Introduction to Graphical Models for Data Mining

Arindam Banerjee
banerjee@cs.umn.edu

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

16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
July 25, 2010
Introduction

• Graphical Models
  – Brief Overview

• Part I: Tree Structured Graphical Models
  – Exact Inference

• Part II: Mixed Membership Models
  – Latent Dirichlet Allocation
  – Generalizations, Applications

• Part III: Graphical Models for Matrix Analysis
  – Probabilistic Matrix Factorization
  – Probabilistic Co-clustering
  – Stochastic Block Models
Graphical Models: What and Why

• Statistical Data Analysis
  – Build diagnostic/predictive models from data
  – Uncertainty quantification based on (minimal) assumptions

• The I.I.D. assumption
  – Data is independently and identically distributed
  – Example: Words in a doc drawn i.i.d. from the dictionary

• Graphical models
  – Assume (graphical) dependencies between (random) variables
  – Closer to reality, domain knowledge can be captured
  – Learning/inference is much more difficult
Flavors of Graphical Models

• Basic nomenclature
  – Node = random variable, maybe observed/hidden
  – Edge = statistical dependency

• Two popular flavors: ‘Directed’ and ‘Undirected’

• Directed Graphs
  – A directed graph between random variables, causal dependencies
  – Example: Bayesian networks, Hidden Markov Models
  – Joint distribution is a product of $P(\text{child}|\text{parents})$

• Undirected Graphs
  – An undirected graph between random variables
  – Example: Markov/Conditional random fields
  – Joint distribution in terms of potential functions
Bayesian Networks

- Joint distribution in terms of $P(X|\text{Parents}(X))$

$$
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
$$

$$
= \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))
$$
Example 1: Burglary Network

This and several other examples are from the Russell-Norvig AI book.
Computing Probabilities of Events

- Probability of any event can be computed:
  \[ P(B, E, A, J, M) = P(B) \cdot P(E|B) \cdot P(A|B, E) \cdot P(J|B, E, A) \cdot P(M|B, E, A, J) \]
  \[ = P(B) \cdot P(E) \cdot P(A|B, E) \cdot P(J|A) \cdot P(M|A) \]

- Example:
  \[ P(b, \neg e, a, \neg j, m) = P(b) \cdot P(\neg e) \cdot P(a|b, \neg e) \cdot P(\neg j|a) \cdot P(m|a) \]
Example II: Rain Network
Example III: “Car Won’t Start” Diagnosis

[Graphical diagram showing the relationship between various components and conditions related to a car not starting, including battery age, alternator, fanbelt, battery dead, no charging, battery flat, no oil, no gas, fuel line blocked, starter broken, lights, oil light, gas gauge, car won’t start, and dipstick.]
Inference

• Some variables in the Bayes net are observed
  – the evidence/data, e.g., John has not called, Mary has called

• Inference
  – How to compute value/probability of other variables
  – Example: What is the probability of Burglary, i.e., \( P(b|\neg j,m) \)
Inference Algorithms

- **Graphs without loops: Tree-structured Graphs**
  - Efficient exact inference algorithms are possible
  - Sum-product algorithm, and its special cases
    - Belief propagation in Bayes nets
    - Forward-Backward algorithm in Hidden Markov Models (HMMs)

- **Graphs with loops**
  - Junction tree algorithms
    - Convert into a graph without loops
    - May lead to exponentially large graph
  - Sum-product/message passing algorithm, ‘disregarding loops’
    - Active research topic, correct convergence ‘not guaranteed’
    - Works well in practice
  - Approximate inference
Approximate Inference

- **Variational Inference**
  - Deterministic approximation
  - Approximate complex true distribution over latent variables
  - Replace with family of simple/tractable distributions
    - Use the best approximation in the family
  - Examples: Mean-field, Bethe, Kikuchi, Expectation Propagation

- **Stochastic Inference**
  - Simple sampling approaches
  - Markov Chain Monte Carlo methods (MCMC)
    - Powerful family of methods
  - Gibbs sampling
    - Useful special case of MCMC methods
Part I: Tree Structured Graphical Models

• The Inference Problem

• Factor Graphs and the Sum-Product Algorithm

• Example: Hidden Markov Models

• Generalizations
The Inference Problem

How can we compute $P(b|j, m)$?
Complexity of Naïve Inference

- Simple query can be answered using Bayes rule
  - From Bayes Rule

\[
P(b|j, m) = \frac{P(b, j, m)}{P(j, m)}
\]
**Bayes Nets to Factor Graphs**

\[
f_A(x_1) = p(x_1) \quad f_B(x_2) = p(x_2) \quad f_C(x_1, x_2, x_3) = p(x_3|x_1, x_2)
\]

\[
f_D(x_3, x_4) = p(x_4|x_3) \quad f_E(x_3, x_5) = p(x_5|x_3)
\]
Factor Graphs: Product of Local Functions

\[ g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5) \]
Marginalize Product of Functions (MPF)

- Marginalize product of functions

\[ g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1)f_B(x_2)f_C(x_1, x_2, x_3)f_D(x_3, x_4)f_E(x_3, x_5) \]

- Computing marginal functions

\[ g_i(x_i) = \sum_{x \sim x_i} g(x_1, x_2, x_3, x_4, x_5) \]

- The “not-sum” notation

\[ \sum_{x \sim x_2} f(x_1, x_2, x_3) = \sum_{x_1, x_3} f(x_1, x_2, x_3) \]
MPF using Distributive Law

- We focus on two examples: \( g_1(x_1) \) and \( g_3(x_3) \)
- Main Idea: Distributive law
  \[ ab + ac = a(b+c) \]
- For \( g_1(x_1) \), we have
  \[
g_1(x_1) = f_A(x_1) \sum_{x_1} \left( f_B(x_2) f_C(x_1, x_2, x_3) \left( \sum_{x_3} f_D(x_3, x_4) \right) \left( \sum_{x_3} f_E(x_3, x_5) \right) \right)
  \]
- For \( g_3(x_3) \), we have
  \[
g_3(x_3) = \left( \sum_{x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) \right) \left( \sum_{x_3} f_D(x_3, x_4) \right) \left( \sum_{x_3} f_E(x_3, x_5) \right)
  \]
Computing Single Marginals

- **Main Idea:**
  - Target node becomes the root
  - Pass messages from leaves up to the root
Message Passing

Compute product of descendants
Example: Computing \( g_1(x_1) \)

\[
g_1(x_1) = f_A(x_1) \sum_{x_1} \left( f_B(x_2) f_C(x_1, x_2, x_3) \left( \sum_{x_3} f_D(x_3, x_4) \right) \left( \sum_{x_3} f_E(x_3, x_5) \right) \right)
\]
Example: Computing $g_3(x_3)$

$$g_3(x_3) = \left( \sum_{x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) \right) \left( \sum_{x_3} f_D(x_3, x_4) \right) \left( \sum_{x_3} f_E(x_3, x_5) \right)$$

Efficient Algorithm is encoded in the structure of the factor graph.
Hidden Markov Models (HMMs)

