Diversity-Promoting Bayesian Learning of Latent Variable Models

Pengtao Xie\textsuperscript{1}, Jun Zhu\textsuperscript{1,2} and Eric Xing\textsuperscript{1}

\textsuperscript{1}Machine Learning Department, Carnegie Mellon University
\textsuperscript{2}Department of Computer Science and Technology, Tsinghua University
Latent Variable Models (LVMs)

Patterns

DATA

Hidden Markov Model, Topic Models, Restricted Boltzmann Machine, Deep Belief Network, Factor Analysis, Neural Network, Sparse Coding, Matrix Factorization, Distance Metric learning, Principal Component Analysis, etc.
Patterns and Components

Latent Patterns Behind Data

Themes in Documents

Politics
- Obama
- Constitution
- Government

Economics
- GDP
- Bank
- Marketing

Education
- University
- Knowledge
- Student

Components in LVMs

Topic Models

Groups in Images

Tiger

Car

Food

Gaussian Mixture Model
Challenges

- How to capture infrequent patterns?

- How to reduce model size (computational complexity) without compromising modeling power?

- How to prevent overfitting?
Diversify LVMs

- Encourage components to be diverse
- Goals

Infrequent-Patterns Extraction

Model Size Reduction without Losing Modeling Power

Overfitting Reduction

[Zou and Adams, 2012]
[Xie, et. al, 2015]
Diversify LVMs in Point Estimation

- Frequentist-style point estimation of LVMs
  - Component vectors (parameters) are deterministic
  - Single “best” estimate of parameters
  - Formulated as an optimization problem

- Diversity-promoting regularization
  - Define a regularizer that encourages components to be diverse
  - Use the regularizer to control LVMs

\[
\max_A L(D; A) + \lambda \Omega(A)
\]

- Examples: determinantal point process, mutual angular regularizer, inverse covariance

[Zou and Adams, 2012]
[Xie, et. al, 2015]
Bayesian Learning of LVMs

- Component vectors are random variables
- Infer a posterior distribution over the components

\[ p(A|D) = \frac{p(D|A)p(A)}{p(D)} \]

- Advantages over point estimation
  - Alleviates overfitting via model averaging
  - Quantify uncertainty
Diversify LVMs in Bayesian Learning

• Prior control: define diversity-biased priors and “propagate” the diversity to posterior via Bayes rule

\[
p(A \mid D) = \frac{p(D \mid A)p(A)}{p(D)}
\]

• Mutual Angular Prior
  • Angle-based notion of diversity: component vectors are deemed to be more diverse if their pairwise angles are larger.
  • Bayesian network and von Mises-Fisher distribution
  • Facilitate efficient (approximate) posterior inference
Mutual Angular Prior

- Decompose a component vector $\mathbf{a}_i$ into its magnitude $g = \|\mathbf{a}_i\|_2$ and direction $\tilde{\mathbf{a}}_i = \mathbf{a}_i / g$

- Develop a mutual angular prior over $\tilde{A} = \{\tilde{\mathbf{a}}_i\}_{i=1}^K$
  - Bayesian network: the parents of $\tilde{\mathbf{a}}_i$ are $\{\tilde{\mathbf{a}}_j\}_{j=1}^{i-1}$

  ![Diagram](image)

- Joint distribution $p(\tilde{A}) = p(\tilde{\mathbf{a}}_1) \prod_{i=2}^K p(\tilde{\mathbf{a}}_i | \text{pa}(\tilde{\mathbf{a}}_i))$

- Local probability distribution
  - Encourage larger mutual angles
  - von Mises-Fisher (vMF) distribution $p(\mathbf{x}) = C_p(\kappa) \exp(\kappa \mathbf{\mu}^T \mathbf{x}) \|\mathbf{\mu}\|_2 = 1 \|\mathbf{x}\|_2 = 1$

  \[
p(\tilde{\mathbf{a}}_i | \text{pa}(\tilde{\mathbf{a}}_i)) = C_p(\kappa) \exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\sum_{j=1}^{i-1} \|\tilde{\mathbf{a}}_j\|_2^2})^T \tilde{\mathbf{a}}_i)
  \]
Mutual Angular Prior (cont’d)

