha_2, \Theta)] + H(q(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2))$$

  – Maximize the lower bound w.r.t $(\phi_1, \gamma_1, \phi_2, \gamma_2)$ for the best lower bound, i.e., minimize the KL divergence between $q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ and $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$
  – Maximize the lower bound w.r.t $(\alpha_1, \alpha_2, \Theta)$

Variational Distribution

- $\mathrm{Dir}(\gamma_1), \mathrm{Disc}(\phi_1)$ for each row, $\mathrm{Dir}(\gamma_2), \mathrm{Disc}(\phi_2)$ for each column





(a) row (b) column

$$q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2) = \left(\prod_{u=1}^{n_1} q(\pi_{1u} | \gamma_{1u}) \right) \times \left(\prod_{v=1}^{n_2} q(\pi_{2v} | \gamma_{2v}) \right)$$
$$\times \left(\prod_{u=1}^{n_1} \prod_{v=1}^{n_2} q(z_{1uv} | \phi_{1u}) q(z_{2uv} | \phi_{2v}) \right)$$

Variational EM for Bayesian Co-clustering

$L(\gamma_1, \gamma_2, \phi_1, \phi_2; \alpha_1, \alpha_2, \Theta)$ = lower bound of log -likelihood

1. **E-step**: Given the model parameters $(\alpha_1^{(t)}, \alpha_2^{(t)}, \Theta^{(t)})$, find the variational parameters

$$(\gamma_1^{(t+1)}, \gamma_2^{(t+1)}, \phi_1^{(t+1)}, \phi_2^{(t+1)}) = \underset{(\gamma_1,\gamma_2,\phi_1,\phi_2)}{\arg\max} \; L(\gamma_1, \gamma_2, \phi_1, \phi_2; \alpha_1^{(t)}, \alpha_2^{(t)}, \Theta^{(t)}) \; .$$

   Then, $L(\gamma_1^{(t+1)}, \gamma_2^{(t+1)}, \phi_1^{(t+1)}, \phi_2^{(t+1)}; \alpha_1, \alpha_2, \Theta)$ serves as the lower bound function for $\log p(X|\alpha_1, \alpha_2, \Theta)$.

2. **M-step**: Obtain an improved estimate of the model parameters:

$$(\alpha_1^{(t+1)}, \alpha_2^{(t+1)}, \Theta^{(t+1)}) = \underset{(\alpha_1,\alpha_2,\Theta)}{\arg\max} \; L(\gamma_1^{(t+1)}, \gamma_2^{(t+1)}, \phi_1^{(t+1)}, \phi_2^{(t+1)}; \alpha_1, \alpha_2, \Theta) \; .$$

EM for Bayesian Co-clustering

- Inference (E-step)

$$\phi_{1ui} \propto \exp\left(\Psi(\gamma_{1ui}) + \frac{\sum_{v=1}^{n_2}\sum_{j=1}^{k_2} \delta_{uv}\phi_{2vj}\log p(x_{uv}|\theta_{ij})}{m_u}\right)$$

$$\phi_{2vj} \propto \exp\left(\Psi(\gamma_{2vj}) + \frac{\sum_{u=1}^{n_1}\sum_{i=1}^{k_1} \delta_{uv}\phi_{1ui}\log p(x_{uv}|\theta_{ij})}{m_v}\right)$$

$$\gamma_{1ui} = \alpha_{1i} + m_u\phi_{1ui}$$

$$\gamma_{2vj} = \alpha_{2j} + m_v\phi_{2vj}$$

- Parameter Estimation (M-step) (Gaussians)

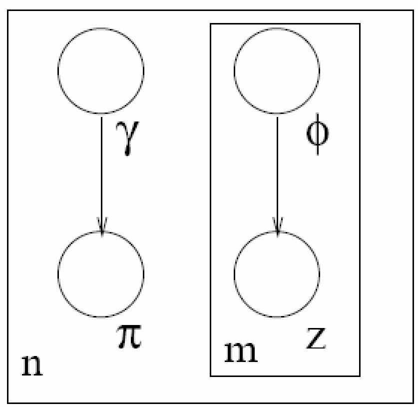
$$\mu_{ij} = \frac{\sum_{u=1}^{n_1}\sum_{v=1}^{n_2} \delta_{uv}x_{uv}\phi_{1ui}\phi_{2vj}}{\sum_{u=1}^{n_1}\sum_{v=1}^{n_2} \delta_{uv}\phi_{1ui}\phi_{2vj}}$$

$$\sigma_{ij}^2 = \frac{\sum_{u=1}^{n_1}\sum_{v=1}^{n_2} \delta_{uv}(x_{uv}-\mu_{ij})^2\phi_{1ui}\phi_{2vj}}{\sum_{u=1}^{n_1}\sum_{v=1}^{n_2} \delta_{uv}\phi_{1ui}\phi_{2vj}}$$

Fast Latent Dirichlet Allocation (FastLDA)

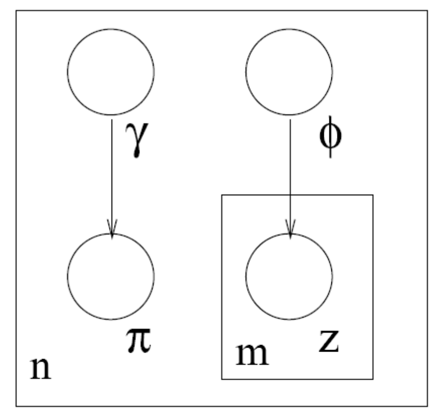
- Introduce a different variational distribution $q(\pi, \mathbf{z}|\gamma, \phi)$ as an approximation of $p(\pi, \mathbf{z}|\alpha, \Theta, \mathbf{x})$.

