Completely Random Measures, Hierarchies and Nesting

Michael I. Jordan
University of California, Berkeley

March 4, 2010

Acknowledgments: Emily Fox, Erik Sudderth, Yee Whye Teh, and Romain Thibaux
Document Corpora

• “Bag-of-words” models of document corpora have a long history in the field of information retrieval
• The bag-of-words assumption is an exchangeability assumption for the words in a document
  – it is generally a very rough reflection of reality
  – but it is useful, for both computational and statistical reasons
• Examples
  – text (bags of words)
  – images (bags of patches)
  – genotypes (bags of polymorphisms)
• Motivated by De Finetti, we wish to find useful latent representations for these “documents”
Outline

- Finite mixture models
- Finite admixture models (*latent Dirichlet allocation*)
- Bayesian nonparametric admixture models (*the hierarchical Dirichlet process*)
- Abstraction hierarchy mixture models (*nested Dirichlet process*)
- Abstraction hierarchy admixture models (*nested beta process*)
Finite Mixture Models

• The mixture components are distributions on individual words in some vocabulary (e.g., for text documents, a multinomial over lexical items)
  – often referred to as “topics”
• The generative model of a document:
  – select a mixture component
  – repeatedly draw words from this mixture component
• The mixing proportions are corpora-specific, not document-specific
• Major drawback: each document can express only a single topic
Finite Admixture Models

• The mixture components are distributions on individual words in some vocabulary (e.g., for text documents, a multinomial over lexical items)
  – often referred to as “topics”
• The generative model of a document:
  – repeatedly select a mixture component
  – draw a word from this mixture component
• The mixing proportions are document-specific
• Now each document can express multiple topics
Finite Admixture Models

- To allow the mixing proportion to be document-specific, it is natural to specify a distribution on mixing proportions
- Leads one toward a Bayesian perspective
  - although one could view the mixing proportions as random effects
  - the posterior distribution on the mixing proportions can be viewed as the representation of a document in the context of a given corpus
- Multiple origins (e.g, Pritchard, Stephens, and Donnelly; Erosheva; Blei, Ng and Jordan)
Latent Dirichlet Allocation

(Blei, Ng, and Jordan, 2003)

- A word is represented as a multinomial random variable $w$
- A topic allocation is represented as a multinomial random variable $z$
- A document is modeled as a Dirichlet random variable $\theta$
- The variables $\alpha$ and $\beta$ are hyperparameters
The Topic Simplex

• Each corner of the simplex corresponds to a topic allocation – a component of the vector $z$

• A document is modeled as a point in the simplex – a set of mixing proportions over topics

• A corpus is modeled as a Dirichlet distribution on the simplex
Some Limitations of Finite Admixture Models

• How many mixture components?
  – the problem is more severe here than in classical finite mixture models because of the capability of capturing additional heterogeneity in a corpus
  – $K$ needs to be the union of the cardinality of the sets of topics expressed across documents

• No notion of abstraction hierarchy
  – documents often mix over words at different levels of abstraction in addition to mixing over topics
Dirichlet Process Mixture Models

• A Bayesian nonparametric approach to letting the number of components be random and subject to inference

• A draw from a Dirichlet process (DP) is (wp1) a probability measure with support on a countably infinite number of atoms:

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \]

where

\[ \pi \sim \text{GEM}(\alpha_0) \]

\[ \phi_k \overset{iid}{\sim} G_0 \]

denoted: \[ G \sim \text{DP}(\alpha_0 G_0) \]

• Convolving a DP with a kernel yields a DP mixture. Note that the DP determines both the mixing proportions and the parameters of the mixing components.
Dirichlet Process Mixture Models

\[
G \sim \text{DP}(\alpha_0 G_0)
\]

\[\theta_i \mid G \sim G \quad i \in 1, \ldots, n\]

\[x_i \mid \theta_i \sim F_{\theta_i} \quad i \in 1, \ldots, n\]
Clustering in DP Mixture Models

- In the model specification, each data point \( x_i \) is associated with one of the atoms \( \phi_k \).
- Data points are said to be in the same cluster if they are associated with the same atom.
- The probability of two independent draws from \( G \) yielding the same atom is:
  \[
  \frac{1}{\alpha_0 + 1}
  \]
- This fact is nicely captured via the Chinese restaurant process.
Chinese Restaurant Process (CRP)