- Prior over magnitude \( G = \{g_i\}_{i=1}^K \)

\[
p(G) = \prod_{i=1}^K p(g_i)
\]

- \( p(g_i) \) is gamma distribution

- Put \( p(\tilde{A}) \) and \( p(G) \) together
  - Generative process
    - Sample \( \tilde{a}_1 \sim vMF(\mu_0, \kappa) \)
    - For \( i = 2, \cdots, K \), sample \( \tilde{a}_i \sim vMF(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2}, \kappa) \)
    - For \( i = 1, \cdots, K \), sample \( g_i \sim \text{Gamma}(\alpha_1, \alpha_2) \)
    - For \( i = 1, \cdots, K \), \( a_i = \tilde{a}_i g_i \)

- Joint distribution

\[
p(A) = C_p(\kappa) \exp(\kappa \mu_0^T \tilde{a}_1) \prod_{i=2}^K C_p(\kappa) \exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \prod_{i=1}^K \frac{\alpha_1^\alpha g_i^{\alpha_1-1} e^{-g_i \alpha_2}}{\Gamma(\alpha_1)}
\]
Posterior Inference

- Posterior Inference: compute $p(A|D)$
- Exact solution is intractable
- Variational inference: an approximate inference method
  - Minimize the KL divergence between the true posterior and a “simple” variational distribution
- Maximize a variational lower bound

$$\sup_{q(A)} \mathbb{E}_{q(A)} \left[ \log p(D|A)p(A) \right] - \mathbb{E}_{q(A)} \left[ \log q(A) \right]$$
Variational Inference

\[ p(\tilde{a}_i | \text{pa}(\tilde{a}_i)) = C_p(\kappa) \exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T\tilde{a}_i) \]

Not amenable for variational inference
Variational Inference (cont’d)

\[ p(\tilde{a}_i | pa(\tilde{a}_i)) = C_p(\kappa)\exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \]

Not amenable for variational inference

Reparametrize \( p(\tilde{a}_i | pa(\tilde{a}_i)) \)

\[
\begin{align*}
p(\tilde{a}_i | pa(\tilde{a}_i)) & \propto \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{a}_j)^T \tilde{a}_i) \\
& \propto \exp(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2 (-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \\
& = C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2)\exp(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2 (-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \\
p(\tilde{a}_i | pa(\tilde{a}_i)) & = C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2)\exp(\kappa(-\sum_{j=1}^{i-1} \tilde{a}_j)^T \tilde{a}_i)
\end{align*}
\]
Variational Inference (cont’d)

\[ p(\tilde{a}_i | \text{pa}(\tilde{a}_i)) = C_p(\kappa)\exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \]

Not amenable for variational inference

Reparametrize \( p(\tilde{a}_i | \text{pa}(\tilde{a}_i)) \)

\[
p(\tilde{a}_i | \text{pa}(\tilde{a}_i)) \propto \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{a}_j)^T \tilde{a}_i) \\
\propto \exp(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \\
= C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2) \exp(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2(-\frac{\sum_{j=1}^{i-1} \tilde{a}_j}{\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2})^T \tilde{a}_i) \\

p(\tilde{a}_i | \text{pa}(\tilde{a}_i)) = C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2) \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{a}_j)^T \tilde{a}_i) \\
\]

\[
C_p(x) = \frac{x^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(x)} \\
I_v(x) = (x/2)^v \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{k!\Gamma(v+k+1)}
\]
Variational Inference (cont’d)