Original



$$q(\pi, \mathbf{z}|\gamma, \phi) = q(\pi|\gamma) \prod_{j=1}^{m} q(z_j|\phi_j)$$

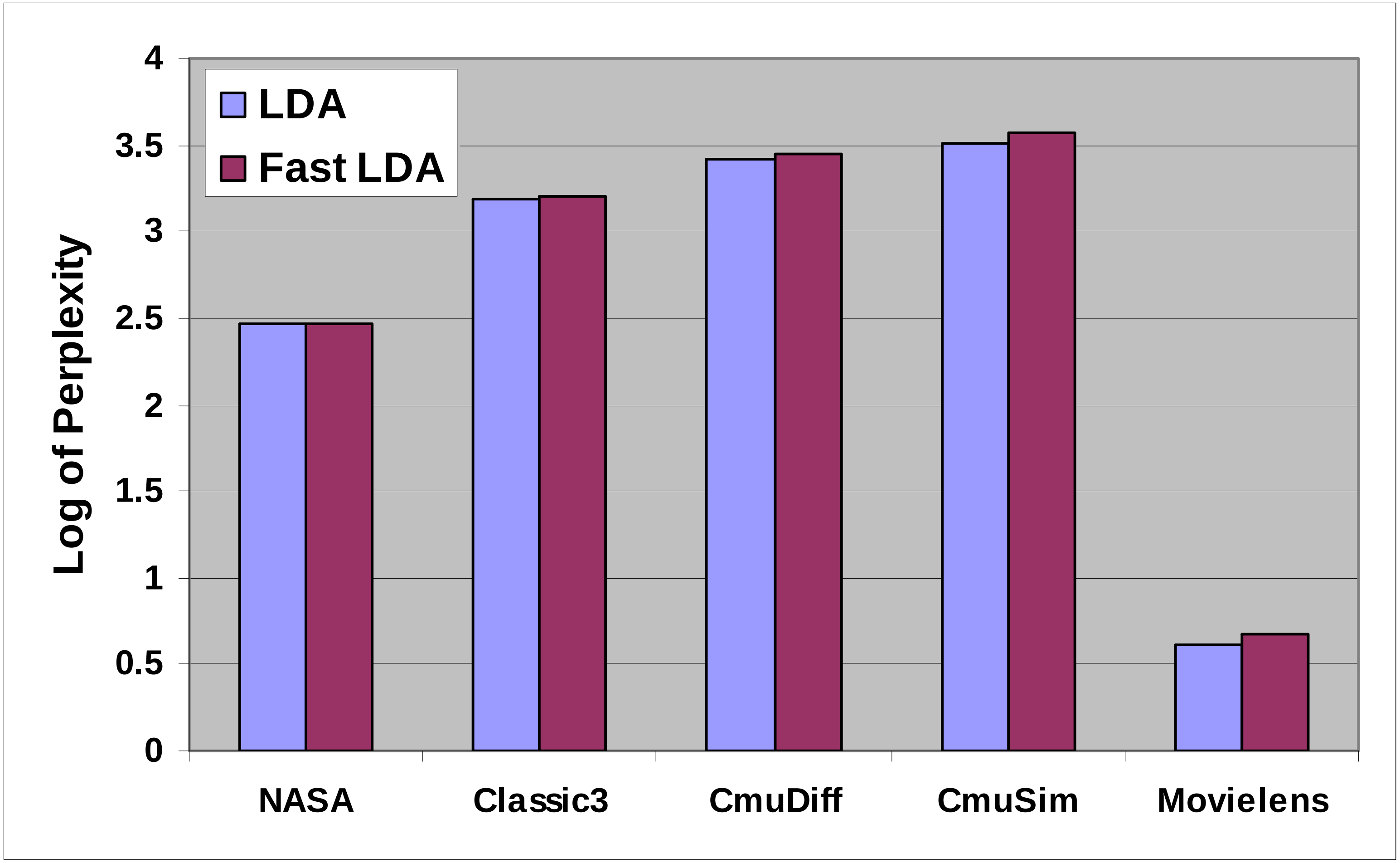
FastLDA



$$q(\pi, \mathbf{z}|\phi, \gamma) = q(\pi|\gamma) \prod_{j=1}^{m} q(z_j|\phi)$$

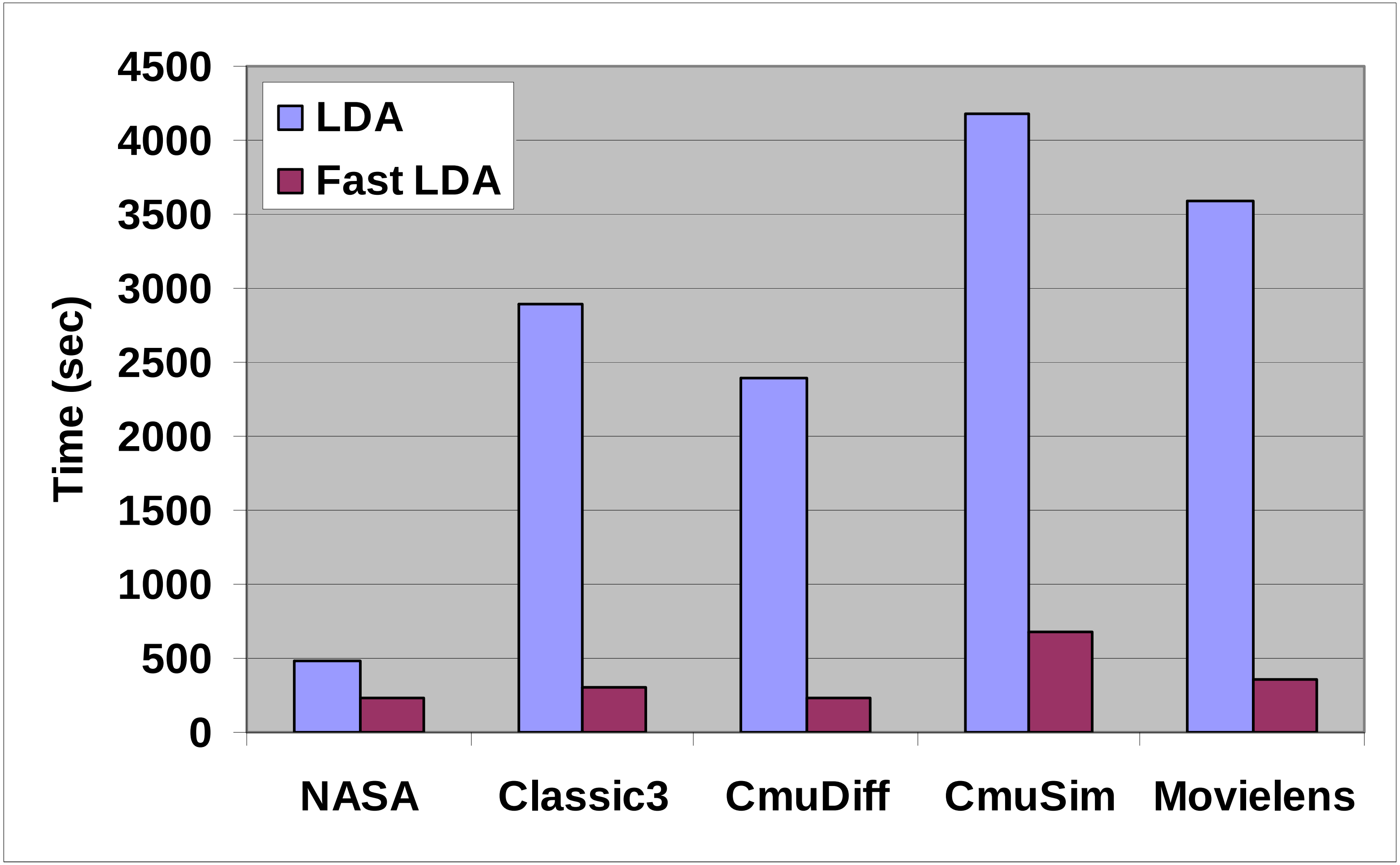
- Number of variational parameters ϕ: $m*n \rightarrow n$.
- Number of optimizations over ϕ: $m*n \rightarrow n$.

FastLDA vs LDA: Perplexity



$$Perplexity(X) = \exp\left\{ - \frac{\sum_{i=1}^{n} \log p(\mathbf{x}_i)}{\sum_{i=1}^{n} m_i} \right\}$$

FastLDA vs LDA: Time



Word List for Topics (Classic3)

LDA

Topic 1	Topic 2	Topic 3
information	patients	flow
library	cells	boundary
system	cases	pressure
data	normal	layer
libraries	growth	number
research	blood	mach
systems	found	results
retrieval	treatment	theory
science	children	heat
scientific	cell	method

Fast LDA

Topic 1	Topic 2	Topic 3
information	patients	flow
library	cells	boundary
system	cases	pressure
libraries	normal	layer
data	growth	number
research	blood	mach
retrieval	treatment	results
systems	found	theory
science	children	shock
scientific	cell	heat

Word List for Three Topics on Classic3 Dataset

Word List for Topics (Newsgroups)