Latent variables:
\[ z_0, z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_T \]

Observed variables:
\[ x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T \]

Inference Problems:
1. Compute \( p(x_{1:T}) \)
2. Compute \( p(z_t|x_{1:T}) \)
3. Find \( \max_{z_{1:T}} p(z_{1:T}|x_{1:T}) \)

Similar problem for chain-structured Conditional Random Fields (CRFs)
The Sum-Product Algorithm

• To compute $g_i(x_i)$, form a tree rooted at $x_i$
• Starting from the leaves, apply the following two rules
  – Product Rule:
    At a variable node, take the product of descendants
  – Sum-product Rule:
    At a factor node, take the product of $f$ with descendants;
    then perform not-sum over the parent node
• To compute all marginals
  – Can be done one at a time; repeated computations, not efficient
  – Simultaneous message passing following the sum-product algorithm
  – Examples: Belief Propagation, Forward-Backward algorithm, etc.
Sum-Product Updates

- Variable to local function:
  \[ \mu_{x\rightarrow f}(x) = \prod_{h \in n(x) \setminus f} \mu_{h\rightarrow x} \]

- Local function to variable:
  \[ \mu_{f\rightarrow x}(x) = \sum_{\sim x} \left( f(x) \prod_{y \in n(f) \setminus \{x\}} \mu_{y\rightarrow f}(y) \right) \]
Sum-Product Updates
Example: Step 1

\[ \mu_{f_A \rightarrow x_1}(x_1) = f_A(x_1) \]
\[ \mu_{f_B \rightarrow x_2}(x_2) = f_B(x_2) \]
\[ \mu_{x_4 \rightarrow f_D}(x_4) = 1 \]
\[ \mu_{x_5 \rightarrow f_E}(x_5) = 1 \]
Example: Step 2

\[
\begin{align*}
\mu_{x_1 \rightarrow f_C}(x_1) &= \mu_{f_A \rightarrow x_1}(x_1) \\
\mu_{x_2 \rightarrow f_C}(x_2) &= \mu_{f_B \rightarrow x_2}(x_2) \\
\mu_{f_D \rightarrow x_3}(x_3) &= \sum_{x_3} f_D(x_3, x_4) \mu_{x_4 \rightarrow f_D}(x_4) \\
\mu_{f_E \rightarrow x_3}(x_3) &= \sum_{x_3} f_D(x_3, x_5) \mu_{x_5 \rightarrow f_E}(x_5)
\end{align*}
\]
Example: Step 3

\[
\mu_{f_C \rightarrow x_3}(x_3) = \sum_{x_1, x_2, x_3} f_C(x_1, x_2, x_3) \mu_{x_1 \rightarrow f_C}(x_1) \mu_{x_2 \rightarrow f_C}(x_2)
\]

\[
\mu_{x_3 \rightarrow f_C}(x_3) = \mu_{f_D \rightarrow x_3}(x_3) \mu_{f_E \rightarrow x_3}(x_3)
\]
Example: Step 4

\[
\mu_{f_{C \rightarrow x_{1}}}(x_{1}) = \sum_{x_{1}} f_{C}(x_{1}, x_{2}, x_{3}) \mu_{x_{2} \rightarrow f_{C}}(x_{2}) \mu_{x_{3} \rightarrow f_{C}}(x_{3})
\]

\[
\mu_{f_{C \rightarrow x_{2}}}(x_{2}) = \sum_{x_{2}} f_{C}(x_{1}, x_{2}, x_{3}) \mu_{x_{1} \rightarrow f_{C}}(x_{1}) \mu_{x_{3} \rightarrow f_{C}}(x_{3})
\]

\[
\mu_{x_{3} \rightarrow f_{D}}(x_{3}) = \mu_{f_{C \rightarrow x_{3}}}(x_{3}) \mu_{f_{E \rightarrow x_{3}}}(x_{3})
\]

\[
\mu_{x_{3} \rightarrow f_{E}}(x_{3}) = \mu_{f_{C \rightarrow x_{3}}}(x_{3}) \mu_{f_{D \rightarrow x_{3}}}(x_{3})
\]
Example: Step 5

\[
\begin{align*}
\mu_{x_1 \to f_A}(x_1) &= \mu_{f_C \to x_1}(x_1) \\
\mu_{x_2 \to f_B}(x_2) &= \mu_{f_C \to x_2}(x_2) \\
\mu_{f_D \to x_4}(x_4) &= \sum_{x_4} f_D(x_3, x_4) \mu_{x_3 \to f_D}(x_4) \\
\mu_{f_E \to x_5}(x_5) &= \sum_{x_5} f_D(x_3, x_5) \mu_{x_3 \to f_E}(x_5)
\end{align*}
\]
Example: Termination

Marginal function is the product of all incoming messages

\[
\begin{align*}
g_1(x_1) &= \mu_{f_A \rightarrow x_1}(x_1) \mu_{f_C \rightarrow x_1}(x_1) \\
g_2(x_2) &= \mu_{f_B \rightarrow x_2}(x_2) \mu_{f_C \rightarrow x_2}(x_2) \\
g_3(x_3) &= \mu_{f_C \rightarrow x_3}(x_3) \mu_{f_D \rightarrow x_3}(x_3) \mu_{f_E \rightarrow x_3}(x_3) \\
g_2(x_2) &= \mu_{f_D \rightarrow x_4}(x_4) \\
g_5(x_5) &= \mu_{f_E \rightarrow x_5}(x_5)
\end{align*}
\]
HMMs Revisited

Latent variables: 
\( z_0, z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_T \)

Observed variables: 
\( x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T \)

Inference Problem:
1. Compute \( p(x_{1:T}) \)
2. Compute \( p(z_t | x_{1:T}) \)

Sum-product algorithm is known as the `forward-backward' algorithm

Smoothing in Kalman Filtering
Distributive Law on Semi-Rings

- Idea can be applied to any commutative semi-ring
- Semi-ring 101
  - Two operations (+, ×): Associative, Commutative, Identity
  - Distributive law: \( a \times b + a \times c = a \times (b+c) \)

<table>
<thead>
<tr>
<th>( K )</th>
<th>( (\cdot, 0) )</th>
<th>( (\cdot, 1) )</th>
<th>short name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( A )</td>
<td>(+, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>2. ( A[x] )</td>
<td>(+, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>3. ( A[x, y, \ldots] )</td>
<td>(+, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>4. ( [0, \infty) )</td>
<td>(min, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>5. ( (0, \infty) )</td>
<td>(max, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>6. ( [0, \infty) )</td>
<td>(min, 0)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>7. ( (-\infty, \infty] )</td>
<td>(max, -\infty)</td>
<td>(·, 1)</td>
<td></td>
</tr>
<tr>
<td>8. ( {0, 1} )</td>
<td>(OR, 0)</td>
<td>(AND, 1)</td>
<td></td>
</tr>
<tr>
<td>9. ( 2^S )</td>
<td>((\cup), 0)</td>
<td>((\cap), 1)</td>
<td></td>
</tr>
<tr>
<td>10. ( \Lambda )</td>
<td>((\lor), 0)</td>
<td>((\land), 1)</td>
<td></td>
</tr>
<tr>
<td>11. ( \Lambda )</td>
<td>((\land), 1)</td>
<td>((\lor), 0)</td>
<td></td>
</tr>
</tbody>
</table>