Upper bound of the log of partition function

\[
Z_i = \frac{1}{C_p(\kappa \| \sum_{j=1}^{i-1} \tilde{a}_j \|_2)}
\]

\[
\gamma + \left[ \log(1 + e^{-\xi}) + \frac{\xi - \gamma}{2} + \frac{1/2 - g(\xi)}{2\xi} (\xi^2 - \gamma^2) \right] \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \\
- \frac{1/2 - g(\xi)}{2\xi} \kappa^2 \| \sum_{j=1}^{i-1} \tilde{a}_j \|_2^2 \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}
\]

Lower bound of \( \mathbb{E}_{q(A)}[\log p(A)] \)

\[
\mathbb{E}_{q(A)}[\log p(A)] \\
\geq \kappa A_p(\hat{\kappa}) \mu_0^T \hat{a}_1 + \sum_{i=2}^{K} (-\kappa A_p^2(\hat{\kappa}) \sum_{j=1}^{i-1} \hat{a}_j^T \hat{a}_i - \gamma_i) \\
- \left[ \log(1 + e^{-\xi_i}) + \frac{\xi_i - \gamma_i}{2} + \frac{1/2 - g(\xi_i)}{2\xi_i} (\xi_i^2 - \gamma_i^2) \right] \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \\
+ \frac{1/2 - g(\xi_i)}{2\xi_i} \kappa^2 (A_p^2(\hat{\kappa}) \sum_{j=1}^{i-1} \sum_{k \neq j} \hat{a}_j^T \hat{a}_k \\
+ \sum_{j=1}^{i-1} (\text{tr}(\frac{h(\hat{\kappa})}{\hat{\kappa}} I + (1 - 2 \frac{\nu+1}{\hat{\kappa}} h(\hat{\kappa}) - h^2(\hat{\kappa})) \hat{a}_j \hat{a}_j^T)) \\
+ A_p^2(\hat{\kappa}) \hat{a}_j^T \hat{a}_j) \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}) + K(\alpha_1 \log \alpha_2 - \log \Gamma(\alpha_1)) \\
+ \sum_{i=1}^{K} (\alpha_1 - 1)(\psi(r_i) - \log(s_i)) - \alpha_2 \frac{r_i}{s_i} + \text{const}
\]
Diversity-Promoting Bayesian Learning of MoE

\[
p(B) = C_p(\kappa) \exp(\kappa \mu_0^T \tilde{\beta}_1) \prod_{i=2}^{K} C_p(\kappa) \exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{\beta}_j}{\|\sum_{j=1}^{i-1} \tilde{\beta}_j\|^2})^T \tilde{\beta}_i) \prod_{i=1}^{K} \alpha_i^{\alpha_1} g_i^{\alpha_i-1} e^{-g_i \alpha_2} / \Gamma(\alpha_1)
\]

\[
p(H) = C_p(\kappa) \exp(\kappa \xi_0^T \tilde{\eta}_1) \prod_{i=2}^{K} C_p(\kappa) \exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{\eta}_j}{\|\sum_{j=1}^{i-1} \tilde{\eta}_j\|^2})^T \tilde{\eta}_i) \prod_{i=1}^{K} \omega_i^{\omega_1} h_i^{\omega_i-1} e^{-h_i \omega_2} / \Gamma(\omega_1)
\]
Results

Diversification improves classification accuracy

Table 1: Classification accuracy (%) on the Adult-9 dataset

<table>
<thead>
<tr>
<th>K</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMoE-G</td>
<td>83.4</td>
<td>84.2</td>
<td>84.9</td>
<td>84.9</td>
</tr>
<tr>
<td>BMoE-MABN-I</td>
<td>87.1</td>
<td>88.3</td>
<td>88.6</td>
<td>88.9</td>
</tr>
<tr>
<td>BMoE-MABN-II</td>
<td>86.4</td>
<td>87.8</td>
<td>88.1</td>
<td>88.4</td>
</tr>
</tbody>
</table>

Table 2: Classification accuracy (%) on SUN-Building dataset

<table>
<thead>
<tr>
<th>K</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMoE-G</td>
<td>76.5</td>
<td>78.6</td>
<td>80.2</td>
<td>80.4</td>
</tr>
<tr>
<td>BMoE-MABN-I</td>
<td>82.1</td>
<td>83.6</td>
<td>85.3</td>
<td>85.2</td>
</tr>
<tr>
<td>BMoE-MABN-II</td>
<td>80.9</td>
<td>82.8</td>
<td>84.9</td>
<td>84.1</td>
</tr>
</tbody>
</table>
Results

Diversification reduces model size without losing modeling power.