- A random process in which \( n \) customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - \( m \)th subsequent customer sits at a table drawn from the following distribution:
    
    \[
    P(\text{previously occupied table } i \mid \mathcal{F}_{m-1}) \propto n_i \\
    P(\text{the next unoccupied table } \mid \mathcal{F}_{m-1}) \propto \alpha_0
    \]
  
  - where \( n_i \) is the number of customers currently at table \( i \) and where \( \mathcal{F}_{m-1} \) denotes the state of the restaurant after \( m - 1 \) customers have been seated
The CRP and Mixture Modeling

- Data points are customers; tables are mixture components
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables
- This prior can be completed with:
  - a likelihood---e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters---the $\phi_k$th customer to sit at table $k$ chooses the parameter vector, $\phi_k$, for that table from a prior $G_0$

- This prior can be viewed as a marginal distribution under the DP mixture model
Dirichlet Process Admixture Models?

- Recall that in an admixture model (aka, topic model), we repeatedly draw the mixing proportions from a prior, once for each document.
- If we were to use the DP to do this, we would obtain (wp1) disjoint sets of atoms for the different documents.
- So a given topic could only be expressed in a single document; that’s not the spirit of admixture models.
- We want to share topics among documents.
Hierarchical Modeling

• Standard nonparametric model:

\[ G \mid \alpha, G_0 \sim \mathcal{P}(\alpha, G_0) \]
\[ \theta_i \mid G \sim G \]
\[ x_i \mid \theta_i \sim F_{\theta_i} \]

• Hierarchical nonparametric model:

\[ G_0 \mid \gamma, H \sim \mathcal{P}(\gamma, H) \]
\[ G \mid \alpha, G_0 \sim \mathcal{Q}(\alpha, G_0) \]
\[ \theta_i \mid G \sim G \]
\[ x_i \mid \theta_i \sim F_{\theta_i} \]
Hierarchical Dirichlet Process Mixtures

(Teh, Jordan, Beal & Blei, JASA, 2006)

\[ G_0 | \gamma, H \sim DP(\gamma H) \]
\[ G_i | \alpha, G_0 \sim DP(\alpha_0 G_0) \]
\[ \theta_{ij} | G_i \sim G_i \]
\[ x_{ij} | \theta_{ij} \sim F(x_{ij} | \theta_{ij}) \]
Chinese Restaurant Franchise (CRF)
Application: Protein Modeling

- A protein is a folded chain of amino acids
- The backbone of the chain has two degrees of freedom per amino acid (phi and psi angles)
- Empirical plots of phi and psi angles are called Ramachandran diagrams
Application: Protein Modeling

- We want to model the density in the Ramachandran diagram to provide an energy term for protein folding algorithms.
- We actually have a linked set of Ramachandran diagrams, one for each amino acid neighborhood.
- We thus have a *linked set* of density estimation problems.
Protein Folding (cont.)

- We have a linked set of Ramachandran diagrams, one for each amino acid neighborhood.
Protein Folding (cont.)

Marginal improvement over finite mixture

- o hdp: right
- ▲ additive model

ALA ARG ASN ASP CPR CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL
Completely Random Measures

(Kingman, 1967)

- Completely random measures are random measures on a set $\Omega$ that assign independent mass to nonintersecting subsets of $\Omega$
  - e.g., Brownian motion, gamma processes, beta processes, compound Poisson processes and limits thereof
- The Dirichlet process is not a completely random measure
- Completely random processes are discrete wp1 (up to a possible deterministic continuous component)
- Completely random processes are random measures, not necessarily random probability measures
Completely Random Measures

(Kingman, 1968)

- Assigns independent mass to nonintersecting subsets of $\Omega$

\[\Omega\]
Completely Random Measures

(Kingman, 1967)

- Consider a non-homogeneous Poisson process on $\Omega \otimes R$ with rate function obtained from some product measure
- Sample from this Poisson process and connect the samples vertically to their coordinates in $\Omega$
Beta Processes