- Belief Propagation in Bayes nets
- MAP inference in HMMs
- Max-product algorithm
- Alternative to Viterbi Decoding
- Kalman Filtering
- Error Correcting Codes
- Turbo Codes
- …
Message Passing in General Graphs

- Tree structured graphs
  - Message passing is guaranteed to give correct solutions
  - Examples: HMMs, Kalman Filters

- General Graphs
  - Active research topic
    - Progress has been made in the past 10 years
  - Message passing
    - May not converge
    - May converge to a ‘local minima’ of ‘Bethe variational free energy’
  - New approaches to convergent and correct message passing

- Applications
  - True Skill: Ranking System for Xbox Live
  - Turbo Codes: 3G, 4G phones, satellite comm, Wimax, Mars orbiter
Part II: Mixed Membership Models

• Mixture Models vs Mixed Membership Models

• Latent Dirichlet Allocation

• Inference
  – Mean-Field and Collapsed Variational Inference
  – MCMC/Gibbs Sampling

• Applications

• Generalizations
Background: Plate Diagrams

Compact representation of large Bayesian networks
Model 1: Independent Features

\[ d=3, \ n=1 \]
Model 2: Naïve Bayes (Mixture Models)
Naïve Bayes Model
Naïve Bayes Model
Model 3: Mixed Membership Model
Mixed Membership Models
Mixed Membership Models

Graphical Models
Mixture Model vs Mixed Membership Model

Single component membership

Multi-component mixed membership
Latent Dirichlet Allocation (LDA)

- Dirichlet priors
- Distribution over topics for each document: $\rho^{(d)} \sim \text{Dirichlet}(\alpha)$
- Topic assignment for each word: $z_i \sim \text{Discrete}(\pi^{(d)})$
- Word generated from assigned topic: $x_i \sim \text{Discrete}(\beta^{(z_i)})$
$z \sim \text{Discrete}(\pi)$

$\pi \sim \text{Dirichlet}(\alpha)$
LDA Generative Model
LDA Generative Model

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

2. For each of $d$ tokens $(x_j, [j]_m^m)$ in $\mathbf{x}$:
   
   (a) Choose a component $z_j \sim \text{Discrete}(\pi)$.

   (b) Choose $x_j$ from $p(x_j | \beta_{z_j})$, a Discrete distribution conditioned on the topic $z_j$. 
Learning: Inference and Estimation

- Learning
  - Estimate model parameters \((\alpha, \beta)\) to maximize log-likelihood
  - Infer ‘mixed-memberships’ of documents
Variational Inference

- Introduce a variational distribution $q(\pi, z | \gamma, \phi)$ to approximate $p(\pi, z | x, \alpha, \beta)$
Variational EM for LDA

\[ L(\gamma, \phi; \alpha, \beta) = \text{lower bound to log-likelihood } L(\alpha, \beta) \]

- E-step: Given model parameters \((\alpha^{(t)}, \beta^{(t)})\), find variational parameters:

\[
(\gamma^{(t+1)}, \phi^{(t+1)}) = \arg\max_{(\gamma,\phi)} L(\gamma, \phi; \alpha^{(t)}, \beta^{(t)})
\]
E-step: Variational Distribution and Updates

- Fully factorized distribution over the latent variables

\[ q(\pi, z | \gamma, \phi) = q_{\text{Dirichlet}}(\pi | \gamma) \prod_{j=1}^{m} q_{\text{discrete}}(z_j | \phi_j) \]
M-step: Parameter Estimation

- For fixed \((\gamma_d, \phi_d)\), the lower bound is optimized over \((\alpha, \beta)\)
- Updates for word distributions
  \[
  \beta_h(v) \propto \sum_{d=1}^{D} \sum_{j=1}^{m} \phi_{d,j}(h) \mathbb{1}_{w_{d,j}=v}
  \]
- \(\alpha\) can be estimated using an efficient Newton method
- Alternate E- and M-steps till convergence
Results: Topics Inferred

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Results: Perplexity Comparison

\[ Perplexity(X) = \exp \left\{ - \frac{\sum_{i=1}^{n} \log p(x_i)}{\sum_{i=1}^{n} m_i} \right\} \]

Graphical Models
### Results: NASA Reports I

<table>
<thead>
<tr>
<th>Arrival Departure</th>
<th>Passenger</th>
<th>Maintenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>runway</td>
<td>passenger</td>
<td>maintenance</td>
</tr>
<tr>
<td>approach</td>
<td>attendant</td>
<td>engine</td>
</tr>
<tr>
<td>departure</td>
<td>flight</td>
<td>mel</td>
</tr>
<tr>
<td>altitude</td>
<td>seat</td>
<td>zzz</td>
</tr>
<tr>
<td>turn</td>
<td>medical</td>
<td>air craft</td>
</tr>
<tr>
<td>tower</td>
<td>captain</td>
<td>installed</td>
</tr>
<tr>
<td>air traffic control</td>
<td>attendants</td>
<td>check</td>
</tr>
<tr>
<td>heading</td>
<td>lavatory</td>
<td>inspection</td>
</tr>
<tr>
<td>taxi way</td>
<td>told</td>
<td>fuel</td>
</tr>
<tr>
<td>flight</td>
<td>police</td>
<td>Work</td>
</tr>
</tbody>
</table>
## Results: NASA Reports II

<table>
<thead>
<tr>
<th>Medical Emergency</th>
<th>Wheel Maintenance</th>
<th>Weather Condition</th>
<th>Departure</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical</td>
<td>tire</td>
<td>knots</td>
<td>departure</td>
</tr>
<tr>
<td>passenger</td>
<td>wheel</td>
<td>turbulence</td>
<td>sid</td>
</tr>
<tr>
<td>doctor</td>
<td>assembly</td>
<td>aircraft</td>
<td>dme</td>
</tr>
<tr>
<td>attendant</td>
<td>nut</td>
<td>degrees</td>
<td>altitude</td>
</tr>
<tr>
<td>oxygen</td>
<td>spacer</td>
<td>ice</td>
<td>climbing</td>
</tr>
<tr>
<td>emergency</td>
<td>main</td>
<td>winds</td>
<td>mean sea level</td>
</tr>
<tr>
<td>paramedics</td>
<td>axle</td>
<td>wind</td>
<td>heading</td>
</tr>
<tr>
<td>flight</td>
<td>bolt</td>
<td>speed</td>
<td>procedure</td>
</tr>
<tr>
<td>nurse</td>
<td>missing</td>
<td>air speed</td>
<td>turn</td>
</tr>
<tr>
<td>aed</td>
<td>tires</td>
<td>conditions</td>
<td>degree</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The pilot flies an owner's airplane with the owner as a passenger. Loses contact with the center during the flight.