Table 1: Classification accuracy (%) on the Adult-9 dataset

<table>
<thead>
<tr>
<th>K</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMoE-G</td>
<td>83.4</td>
<td>84.2</td>
<td>84.9</td>
<td>84.9</td>
</tr>
<tr>
<td>BMoE-MABN-I</td>
<td>87.1</td>
<td>88.3</td>
<td>88.6</td>
<td>88.9</td>
</tr>
<tr>
<td>BMoE-MABN-II</td>
<td>86.4</td>
<td>87.8</td>
<td>88.1</td>
<td>88.4</td>
</tr>
</tbody>
</table>

Table 2: Classification accuracy (%) on SUN-Building dataset

<table>
<thead>
<tr>
<th>K</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMoE-G</td>
<td>76.5</td>
<td>78.6</td>
<td>80.2</td>
<td>80.4</td>
</tr>
<tr>
<td>BMoE-MABN-I</td>
<td>82.1</td>
<td>83.6</td>
<td>85.3</td>
<td>85.2</td>
</tr>
<tr>
<td>BMoE-MABN-II</td>
<td>80.9</td>
<td>82.8</td>
<td>84.9</td>
<td>84.1</td>
</tr>
</tbody>
</table>
Results

- Diversification effectively captures infrequent patterns

<table>
<thead>
<tr>
<th>Category ID</th>
<th>C18</th>
<th>C17</th>
<th>C12</th>
<th>C14</th>
<th>C22</th>
<th>C34</th>
<th>C23</th>
<th>C32</th>
<th>C16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of Docs</td>
<td>5281</td>
<td>4125</td>
<td>1194</td>
<td>741</td>
<td>611</td>
<td>483</td>
<td>262</td>
<td>208</td>
<td>192</td>
</tr>
<tr>
<td>BMEM-G Accuracy (%)</td>
<td>87.3</td>
<td>88.5</td>
<td>75.7</td>
<td>70.1</td>
<td>71.6</td>
<td>64.2</td>
<td>55.9</td>
<td>57.4</td>
<td>51.3</td>
</tr>
<tr>
<td>BMEM-MABN-I Accuracy (%)</td>
<td>88.1</td>
<td>86.9</td>
<td>74.7</td>
<td>72.2</td>
<td>70.5</td>
<td>67.3</td>
<td>68.9</td>
<td>70.1</td>
<td>65.5</td>
</tr>
<tr>
<td>Relative Improvement (%)</td>
<td>1.0</td>
<td>-1.8</td>
<td>-1.3</td>
<td>2.9</td>
<td>-1.6</td>
<td>4.6</td>
<td>18.9</td>
<td>18.1</td>
<td>21.7</td>
</tr>
</tbody>
</table>

Table 5: Classification accuracy on 9 sub-categories of the CCAT category in the RCV1.Binary dataset

Frequent
Infrequent
Conclusions

- Promoting diversity in Bayesian latent variable models
  - Capture infrequent patterns
  - Reduce model size without sacrificing modeling power
  - Prevent overfitting
- Mutual angular prior
  - Angle-based notion of diversity
  - Bayesian network and von Mises-Fisher distribution
  - Facilitate approximate posterior inference
- Variational inference
  - Reparametrize local probability distribution
  - Derive upper bound of the log partition function
- Experiments corroborate the efficacy of our methods
Thank you!