Bayesian Learning of Joint Distributions of Objects

Anjishnu Banerjee    Jared Murray    David B. Dunson

Statistical Science, Duke University

02 May 2014

*Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (2013)*

Presented by Kyle Ulrich
Motivation

- Dependent data of disparate data types is often observed
  - Real numbers
  - Counts
  - Categorical data
  - More complex objects – functions, shapes, and images

- This domain is known as *mixed domain modeling* (MDM)

- The modelling objectives for these types of observations are diverse
  - Learn dependence between data types
  - Co-clustering
  - Prediction

- Recent interest in modelling such data sets has had a broad focus on joint Dirichlet Process mixture models
Overview

- Review of Sethuraman’s representation of Dirichlet Process
- Joint Dirichlet process mixtures
- Problems with joint DP mixtures
- *Infinite tensor factorization mixture* as an alternative
- Results
The Sethuraman (1994) stick-breaking representation of the Dirichlet process mixture model is written as:

\[
f = \sum_{h=1}^{\infty} \pi_h \mathcal{K}(\theta^*_h) \tag{1}
\]

\[
\pi_h = V_h \prod_{l<h} (1 - V_l) \quad V_h \sim \text{Beta}(1, \alpha) \tag{2}
\]

\[
\theta^*_h \sim P_0 \tag{3}
\]

and an observation \( y_i \sim f \) for subjects \( i = 1, \ldots, N \)

Here, \( \pi = \{\pi_h\} \) is drawn from a stick-breaking process, \( \pi \sim \text{GEM}(\alpha) \)

For efficient computation, a latent variable model is crucial:

\[
y_i \sim \mathcal{K}(\theta^*_{C_i}), \quad \theta^*_h \sim P_0, \quad \text{pr}(C_i = h) = \pi_h \tag{4}
\]

where \( C_i \) is the cluster index for subject \( i \)
Joint Dirichlet Process Mixture (joint DPM)

- Assume we have observations $y_i = \{y_{i1}, \ldots, y_{iJ}\}$ for subjects $i = 1, \ldots, N$
- Furthermore, the elements of $y_i = \{y_{ij}\}$ are potentially drawn from different parametric densities
- We wish to cluster subjects
- The **joint Dirichlet process mixture** (Dunson & Bhattacharya, 2012) is one possible formulation:

\[
y_{ij} \sim K_j(\theta_{C_i}^*), \quad \theta_h^* \sim P_{0j}, \quad \Pr(C_i = h) = \pi_h
\]  

(5)

- Two subjects are either allocated to the same cluster *globally*, or not clustered
- *Conditional independence* given a single latent class variable
Problems with Joint DPM Approach

• Underlying problem: conditional independence assumption
• There are several disadvantages to this assumption:
  • To realistically model joint distributions across many variables, unnecessary clusters may be introduced
  • Poor performance for small sample sizes
  • Over-clustering leads to misleading inferences and poor prediction
  • Tends to favor certain components of the data
• Overcome these problems by allowing a separate, but dependent index for disparate data types:

\[
\text{pr}(C_{i1} = h_1, \ldots, C_{ip} = h_p) = \pi_{h_1,\ldots,h_p}
\]

(6)

where \( h_j = 1, \ldots, \infty \) and \( j = 1, \ldots, p \)
• Here, \( \pi = \{\pi_{h_1,\ldots,h_p}\} \in \prod_p^\infty \) is an infinite \( p \)-way probability tensor
Probabilistic Parafac Factorization