While performing a sky diving, a jet approaches at the same altitude, but an accident is avoided.
Altimeter has a problem, but the pilot overcomes the difficulty during the flight. During acceleration, a flap retraction issue happens. The pilot then returns to base and lands. The mechanic finds out the problem.
The pilot has a landing gear problem. Maintenance crew joins radio conversation to help.

The captain has a medical emergency.

Red: Flight crew  Blue: Passenger  Green: Maintenance
Mixed Membership of Reports

Red: Flight Crew
Blue: Passenger
Green: Maintenance

Flight Crew: 0.7039
Passenger: 0.0009
Maintenance: 0.2953

Flight Crew: 0.1405
Passenger: 0.0663
Maintenance: 0.7932

Flight Crew: 0.0013
Passenger: 0.0013
Maintenance: 0.9973

Graphical Models
Smoothed Latent Dirichlet Allocation

- Dirichlet priors
- distribution over words for each topic: $\phi^{(i)} \sim \text{Dirichlet}(\beta)$
- distribution over topics for each document: $\rho^{(d)} \sim \text{Dirichlet}(\alpha)$
- topic assignment for each word: $z_i \sim \text{Discrete}(\rho^{(d)})$
- word generated from assigned topic: $x_i \sim \text{Discrete}(\phi^{(z_i)})$
Stochastic Inference using Markov Chains

- Powerful family of approximate inference methods
  - Markov Chain Monte Carlo, Gibbs Sampling

- The basic idea
  - Need to marginalize over complex latent variable distribution
    \[ p(x|\theta) = \int_z p(x,z|\theta) = \int_z p(x|\theta) \, p(z|x,\theta) = \mathbb{E}_{z \sim p(z|x,\theta)}[p(x|\theta)] \]
    - Draw ‘independent’ samples from \( p(z|x,\theta) \)
    - Compute sample based average instead of the full integral

- Main Issue: How to draw samples?
  - Difficult to directly draw samples from \( p(z|x,\theta) \)
  - Construct a Markov chain whose stationary distribution is \( p(z|x,\theta) \)
  - Run chain till ‘convergence’
  - Obtain samples from \( p(z|x,\theta) \)
The Metropolis-Hastings Algorithm

- Most popular MCMC method
- Based on a proposal distribution $q(x^* | x)$
The Metropolis-Hastings Algorithm (Contd)
The Gibbs Sampler

- For a $d$-dimensional vector $x$, assume we know

$$p(x_j|x_{-j}) = p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$$
Collapsed Gibbs Sampling for LDA

- Naive MCMC would sample all latent variables: \((z, \phi, \theta)\)
- Observation: \((\phi, \theta)\) can be marginalized in closed form
Collapsed Variational Inference for LDA

- Recall that $p(x, z|\alpha, \beta)$ can be obtained in closed form.
- However, we cannot marginalize over $z$.
- We approximate $p(z|x, \alpha, \beta)$ with $q(z|x, \alpha, \beta)$.
Collapsed Variational Inference for LDA

- With these approximations, the variational update is

\[
\gamma_{d,j}(h) \propto \frac{n_{w,h}^{-ij} + \beta}{n_{.,h}^{-ij} + D \beta} (n_{h,j}^{-ij} + \alpha) \exp \left( -\frac{v_{h,j}^{-ij}}{2(n_{h,j}^{-ij} + \alpha)^2} \right)
\]

- \[
- \frac{v_{wh}^{-ij}}{2(n_{wh}^{-ij} + \beta)^2} + \frac{v_{.,h}^{-ij}}{2(n_{.,h}^{-ij} + D \beta)^2}
\]

where, not including the current token,

- \( n_{h,j}^{-ij} = \sum_{i' \neq i} \gamma_{i'j|h} \), the expected number of tokens in document \( j \) assigned to topic \( h \);

- \( v_{h,j}^{-ij} = \sum_{i' \neq i} \gamma_{i'j|h}(1 - \gamma_{i'j|h}) \), the variance associated with the expected count; and similarly for other terms
Results: Comparison of Inference Methods

![Bar chart showing perplexity comparison for different datasets and methods: CGS, VB, CVB, and CVB0. The x-axis represents different data sources (CRAN, KOS, MED, NIPS, NEWS, NYT, PAT), and the y-axis represents perplexity ranging from 0 to 3500. Each dataset has bars for different inference methods, with CGS having the highest perplexity for most datasets, followed by VB and CVB, and CVB0 having the lowest perplexity.]
Results: Comparison of Inference Methods

![Graph showing Perplexity vs Iteration for CGS, VB, CVB, and CVB0 methods.](image)

- CGS
- VB
- CVB
- CVB0

Graphical Models
Generalizations

• Generalized Topic Models
  – Correlated Topic Models
  – Dynamic Topic Models, Topics over Time
  – Dynamic Topics with birth/death

• Mixed membership models over non-text data, applications
  – Mixed membership naïve-Bayes
  – Discriminative models for classification
  – Cluster Ensembles

• Nonparametric Priors
  – Dirichlet Process priors: Infer number of topics
  – Hierarchical Dirichlet processes: Infer hierarchical structures
  – Several other priors: Pachinko allocation, Gaussian Processes, IBP, etc.
CTM Results
DTM Results

Graphical Models
DTM Results II

Graphical Models
Mixed Membership Naïve Bayes

- For each data point,
  - Choose $\pi \sim \text{Dirichlet}(\alpha)$

- For each of observed features $f_n$:
  - Choose a class $z_n \sim \text{Discrete}(\pi)$
  - Choose a feature value $x_n$ from $p(x_n|z_n,f_n,\Theta)$, which could be Gaussian, Poisson, Bernoulli…
MMNB vs NB: Perplexity Surfaces

- MMNB typically achieves a lower perplexity than NB

- On test set, NB shows overfitting, but MMNB is stable and robust.
Discriminative Mixed Membership Models

(a) DLDA

(b) DMNB
Results: DLDA for text classification

<table>
<thead>
<tr>
<th></th>
<th>Nasa</th>
<th>Classic3</th>
<th>Diff</th>
<th>Sim</th>
<th>Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast DLDA</td>
<td>0.9301±0.0128</td>
<td>0.6866±0.0245</td>
<td>0.9823±0.0083</td>
<td>0.8718±0.0182</td>
<td>0.8468±0.0190</td>
</tr>
<tr>
<td>vMF</td>
<td>0.9216±0.0113</td>
<td>0.6509±0.0246</td>
<td>0.9330±0.0071</td>
<td>0.7447±0.0214</td>
<td>0.7600±0.0347</td>
</tr>
<tr>
<td>NB</td>
<td><strong>0.9334±0.0094</strong></td>
<td>0.6766±0.0230</td>
<td>0.9813±0.0069</td>
<td>0.8613±0.0216</td>
<td>0.8410±0.0262</td>
</tr>
<tr>
<td>LR</td>
<td>0.9209±0.0157</td>
<td>0.6396±0.0252</td>
<td>0.9553±0.0157</td>
<td>0.6750±0.1330</td>
<td>0.4823±0.1283</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9192±0.0146</td>
<td>0.6854±0.0278</td>
<td>0.9563±0.0105</td>
<td>0.8357±0.0156</td>
<td>0.8120±0.2030</td>
</tr>
</tbody>
</table>