LDA

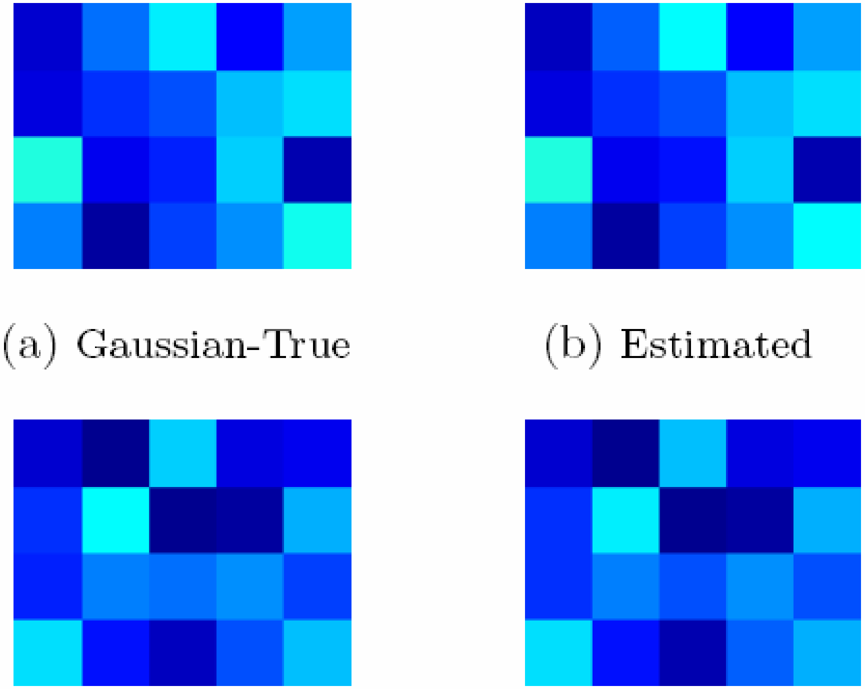
Topic 1	Topic 2	Topic 3
god	space	year
people	earth	game
don	nasa	don
time	launch	team
good	orbit	baseball
religion	system	good
make	shuttle	time
objective	moon	games
point	time	hit
evidence	mission	players

Fast LDA

Topic 1	Topic 2	Topic 3
god	space	year
people	earth	game
don	nasa	don
religion	launch	team
time	time	baseball
objective	orbit	good
good	system	games
moral	don	time
make	shuttle	hit
point	moon	players

Word List for Three Topics on CmuDiff Dataset

BCC Results: Simulated Data



(a) Gaussian-True (b) Estimated

(c) Bernoulli-True (d) Estimated

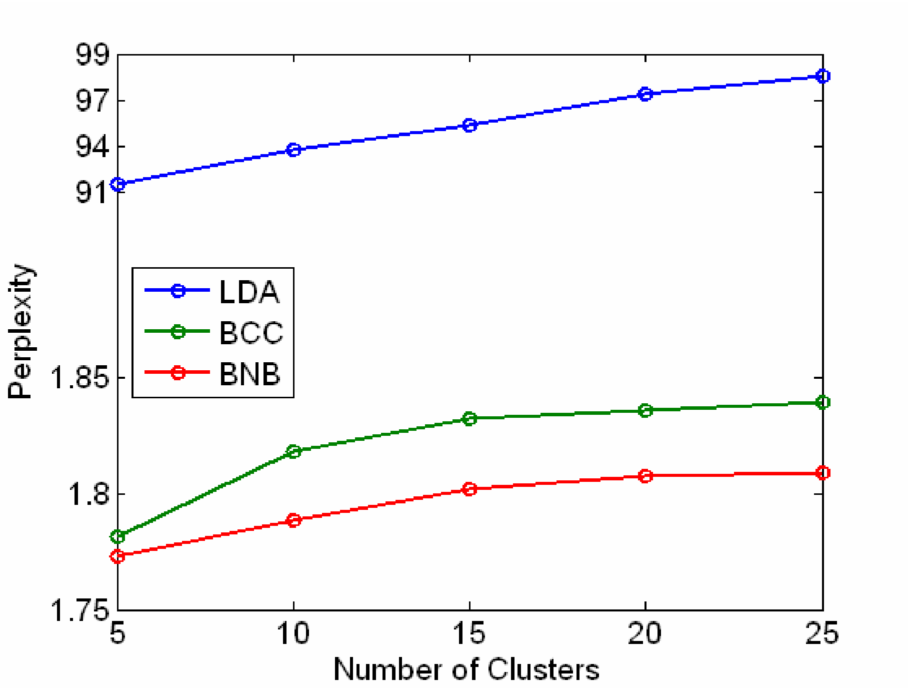
	Gaussian	Bernoulli	Poisson
Row	100%	99.5833%	100%
Column	100%	98.5833%	100%

Table 1: Cluster accuracy on simulated data.