- The product measure is called a *Levy measure*
- For the beta process, this measure lives on $\Omega \otimes (0, 1)$ and is given as follows:

  \[ \nu(d\omega, dp) = cp^{-1}(1 - p)^{c-1} dp B_0(d\omega) \]

  degenerate Beta(0,c) distribution  Base measure

- And the resulting random measure can be written simply as:

  \[ B = \sum_i p_i \delta_{\omega_i} \]
Beta Processes

\[ B = \sum_i p_i \delta \omega_i \]
Gamma Processes, Dirichlet Processes and Bernoulli Processes

- The gamma process uses an improper gamma density in the Levy measure
- Once again, the resulting random measure is an infinite weighted sum of atoms
- To obtain the Dirichlet process, simply normalize the gamma process
- The Bernoulli process is obtained by treating the beta process as an infinite collection of coins and tossing those coins:

\[ G = \sum_{k=1}^{\infty} z_k \delta_{\omega_k} \quad \text{where} \quad z_k \sim \text{Be}(p_k) \]
Beta Process and Bernoulli Process

Concentration $c = 10$  Mass $\gamma = 2$
BP and BeP Sample Paths

\( \gamma = 2 \) \hspace{1cm} \gamma = 5 \hspace{1cm} \gamma = 10 \\

\( c = .1 \) \\

\( c = 1 \) \\

\( c = 10 \)
Indian Buffet Process (IBP)
(Griffiths & Ghahramani, 2002)

• Indian restaurant with infinitely many dishes in a buffet line

• Customers 1 through $n$ enter the restaurant
  – the first customer samples $\text{Poisson}(\alpha)$ dishes
  – the $i$th customer samples a previously sampled dish with probability $m_k/(i+1)$ then samples $\text{Poisson}(\alpha/i)$ new dishes
Indian Buffet Process (IBP)
(Griffiths & Ghahramani, 2002)

- Indian restaurant with infinitely many dishes in a buffet line
- Customers 1 through $n$ enter the restaurant
  - the first customer samples $\text{Poiss}(\alpha)$ dishes
  - the $i$th customer samples a previously sampled dish with probability $m_k/(i + 1)$ then samples $\text{Poiss}(\alpha/i)$ new dishes
Indian Buffet Process (IBP)  
(Griffiths & Ghahramani, 2002)

• Indian restaurant with infinitely many dishes in a buffet line

• Customers 1 through $n$ enter the restaurant
  – the first customer samples $\text{Poiss}(\alpha)$ dishes
  – the $i$th customer samples a previously sampled dish with probability $m_k/(i+1)$ then samples $\text{Poiss}(\alpha/i)$ new dishes
Indian Buffet Process (IBP)
(Griffiths & Ghahramani, 2002)

- Indian restaurant with infinitely many dishes in a buffet line
- Customers 1 through $n$ enter the restaurant
  - the first customer samples $\text{Poiss}(\alpha)$ dishes
  - the $i$th customer samples a previously sampled dish with probability $m_k/(i+1)$ then samples $\text{Poiss}(\alpha/i)$ new dishes
Indian Buffet Process (IBP)
(Griffiths & Ghahramani, 2002)

- Indian restaurant with infinitely many dishes in a buffet line
- Customers 1 through $n$ enter the restaurant
  - the first customer samples $\text{Poiss}(\alpha)$ dishes
  - the $i$th customer samples a previously sampled dish with probability $m_k/(i + 1)$ then samples $\text{Poiss}(\alpha/i)$ new dishes
The IBP and the Beta Process
(Thibaux & Jordan, 2007)

• *Theorem*: The beta process is the De Finetti mixing measure underlying the Indian buffet process (IBP)
Motion Capture Analysis
Multiple Time Series

• Goals:
  – transfer knowledge among related time series in the form of a library of “behaviors”
  – allow each time series model to make use of an arbitrary subset of the behaviors

• Method:
  – represent behaviors as states in an autoregressive HMM
  – use the beta/Bernoulli process to pick out subsets of states
BP-AR-HMM

- Bernoulli process determines which states are used

- Beta process prior:
  - encourages sharing
  - allows variability

\[ \pi_j^{(i)} \mid f_i, \gamma, \kappa \sim \text{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots] \otimes f_i) \]

\[ z_t^{(i)} \sim \pi_{z_{t-1}}^{(i)} \]

\[ y_t^{(i)} = \sum_{j=1}^{r} A_{j,z_t^{(i)}} y_{t-j}^{(i)} + e_t^{(i)}(z_t^{(i)}) \]
Motion Capture Results
Hierarchical Beta Processes