Generally, Fast DLDA has a higher accuracy on most of the datasets.
Topics from DLDA
Cluster Ensembles

• Combining multiple base clusterings of a dataset

• Robust and stable
• Distributed and scalable
• Knowledge reuse, privacy preserving
### Problem Formulation

- **Input & Output**

<table>
<thead>
<tr>
<th>Data points</th>
<th>Base clusterings</th>
<th>Consensus clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( \lambda_1 )</td>
<td>( \lambda^* )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( 2 \quad 1 \quad \ldots \quad 3 )</td>
<td>( 1 \quad 3 \quad \ldots )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots \quad \vdots \quad \vdots \quad \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( x_N )</td>
<td>( 3 \quad 2 \quad \ldots \quad 6 )</td>
<td>( 6 )</td>
</tr>
</tbody>
</table>
## Results: State-of-the-art vs Bayesian Ensembles

<table>
<thead>
<tr>
<th>Dataset</th>
<th>The results of base clusterings K-means</th>
<th>MCLA</th>
<th>CSPA</th>
<th>HGPA</th>
<th>MM</th>
<th>K-means cluster ensemble</th>
<th>G-BCE</th>
<th>V-BCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>average</td>
<td>Max</td>
<td>average</td>
<td>Max</td>
<td>average</td>
<td>Max</td>
<td>average</td>
</tr>
<tr>
<td>ins</td>
<td>0.9167</td>
<td>0.9167</td>
<td>0.9533</td>
<td>0.9533</td>
<td>0.9067</td>
<td>0.9067</td>
<td>0.867</td>
<td>0.867</td>
</tr>
<tr>
<td>wdbc</td>
<td>0.8840</td>
<td>0.8840</td>
<td>0.8840</td>
<td>0.8840</td>
<td>0.5818</td>
<td>0.5818</td>
<td>0.940</td>
<td>0.940</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.7123</td>
<td>0.7123</td>
<td>0.7046</td>
<td>0.6952</td>
<td>0.6353</td>
<td>0.6353</td>
<td>0.7094</td>
<td>0.7094</td>
</tr>
<tr>
<td>glass</td>
<td>0.4393</td>
<td>0.4393</td>
<td>0.4393</td>
<td>0.4393</td>
<td>0.5748</td>
<td>0.5748</td>
<td>0.5519</td>
<td>0.5519</td>
</tr>
<tr>
<td>bupa</td>
<td>0.5710</td>
<td>0.5710</td>
<td>0.5710</td>
<td>0.5710</td>
<td>0.5188</td>
<td>0.5188</td>
<td>0.5075</td>
<td>0.5075</td>
</tr>
<tr>
<td>pima</td>
<td>0.5653</td>
<td>0.5653</td>
<td>0.5653</td>
<td>0.5653</td>
<td>0.6654</td>
<td>0.6654</td>
<td>0.5029</td>
<td>0.5029</td>
</tr>
<tr>
<td>wine</td>
<td>0.7416</td>
<td>0.7416</td>
<td>0.7416</td>
<td>0.7416</td>
<td>0.5562</td>
<td>0.5562</td>
<td>0.7129</td>
<td>0.7129</td>
</tr>
<tr>
<td>magic04</td>
<td>0.6491</td>
<td>0.6252</td>
<td>0.6491</td>
<td>0.6491</td>
<td>0.6530</td>
<td>0.6530</td>
<td>0.6231</td>
<td>0.6231</td>
</tr>
<tr>
<td>balance</td>
<td>0.6016</td>
<td>0.6016</td>
<td>0.6016</td>
<td>0.6016</td>
<td>0.5824</td>
<td>0.5824</td>
<td>0.5714</td>
<td>0.5714</td>
</tr>
<tr>
<td>segmentation</td>
<td>0.5657</td>
<td>0.5657</td>
<td>0.5657</td>
<td>0.5657</td>
<td>0.5810</td>
<td>0.5810</td>
<td>0.5419</td>
<td>0.5419</td>
</tr>
</tbody>
</table>
Part III: Graphical Models for Matrix Analysis

- Probabilistic Matrix Factorizations
- Probabilistic Co-clustering
- Stochastic Block Structures
Matrix Factorization

- Singular value decomposition

- Problems
  - Large matrices, with millions of row/columns
    - SVD can be rather slow
  - Sparse matrices, most entries are missing
    - Traditional approaches cannot handle missing entries
Matrix Factorization: “Funk SVD”

- Model $X \in \mathbb{R}^{n \times m}$ as $UV^T$ where
  - $U$ is a $\mathbb{R}^{n \times k}$, $V$ is $\mathbb{R}^{m \times k}$
  - Alternatively optimize $U$ and $V$

\[
\hat{X}_{ij} = u_i^T v_j = (X_{ij} - \hat{X}_{ij})^2 = (X_{ij} - u_i^T v_j)^2
\]
Matrix Factorization (Contd)

- Gradient descent updates
  
  \[ u_{ik}^{(t+1)} = u_{ik}^{(t)} + \eta \left( X_{ij} - \hat{X}_{ij} \right) v_{jk}^{(t)} \]

  \[ v_{jk}^{(t+1)} = v_{jk}^{(t)} + \eta \left( X_{ij} - \hat{X}_{ij} \right) u_{jk}^{(t)} \]

  \[ \hat{X}_{ij} = u_{i}^{T} v_{j} = \]

  \[ \text{error} = (X_{ij} - \hat{X}_{ij})^2 \]
Probabilistic Matrix Factorization (PMF)

\[ X_{ij} \sim N(u_i^T v_j, \sigma^2) \]

\[ u_i^T \sim N(0, \sigma_u^2 I) \]

\[ v_j \sim N(0, \sigma_v^2 I) \]

\[ R_{ij} \sim N(u_i^T v_j, \sigma^2) \]

\[ N(0, \sigma_v^2 I) \]

Inference using gradient descent
Bayesian Probabilistic Matrix Factorization

\[
X_{ij} \sim N(u_i^T v_j, \sigma^2)
\]

\[
\begin{align*}
\mu_u & \sim N(\mu_0, \Lambda_u), \quad \Lambda_u \sim W(\nu_0, W_0) \\
\mu_v & \sim N(\mu_0, \Lambda_v), \quad \Lambda_v \sim W(\nu_0, W_0) \\
u_i & \sim N(\mu_u, \Lambda_u) \\
v_j & \sim N(\mu_v, \Lambda_v) \\
R_{ij} & \sim N(u_i^T v_j, \sigma^2)
\end{align*}
\]