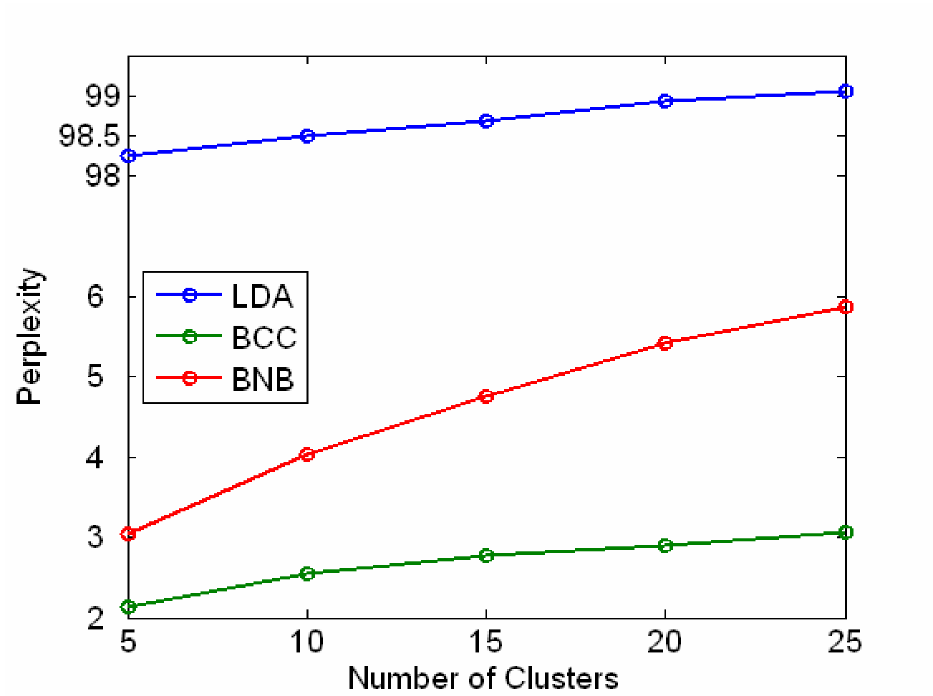
BCC Results: Real Data

- Movielens: Movie recommendation data
  - 100,000 ratings (1-5) for 1682 movies from 943 users (6.3%)
  - Binarize: 0 (1-3), 1(4-5).
  - Discrete (original), Bernoulli (binary)
- Foodmart: Transaction data
  - 164,558 sales records for 7803 customers and 1559 products (1.35%)
  - Binarize: 0 (less than median), 1(higher than median)
  - Poisson (original), Bernoulli (binary)
- Jester: Joke rating data
  - 100,000 ratings (-10.00 - +10.00) for 100 jokes from 1000 users (100%)
  - Binarize: 0 (lower than 0), 1 (higher than 0)
  - Gaussian (original), Bernoulli (binary)

BCC vs BNB vs LDA (Binary data)



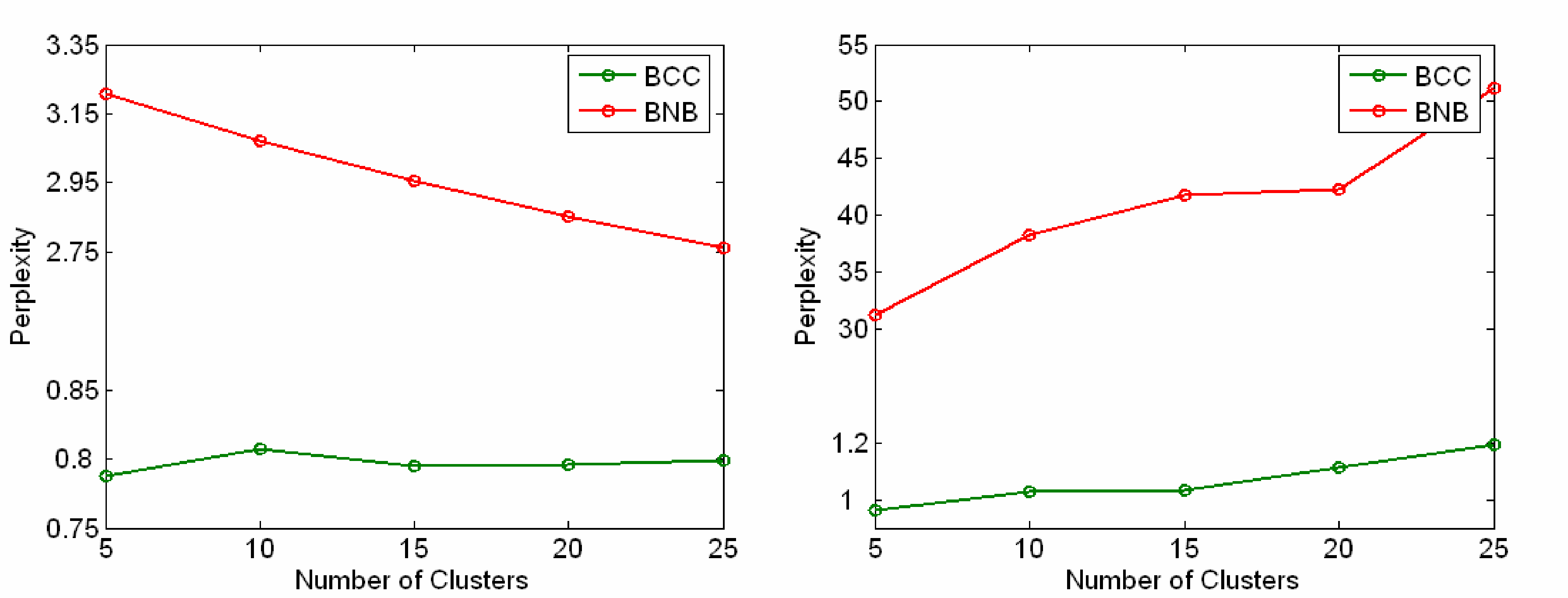
Training Set



Test Set

Perplexity on Binary Jester Dataset with Different Number of User Clusters

BCC vs BNB (Original data)