- A hierarchical beta process is a beta process whose base measure is itself random and drawn from a beta process.

\[ B \sim \text{BP}(c_0, B_0) \]

\[ j = 1, \ldots, n \]

\[ A_j \sim \text{BP}(c_j, B) \]

\[ i = 1, \ldots, n_j \]

\[ X_{ij} \sim \text{BeP}(A_j) \]
Abstraction Hierarchies

• Words in documents are organized not only by topics but also by level of abstraction
• Admixture models such as LDA don’t capture this notion; common words often appear repeatedly across many LDA topics
• Idea: *Let documents be represented as paths down a tree of topics, placing topics focused on common words near the top of the tree*
Nesting

- In the hierarchical models, all of the atoms are available to all of the “groups”
- In nested models, the atoms are subdivided into smaller collections, and the groups select among the collections
- E.g., the nested CRP of Blei, Griffiths and Jordan
  - each table in the Chinese restaurant indexes another Chinese restaurant
- E.g., the nested DP of Rodriguez, Dunson and Gelfand
  - each atom in a high-level DP is itself a DP
Nested Chinese Restaurant Process

(Blei, Griffiths and Jordan, JACM, 2010)
Nested Chinese Restaurant Process
Nested Chinese Restaurant Process
Nested Chinese Restaurant Process
Nested Chinese Restaurant Process
Nested Chinese Restaurant Process
Hierarchical Latent Dirichlet Allocation

• The generative model for a document is as follows:
  – use the nested CRP to select a path down the infinite tree
  – draw $\lambda \sim \text{GEM}(\lambda_0)$ to obtain a distribution on levels along the path
  – repeatedly draw a level from $\lambda$ and draw a word from the topic distribution at that level
Nested DP

(Rodriguez, Dunson and Gelfand, 2005)

• A mixture model in which each component is itself a mixture model:

\[ G \sim \sum_{k=1}^{\infty} \pi_k^* \delta G_k \]

\[ G_k = \sum_{j=1}^{\infty} \pi_{kj} \delta \phi_{kj} \]

• Note that the atoms in \( G_k \) are distinct from those in \( G_{k'} \).

• Extending the nested DP to an arbitrary recursion and integrating out the random measures yields the nested CRP.
Pros and Cons of Hierarchical LDA

- The hierarchical LDA model yields more interpretable topics than flat LDA.
- The maximum a posteriori tree can be interesting.
- But the model is out of the spirit of classical topic models because it only allows one “theme” per document.
- We would like a model that can both mix over levels in a tree and mix over paths within a single document.
Nested Beta Processes, Nested Gamma Processes

- In the nested DP/nested CRP, at each level of the tree the procedure selects only a single outgoing branch.
- Instead of using the DP/CRP, use the beta process (or the gamma process) together with the Bernoulli process (or the Poisson process) to allow multiple branches to be selected:

\[ B \sim \text{BeP} \left( \sum_{k=1}^{\infty} p_k^* \delta_{B_k^*} \right) \]

\[ B_k^* = \sum_{j=1}^{\infty} p_{k,j} \delta_{\theta_{k,j}} \]
Nested Beta Processes, Nested Gamma Processes
Nested Beta Processes, Nested Gamma Processes
Nested Beta Processes, Nested Gamma Processes
Nested Beta Processes, Nested Gamma Processes

- Results on a corpus of C. Elegans abstracts
- Held-out marginal log likelihood

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDP</td>
<td>-1011.34</td>
</tr>
<tr>
<td>nCRP</td>
<td>-1019.28</td>
</tr>
<tr>
<td>nGP</td>
<td>-1007.92</td>
</tr>
</tbody>
</table>
Conclusions

• Hierarchies and nesting have at least an important role to play in Bayesian nonparametrics as they play in classical Bayesian parametric modeling.

• In particular, infinite-dimensional parameters can be controlled with these strategies; this yields desirable statistical control.

• Completely random measures are an important tool in constructing nonparametric priors.

• For more details:
  
  www.cs.berkeley.edu/~jordan/publications.html