Graphical Models

Inference using MCMC
Results: PMF on the Netflix Dataset

![Graph showing RMSE vs Epochs for 10D and 30D Netflix Dataset]

10D

30D

Graphical Models 95
Results: PMF on the Netflix Dataset
Results: Bayesian PMF on Netflix

![Graph showing anomaly scores over epochs for Netflix, Baseline Score, SVD, PMF, and Logistic PMF.]
Results: Bayesian PMF on Netflix

Bayesian PMF

![Graph showing RMSE vs. Number of Samples for Bayesian PMF on Netflix with different sample sizes and associated training times.](image)
Results: Bayesian PMF on Netflix

![Graphical representation of RMSE vs. number of observed ratings for different models including Movie Average, Logistic PMF, and Bayesian PMF. The graph shows a decrease in RMSE as the number of observed ratings increases.]
Co-clustering: Gene Expression Analysis

Original

Co-clustered
Co-clustering and Matrix Approximation

<table>
<thead>
<tr>
<th>(U, V)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-66</td>
<td>54</td>
<td>-63</td>
<td>93</td>
<td>51</td>
<td>96</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>87</td>
<td>37</td>
<td>-26</td>
<td>84</td>
<td>-22</td>
</tr>
<tr>
<td>3</td>
<td>-68</td>
<td>56</td>
<td>-64</td>
<td>92</td>
<td>52</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>83</td>
<td>32</td>
<td>-24</td>
<td>80</td>
<td>-21</td>
</tr>
<tr>
<td>5</td>
<td>-63</td>
<td>55</td>
<td>-60</td>
<td>92</td>
<td>53</td>
<td>95</td>
</tr>
</tbody>
</table>

Original Matrix \(Z\)

Row Clustering

<table>
<thead>
<tr>
<th>(U, \hat{U})</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Low Parameter Matrix

<table>
<thead>
<tr>
<th>(\hat{U}, \hat{V})</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33.5</td>
<td>83.5</td>
<td>-23.3</td>
</tr>
<tr>
<td>2</td>
<td>-64.0</td>
<td>53.5</td>
<td>93.7</td>
</tr>
</tbody>
</table>

Column Clustering

<table>
<thead>
<tr>
<th>(V, \hat{V})</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Graphical Models
Probabilistic Co-clustering
Probabilistic Co-clustering
Generative Process

- Assume a mixed membership for each row and column
- Assume a Gaussian for each co-cluster

1. Pick row/column clusters
2. Generate each entry of the matrix
Reduction to Mixture Models
Reduction to Mixture Models
Generative Process

- Assume a mixed membership for each row and column
- Assume a Gaussian for each co-cluster

1. Pick row/column clusters
2. Generate each entry of the matrix
Bayesian Co-clustering (BCC)

- A Dirichlet distribution over all possible mixed memberships
Bayesian Co-clustering (BCC)

1. For each row $u, [u]_{1}^{n_1}$, choose $\pi_{1u} \sim \text{Dir}(\alpha_1)$.

2. For each column $v, [v]_{1}^{n_2}$, choose $\pi_{2v} \sim \text{Dir}(\alpha_2)$.

3. For each non-missing entry in row $u$ and column $v$:

   (a) Choose $z_1 \sim \text{Discrete}(\pi_{1u})$.

   (b) Choose $z_2 \sim \text{Discrete}(\pi_{2v})$.

   (c) Choose $x_{uv} \sim p(x | \theta_{z_1z_2})$.

\[
\log p(X | \alpha_1, \alpha_2, \Theta) \neq \sum_{n=1}^{N} \log p(x_n | \alpha_1, \alpha_2, \Theta)
\]
Learning: Inference and Estimation

• Learning
  – Estimate model parameters \((\alpha_1, \alpha_2, \theta)\)
  – Infer ‘mixed memberships’ of individual rows and columns

• Expectation Maximization
  - E-step: Calculate posterior probability \(p(\pi_1, \pi_2, z_1, z_2 | \alpha_1, \alpha_2, \Theta, X)\)
    to obtain log-likelihood \(L(\alpha, \Theta)\).
  - M-step: Maximize \(L(\alpha, \Theta)\) w.r.t \(\alpha, \Theta\).

• Issues
  – Posterior probability cannot be obtained in closed form
  – Parameter estimation cannot be done directly

• Approach: Approximate inference
  – Variational Inference
  – Collapsed Gibbs Sampling, Collapsed Variational Inference
Variational EM

- Introduce a variational distribution $q(\pi_1, \pi_2, z_1, z_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ to approximate $p(\pi_1, \pi_2, z_1, z_2 | \alpha_1, \alpha_2, \Theta, X)$

- Use Jensen’s inequality to get a tractable lower bound
  \[
  \log p(X | \alpha_1, \alpha_2, \Theta) \geq E_q [\log p(X, z_1, z_2, \pi_1, \pi_2 | \alpha_1, \alpha_2, \Theta)]
  + H(q(z_1, z_2, \pi_1, \pi_2))
  \]

- Maximize the lower bound w.r.t. $(\phi_1, \gamma_1, \phi_2, \gamma_2)$
  – Alternatively minimize the KL divergence between
    $q(\pi_1, \pi_2, z_1, z_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ and $p(\pi_1, \pi_2, z_1, z_2 | \alpha_1, \alpha_2, \Theta, X)$

- Maximize the lower bound w.r.t. $(\alpha_1, \alpha_2, \Theta)$
Variational Distribution

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

\[
q(\pi_1, \pi_2, z_1, z_2 | \gamma_1, \gamma_2, \phi_1, \phi_2) = \left( \prod_{u=1}^{n_1} q(\pi_{1u} | \gamma_{1u}) \right) \times \left( \prod_{v=1}^{n_2} q(\pi_{2v} | \gamma_{2v}) \right) \\
\times \left( \prod_{u=1}^{n_1} \prod_{v=1}^{n_2} q(z_{1uv} | \phi_{1u})q(z_{2uv} | \phi_{2v}) \right)
\]
Collapsed Inference

• Latent distribution can be exactly marginalized over \((\pi_1, \pi_2)\)
  – Obtain \(p(X,z_1,z_2|\alpha_1, \alpha_2, \beta)\) in closed form
  – Analysis assumes discrete/categorical entries
  – Can be generalized to exponential family distributions

• Collapsed Gibbs Sampling
  – Conditional distribution of \((z_{1uv}, z_{2uv})\) in closed form
    \[
p(z_{1uv}=i, z_{2uv}=j \mid X, z_{1-uv}, z_{2-uv}, \alpha_1, \alpha_2, \beta)
    \]
  – Sample states, run sampler till convergence

• Collapsed Variational Bayes
  – Variational distribution \(q(z_1, z_2|\gamma) = \prod_{u,v} q(z_{1uv}, z_{2uv}|\gamma^{uv})\)
  – Gaussian and Taylor approximation to obtain updates for \(\gamma^{uv}\)
Residual Bayesian Co-clustering (RBC)

\[ x_{uv} \sim N(x|\mu_{z_1z_2}, \sigma^2_{z_1z_2}) \]