- Parafac = parallel factor analysis = canonical polyadic (CP) decomposition = generalization of SVD to tensors
- A possible Parafac tensor factorization of $\pi$ (Dunson & Xing, 2009):
  \[
  \text{pr}(C_{i1} = h_1, \ldots, C_{ip} = h_p) = \pi_{h_1,\ldots,h_p} \tag{7}
  \]
  \[
  \pi = \sum_{h=1}^{k} \lambda_h \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)} \tag{8}
  \]
  \[
  \lambda \sim \text{GEM}(\alpha) \tag{9}
  \]
  where $C_{ij} \in \{1, \ldots, d_j\}$, $d_j$ is known, and $\psi^{(j)}_h = (\psi^{(j)}_{h1}, \ldots, \psi^{(j)}_{hd_j})^T$ represents the probability vector for component $h$ and outcome $j$
- What if $C_i$ is unobserved and we allow $d_j \to \infty$?
With a small change, the authors allow infinitely many levels:

\[
\text{pr}(C_{i1} = h_1, \ldots, C_{ip} = h_p) = \pi_{h_1, \ldots, h_p} \tag{10}
\]

\[
\pi = \sum_{h=1}^{k} \lambda_h \bigotimes_{j=1}^{p} \psi_h^{(j)} \tag{11}
\]

\[
\lambda \sim \text{GEM}(\alpha) \tag{12}
\]

\[
\psi_h^{(j)} \sim \text{GEM}(\beta_j) \tag{13}
\]

which is a stick-breaking mixture of outer products of stick-breaking processes

- Elements of \( \pi \) are stochastically larger in the smallest indexed cells
- For shorthand, the above model is written \( \pi \sim \text{ITF}(\alpha, \beta) \)
Here, $y_i$ is drawn from an infinite tensor factorization mixture model:

$$y_{ij} \sim \mathcal{K}_j(\theta^*_{ij})$$  \hspace{1cm} (14)

$$\theta^*_i \sim \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1,\ldots,h_p} \prod_{j=1}^{p} \delta_{\theta_j,h_j}$$  \hspace{1cm} (15)

$$\pi \sim \text{ITF}(\alpha, \beta)$$  \hspace{1cm} (16)

$$\theta_{j,h_j} \sim P_{0j}$$  \hspace{1cm} (17)

where $\theta^*_i = (\theta^*_{i1}, \ldots, \theta^*_{ip})$

Now, the elements of $y_i = \{y_{ij}\}$ are dependent
Simulated Data Examples

- Data was generated for 1000 individuals comprising of
  - $T$, a time series
  - $R$, a multivariate real-valued response ($\in \mathbb{R}^4$)
  - $C_1, C_2, C_3$, 3 different categorical variables

- In Scenario 1, the entire ensemble ($T, R, C_1, C_2, C_3$) was drawn from a single component of a 3-component mixture

- In Scenario 2, a dependency structure is introduced

- The goal is to test both prediction and inference of dependency structure

### Table 1: Simulation Example, Scenario 1: Prediction error (top), tests of independence (bottom)

<table>
<thead>
<tr>
<th></th>
<th>ITF</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>1.79</td>
<td>1.43</td>
</tr>
<tr>
<td>C2</td>
<td>31%</td>
<td>23%</td>
</tr>
<tr>
<td>C3</td>
<td>37%</td>
<td>36%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ITF</th>
<th>DPM</th>
<th>“Truth”</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$ vs $T$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$C_2$ vs $T$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$C_3$ vs $T$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$C_2$ vs $R$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Table 2: Simulation Example, Scenario 2: Prediction error (top), tests of independence (bottom)

<table>
<thead>
<tr>
<th></th>
<th>ITF</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>4.61</td>
<td>10.82</td>
</tr>
<tr>
<td>C2</td>
<td>27%</td>
<td>55%</td>
</tr>
<tr>
<td>C3</td>
<td>34%</td>
<td>57%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ITF</th>
<th>DPM</th>
<th>“Truth”</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$ vs $T$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$C_2$ vs $T$</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>$C_3$ vs $T$</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>$C_2$ vs $R$</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Political Blog Example

- Data collected from 105 political blogs
- Includes the network (web link connectivity), blog's ideology label, and binary indicators for 7 labelling sources

**Figure:** Left: True clustering. Right: Recovered clustering.
Questions?