Training Set | Test Set

Perplexity on Movielens Dataset with Different Number of User Clusters

Perplexity Comparison with 10 User Clusters

Training Set

	BNB	BCC	LDA
Jester	**1.7883**	1.8186	98.3742
Movielens	**1.6994**	1.9831	439.6361
Foodmart	**1.8691**	1.9545	1461.7463

Test Set

	BNB	BCC	LDA
Jester	4.0237	**2.5498**	98.9964
Movielens	3.9320	**2.8620**	1557.0032
Foodmart	6.4751	**2.1143**	6542.9920

On Binary Data

Training Set

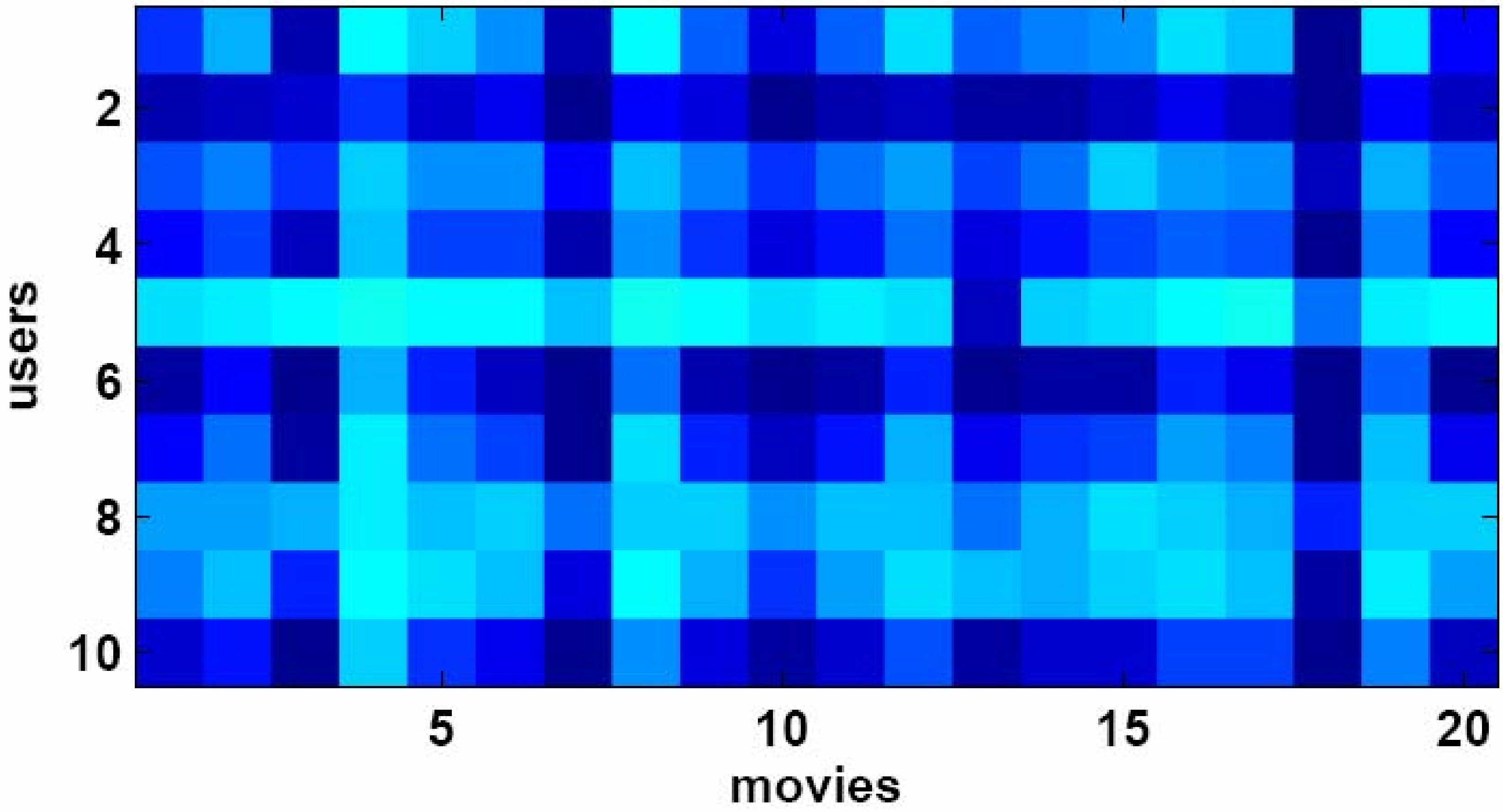
	BNB	BCC
Jester	**15.4620**	18.2495
Movielens	3.1495	**0.8068**
Foodmart	**4.5901**	4.5938