\( (z_1, z_2) \) determines the distribution

\( \text{Users/movies may have bias} \)

\[ x_{uv} \sim N(x|\mu_{z_1z_2} + bm_{1u} + bm_{2v}, \sigma^2_{z_1z_2}) \]

\( (m_1, m_2): \) row/column means

\( (bm_1, bm_2): \) row/column bias
Results: Datasets

- **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), Real (z-scored)

- **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), Real (z-scored)

- **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), Real (z-scored)
## Perplexity Comparison with 10 Clusters

### On Binary Data

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MMNB</td>
<td>BCC</td>
</tr>
<tr>
<td>Jester</td>
<td>1.7883</td>
<td>1.8186</td>
</tr>
<tr>
<td>Movielens</td>
<td>1.6994</td>
<td>1.9831</td>
</tr>
<tr>
<td>Foodmart</td>
<td>1.8691</td>
<td>1.9545</td>
</tr>
</tbody>
</table>

### On Original Data

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MMNB</td>
<td>BCC</td>
</tr>
<tr>
<td>Jester</td>
<td>15.4620</td>
<td>18.2495</td>
</tr>
<tr>
<td>Movielens</td>
<td>3.1495</td>
<td>0.8068</td>
</tr>
<tr>
<td>Foodmart</td>
<td>4.5901</td>
<td>4.5938</td>
</tr>
</tbody>
</table>
Co-embedding: Users

User signatures

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>39</td>
<td>F</td>
<td>administrator</td>
</tr>
<tr>
<td>374</td>
<td>36</td>
<td>M</td>
<td>executive</td>
</tr>
<tr>
<td>470</td>
<td>24</td>
<td>M</td>
<td>programmer</td>
</tr>
<tr>
<td>933</td>
<td>28</td>
<td>M</td>
<td>student</td>
</tr>
</tbody>
</table>

User profiles.
Co-embedding: Movies

Movies

Movie signatures

<table>
<thead>
<tr>
<th>ID</th>
<th>Movie</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>498</td>
<td>The African Queen</td>
<td>American Expatriate, Boat, Mission, African Tribe</td>
</tr>
<tr>
<td>995</td>
<td>Kiss Me, Guido</td>
<td>Italian Food, Homosexual, Pizza, Gay Interest</td>
</tr>
<tr>
<td>1023</td>
<td>Fathers’ Day</td>
<td>Seduction, Con, Box Office Flop, Friendship</td>
</tr>
<tr>
<td>1233</td>
<td>Nénette et Boni</td>
<td>Brother Sister Relationship, Teen, Pregnancy, Teenage Pregnancy</td>
</tr>
</tbody>
</table>

Movie names and keywords.
RBC vs. other co-clustering algorithms

- RBC and RBC-FF perform better than BCC
- RBC and RBC-FF are also the best among others
RBC vs. other co-clustering algorithms

<table>
<thead>
<tr>
<th>$k_1, k_2$</th>
<th>SpecC2</th>
<th>SpecC5</th>
<th>BregC1</th>
<th>BregC2</th>
<th>BregC3</th>
<th>BregC4</th>
<th>BregC5</th>
<th>BregC6</th>
<th>BCC</th>
<th>RBC</th>
<th>RBC-FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,10</td>
<td>0.1175</td>
<td>0.0979</td>
<td>0.0956</td>
<td>0.1073</td>
<td>0.0949</td>
<td>0.1201</td>
<td>0.1073</td>
<td>0.1715</td>
<td>0.0957</td>
<td>0.0943</td>
<td>±0.0012</td>
</tr>
<tr>
<td></td>
<td>±0.0019</td>
<td>±0.0013</td>
<td>±0.0015</td>
<td>±0.0026</td>
<td>±0.0015</td>
<td>±0.0033</td>
<td>±0.0026</td>
<td>±0.0080</td>
<td>±0.0012</td>
<td>±0.0010</td>
<td></td>
</tr>
<tr>
<td>10,15</td>
<td>0.1141</td>
<td>0.0963</td>
<td>0.0959</td>
<td>0.0942</td>
<td>0.1173</td>
<td>0.1090</td>
<td>0.2603</td>
<td>0.0953</td>
<td>0.0935</td>
<td>0.0935</td>
<td>±0.0010</td>
</tr>
<tr>
<td></td>
<td>±0.0016</td>
<td>±0.0013</td>
<td>±0.0013</td>
<td>±0.0012</td>
<td>±0.0040</td>
<td>±0.0037</td>
<td>±0.0084</td>
<td>±0.0011</td>
<td>±0.0010</td>
<td>±0.0011</td>
<td></td>
</tr>
<tr>
<td>15,20</td>
<td>0.1136</td>
<td>0.0960</td>
<td>0.0944</td>
<td>0.1100</td>
<td>0.0954</td>
<td>0.1178</td>
<td>0.3399</td>
<td>0.0952</td>
<td>0.0931</td>
<td>0.0931</td>
<td>±0.0013</td>
</tr>
<tr>
<td></td>
<td>±0.0014</td>
<td>±0.0009</td>
<td>±0.0010</td>
<td>±0.0040</td>
<td>±0.0012</td>
<td>±0.0048</td>
<td>±0.0040</td>
<td>±0.1112</td>
<td>±0.0013</td>
<td>±0.0013</td>
<td></td>
</tr>
</tbody>
</table>

Movielens

<table>
<thead>
<tr>
<th>$k_1, k_2$</th>
<th>SpecC2</th>
<th>SpecC5</th>
<th>BregC1</th>
<th>BregC2</th>
<th>BregC3</th>
<th>BregC4</th>
<th>BregC5</th>
<th>BregC6</th>
<th>BCC</th>
<th>RBC</th>
<th>RBC-FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,5</td>
<td>0.9758</td>
<td>0.9159</td>
<td>0.9123</td>
<td>0.9765</td>
<td>0.9819</td>
<td>0.9855</td>
<td>0.9415</td>
<td>1.4148</td>
<td>0.9591</td>
<td>0.9119</td>
<td>±0.0196</td>
</tr>
<tr>
<td></td>
<td>±0.0221</td>
<td>±0.0199</td>
<td>±0.0194</td>
<td>±0.0212</td>
<td>±0.0217</td>
<td>±0.0221</td>
<td>±0.0154</td>
<td>±0.0168</td>
<td>±0.0212</td>
<td>±0.0197</td>
<td></td>
</tr>
<tr>
<td>15,10</td>
<td>0.9767</td>
<td>0.9170</td>
<td>0.9126</td>
<td>1.0178</td>
<td>1.0206</td>
<td>1.0269</td>
<td>0.9875</td>
<td>2.0442</td>
<td>0.9582</td>
<td>0.9111</td>
<td>±0.0202</td>
</tr>
<tr>
<td></td>
<td>±0.0214</td>
<td>±0.0191</td>
<td>±0.0200</td>
<td>±0.0236</td>
<td>±0.0237</td>
<td>±0.0239</td>
<td>±0.0229</td>
<td>±0.0262</td>
<td>±0.0217</td>
<td>±0.0204</td>
<td></td>
</tr>
<tr>
<td>20,15</td>
<td>0.9785</td>
<td>0.9187</td>
<td>0.9129</td>
<td>1.0561</td>
<td>1.0648</td>
<td>1.0670</td>
<td>1.0304</td>
<td>2.9876</td>
<td>0.9580</td>
<td>0.9106</td>
<td>±0.0198</td>
</tr>
<tr>
<td></td>
<td>±0.0209</td>
<td>±0.0192</td>
<td>±0.0196</td>
<td>±0.0227</td>
<td>±0.0222</td>
<td>±0.0164</td>
<td>±0.0217</td>
<td>±0.0505</td>
<td>±0.0215</td>
<td>±0.0217</td>
<td></td>
</tr>
</tbody>
</table>

Foodmart

Graphical Models 120
RBC vs. SVD, NNMF, and CORR

- RBC and RBC-FF are competitive with other algorithms

![Graph](image.png)

Jester

Graphical Models 121
### RBC vs. SVD, NNMF and CORR

<table>
<thead>
<tr>
<th>$k_1, k_2$</th>
<th>SVD</th>
<th>NNMF</th>
<th>CORR</th>
<th>RBC</th>
<th>RBC-FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,10</td>
<td>0.0986 ± 0.0012</td>
<td>0.1086 ± 0.0012</td>
<td>0.4118 ± 0.0061</td>
<td><strong>0.0943 ± 0.0012</strong></td>
<td><strong>0.0943 ± 0.0010</strong></td>
</tr>
<tr>
<td>10,15</td>
<td>0.0988 ± 0.0011</td>
<td>0.1078 ± 0.0013</td>
<td>0.4118 ± 0.0061</td>
<td><strong>0.0935 ± 0.0010</strong></td>
<td><strong>0.0935 ± 0.0011</strong></td>
</tr>
<tr>
<td>15,20</td>
<td>0.0991 ± 0.0011</td>
<td>0.1080 ± 0.0012</td>
<td>0.4118 ± 0.0061</td>
<td><strong>0.0931 ± 0.0013</strong></td>
<td><strong>0.0931 ± 0.0013</strong></td>
</tr>
</tbody>
</table>

### Movielens

<table>
<thead>
<tr>
<th>$k_1, k_2$</th>
<th>SVD</th>
<th>NNMF</th>
<th>CORR</th>
<th>RBC</th>
<th>RBC-FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,5</td>
<td><strong>0.8998 ± 0.0210</strong></td>
<td>0.9197 ± 0.0212</td>
<td>1.4528 ± 0.0281</td>
<td>0.9119 ± 0.0196</td>
<td>0.9136 ± 0.0197</td>
</tr>
<tr>
<td>15,10</td>
<td><strong>0.8995 ± 0.0208</strong></td>
<td>0.9216 ± 0.0207</td>
<td>1.4528 ± 0.0281</td>
<td>0.9111 ± 0.0202</td>
<td>0.9113 ± 0.0204</td>
</tr>
<tr>
<td>20,15</td>
<td><strong>0.9021 ± 0.0211</strong></td>
<td>0.9202 ± 0.0208</td>
<td>1.4528 ± 0.0281</td>
<td>0.9106 ± 0.0198</td>
<td>0.9112 ± 0.0217</td>
</tr>
</tbody>
</table>

### Foodmart
SVD vs. Parallel RBC

Parallel RBC scales well to large matrices
Inference Methods: VB, CVB, Gibbs

<table>
<thead>
<tr>
<th></th>
<th>Gibbs</th>
<th>CVB</th>
<th>VB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MovieLens</td>
<td>3.247</td>
<td>4.553</td>
<td>5.849</td>
</tr>
<tr>
<td>Binarized Jester</td>
<td>2.954</td>
<td>3.216</td>
<td>4.023</td>
</tr>
</tbody>
</table>

Graphical Models
Mixed Membership Stochastic Block Models

• Network data analysis
  – Relational View: Rows and Columns are the same entity
  – Example: Social networks, Biological networks
  – Graph View: (Binary) adjacency matrix

• Model

• For each node $p \in \mathcal{N}$:
  – Draw a $K$ dimensional mixed membership vector $\tilde{\pi}_p \sim \text{Dirichlet (} \tilde{\alpha} \text{)}$.

• For each pair of nodes $(p, q) \in \mathcal{N} \times \mathcal{N}$:
  – Draw membership indicator for the initiator, $\tilde{z}_{p \rightarrow q} \sim \text{Multinomial (} \tilde{\pi}_p \text{)}$.
  – Draw membership indicator for the receiver, $\tilde{z}_{q \rightarrow p} \sim \text{Multinomial (} \tilde{\pi}_q \text{)}$.
  – Sample the value of their interaction, $Y(p, q) \sim \text{Bernoulli (} \tilde{z}_{p \rightarrow q}^\top B \tilde{z}_{p \leftarrow q} \text{)}$.
MMB Graphical Model
Variational Inference

- Variational lower bound

\[ \log p(Y | \alpha, B) \geq \mathbb{E}_q \left[ \log p(Y, \vec{\pi}_{1:N}, Z_-, Z_\rightarrow | \alpha, B) \right] - \mathbb{E}_q \left[ \log q(\vec{\pi}_{1:N}, Z_-, Z_\rightarrow) \right] \]

- Fully factorized variational distribution

\[ q(\vec{\pi}_{1:N}, Z_-, Z_\rightarrow | \vec{\gamma}_{1:N}, \Phi_-, \Phi_\rightarrow) = \prod_p q_1(\vec{\pi}_p | \vec{\gamma}_p) \prod_{p,q} \left( q_2(\vec{z}_{p \rightarrow q} | \vec{\phi}_{p \rightarrow q}) q_2(\vec{z}_{p \leftarrow q} | \vec{\phi}_{p \leftarrow q}) \right) \]

- Variational EM
  - E-step: Update variational parameters (\(\gamma, \phi\))
  - M-step: Update model parameters (\(\alpha, B\))

Graphical Models 127
Results: Inferring Communities

Original friendship matrix

Friendships inferred from the posterior, respectively based on thresholding $\pi_p^T B \pi_q$ and $\varphi_p^T B \varphi_q$
Results: Protein Interaction Analysis

“Ground truth”: MIPS collection of protein interactions (yellow diamond)

Comparison with other models based on protein interactions and microarray expression analysis
Non-parametric Bayes

- Dirichlet Process Mixtures
- Gaussian Processes
- Hierarchical Dirichlet Processes
- Chinese Restaurant Processes
- Pittman-Yor Processes
- Mondrain Processes
- Indian Buffet Processes
References: Graphical Models

References: Inference


References: Mixed-Membership Models


References: Matrix Factorization

References: Co-clustering, Block Structures

Acknowledgements

Hanhua Shan

Amrudin Agovic
Thank you!