Test Set

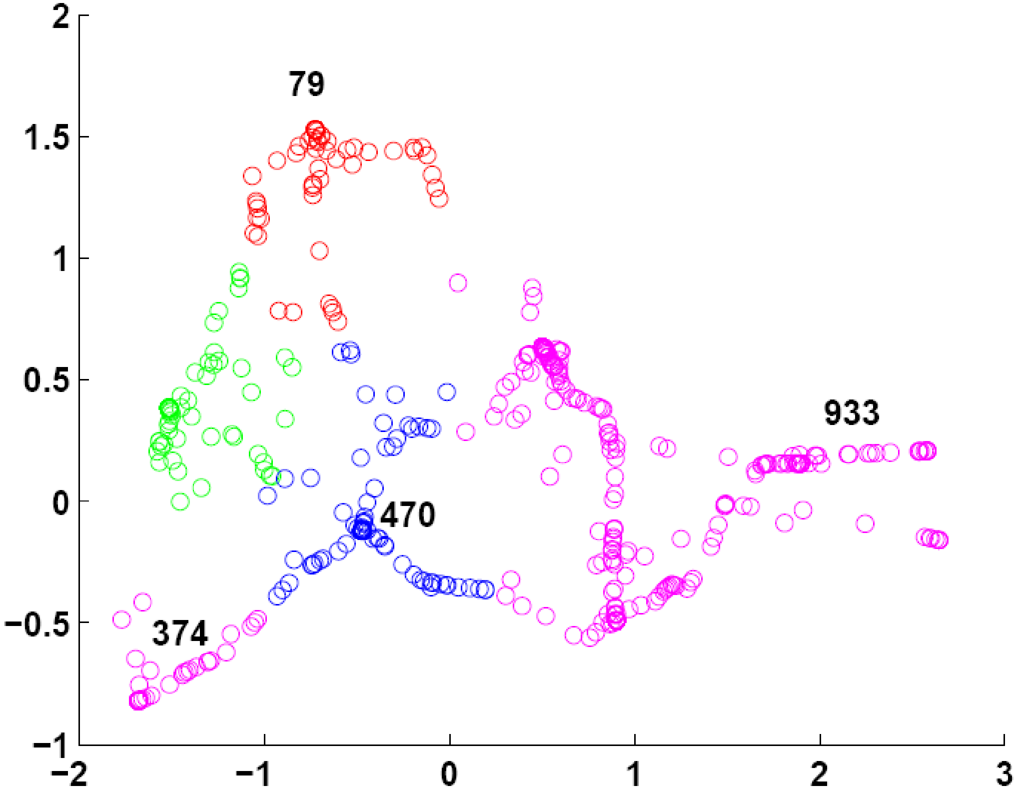
	BNB	BCC
Jester	39.9395	**24.8239**
Movielens	38.2377	**1.0265**
Foodmart	4.6681	**4.5964**

On Original Data

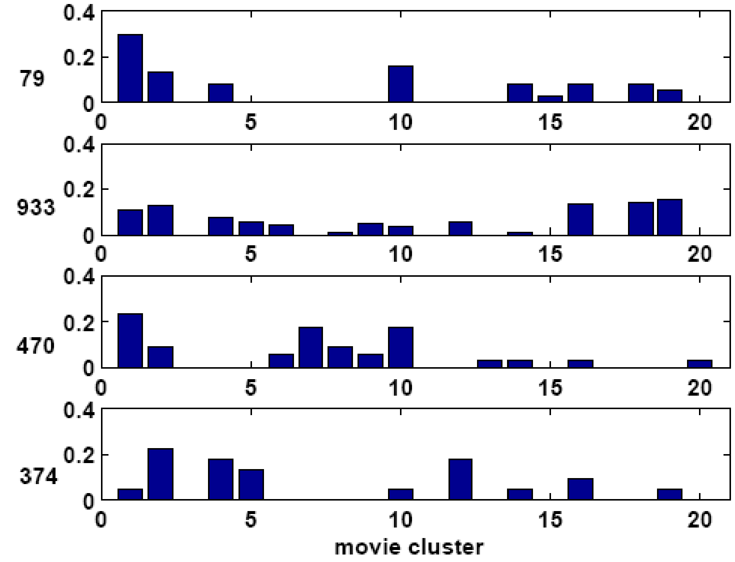
Co-cluster Parameters (Movielens)



Co-embedding: Users



Users

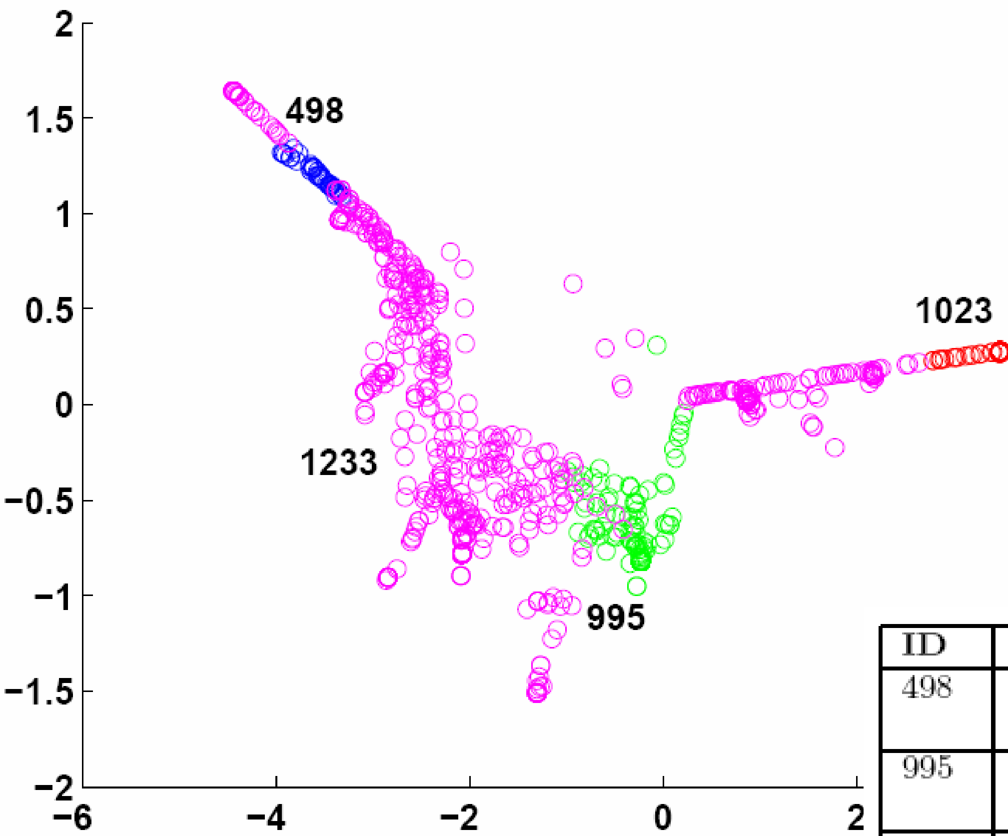


User signatures

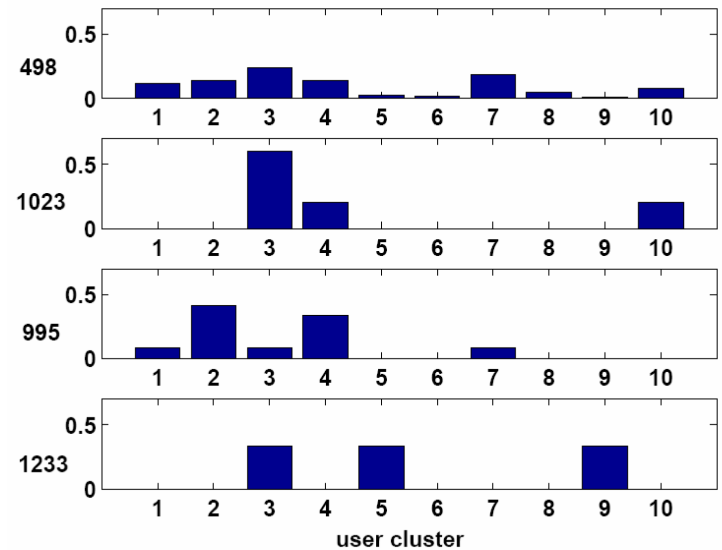
ID	Age	Sex	Occupation
79	39	F	administrator
374	36	M	executive
470	24	M	programmer
933	28	M	student

User profiles.

Co-embedding: Movies



Movies



Movie signatures

ID	Movie	Keywords
498	The African Queen	American Expatriate, Boat, Mission, African Tribe
995	Kiss Me, Guido	Italian Food, Homosexual, Pizza, Gay Interest
1023	Fathers' Day	Seduction, Con, Box Office Flop, Friendship
1233	Nënette et Boni	Brother Sister Relationship, Teen, Pregnancy, Teenage Pregnancy

Movie names and keywords.

Summary

- Bayesian co-clustering
  - Mixed membership co-clustering for dyadic data
  - Flexible Bayesian priors over memberships
  - Applicable to variety of data types
  - Stable performance, consistently better in test set
- Fast variational inference algorithm
  - One variational parameter for each row/column
  - Maintains coupling between row/column cluster memberships
  - Same idea leads to FastLDA (try it at home)
- Future work
  - Open problem: Joint decoding of missing entries
  - Predictive models based on mixed membership co-clusters
  - Multi-relational clustering

References

- **A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation**
  A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D. Modha.
  Journal of Machine Learning Research (JMLR), (2007) .
- **Latent Dirichlet Conditional Naive Bayes Models**
  A. Banerjee and H. Shan.
  IEEE International Conference on Data Mining (ICDM), (2007).
- **Latent Dirichlet Allocation**
  D. Blei, A. Ng, M. Jordan.
  Journal of Machine Learning Research (JMLR), (2003).
- **Bayesian Co-clustering**
  H. Shan, A. Banerjee.
  Tech Report, University of Minnesota, *Twin Cities,* (2008).