A Short Course on Bayesian Nonparametrics

Lecture 3 - Computation for Dirichlet process mixture models

Abel Rodriguez - UC, Santa cruz

Universidade Federal Do Rio de Janeiro
March, 2011

Challenges in working with posterior samples from a DPM

- Identifiability (label switching).
    - Computation: Multimodality.
    - Interpretation: Focus on inferences on identifiable parameters.
- How to summarize the information in the posterior:
    - Posterior inference for functionals of G, including $F(x) = \int \psi(x|\theta)dG(\theta)$.
    - Clustering structure $\Rightarrow$ What is a cluster?

Label switching

- The likelihood is invariant to the labels used for the components. Stephens (2000) provides an excellent review in the context of finite mixtures.
  - For example, if $n = 5$, $\xi_1 = 1$, $\xi_2 = 1$, $\xi_3 = 2$, $\xi_4 = 2$, $\xi_5 = 3$ implies exactly the same model as $\xi_1 = 2$, $\xi_2 = 2$, $\xi_3 = 3$, $\xi_4 = 3$, $\xi_5 = 1$.
  - For fixed K, the posterior has $K!$ identical modes, corresponding to each of the $K!$ copies of the space (one for each order of the labels).

Label switching

- Computational implications: Are we exploring all these modes? A two-part answer:
  - The collapsed Gibbs sampler (and some of the other samplers we will discuss today) act on the equivalence classes associated with the label switching ⇒ Two labelings $\{\xi_i\}$ and $\{\xi_i^*\}$ belong to the equivalence class if the induce the same partition of the observations.
  - Even if we are not, who cares? (as long as inference is made on identifiable functions of the parameters).
- In general, the interpretation of the parameters is not consistent from one iteration of the MCMC to the next ⇒ This is true even if "identifiability" constrains are introduced in the mixture!!!

Summarizing the posterior

- The collapsed Gibbs sampler integrates G out of the model.
- If we need to do inferences for functionals of G (which are identifiable!!!), note that G follows a mixture of Dirichlet processes (MDP) (Antoniak, 1974)

$$G|y_1, \ldots, y_n \sim \int \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n} H_\eta + \sum_{k=1}^{K} \frac{m_k}{\alpha+n}\delta_{\vartheta_k}\right) dP(\{\xi_i\}, \{\vartheta_k\}, \alpha, \eta | y_1, \ldots, y_n)$$

- The DPM and the MDP are different!!!

DPM	MDP
$y_i \sim \int \psi(y_i|\theta)dG(\theta) \quad G \sim DP(\alpha, H)$	$y_i \sim G \quad G \sim \int \mathrm{DP}(\alpha, H_\eta)dP(\alpha, \eta)$
$\Downarrow$	$\Downarrow$
G models the parameters	G models the observations

Linear functionals of G

- Consider first computing summaries for linear functionals of G. For example, using Fubinni's rule

$$
\begin{aligned}
\hat{f}(y) &= \mathsf{E}_{G|y_1,\ldots,y_n}\left\{ \int \psi(y|\theta) dG(\theta) \right\} \\
&= \int \int_{\Theta} \psi(y|\theta) \left\{ \frac{\alpha}{\alpha+n} H_\eta(d\theta) + \sum_{k=1}^{K} \frac{m_k}{\alpha+n} \delta_{\vartheta_k}(d\theta) \right\} \\
&\qquad\qquad dP(\{\xi_i\}, \{\vartheta_k\}, \alpha, \eta | y_1, \ldots, y_n) \\
&= \int \left\{ \sum_{k=1}^{K} \frac{m_k}{\alpha+n} \psi(y|\vartheta_k) + \frac{\alpha}{\alpha+n} \int \psi(y|\vartheta_k) H_\eta(d\theta) \right\} \\
&\qquad\qquad dP(\{\xi_i\}, \{\vartheta_k\}, \alpha, \eta | y_1, \ldots, y_n)
\end{aligned}
$$

- Same expression we obtained for $p(y_{n+1|}|y_1, \ldots, y_n)$!!!

Non-linear functionals of G

- Hence, point estimates for $F(y)$ can be easily obtained without any need to explicitly sample G.
- Similar story for (for example) $\mathrm{E}(y) = \int y dF(y)$ or, more generally, $\mathrm{E}(y^d) = \int y^d dF(y)$
- How about non-linear functionals (for example, $F^{-1}(\gamma)$ the $\gamma \in (0,1)$ quantile)?
  - Note that $\mathrm{E}_{G|y_1,\ldots,y_n}\{F^{-1}(\gamma)\} \neq \hat{F}^{-1}(\gamma)$!!!
  - If we had samples from $p(G|y_1, \ldots y_n)$, we could transform them into samples from $p(\lambda_\gamma|y_1, \ldots, y_n)$ where $\lambda_\gamma = F^{-1}(\gamma|G)$

Non-linear functionals of G

- Samples from G can in principle be obtained by using the stick-breaking construction and the fact that

$$G|y_1,\ldots,y_n \sim \int \mathrm{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}H_\eta + \sum_{k=1}^{K}\frac{m_k}{\alpha+n}\delta_{\vartheta_k}\right) dP(\{\xi_i\},\{\vartheta_k\},\alpha,\eta|y_1,\ldots,y_n)$$

- Since G involves an infinite number of atoms, use instead a finite approximation G_{N_ϵ} for small ϵ (Kottas & Gelfand, 2002).

$$G^{(b)}_{N_\epsilon} = \sum_{k=1}^{N_\epsilon^{(b)}} \omega_k^{*(b)}\delta_{\vartheta_k^{*(b)}} \qquad \vartheta_k^{*(b)} \sim_{iid} \frac{\alpha^{(b)}}{\alpha^{(b)}+n}H_{\eta^{(b)}} + \sum_{k=1}^{K^{(b)}}\frac{m_k^{(b)}}{\alpha^{(b)}+n}\delta_{\vartheta_k^{(b)}}$$

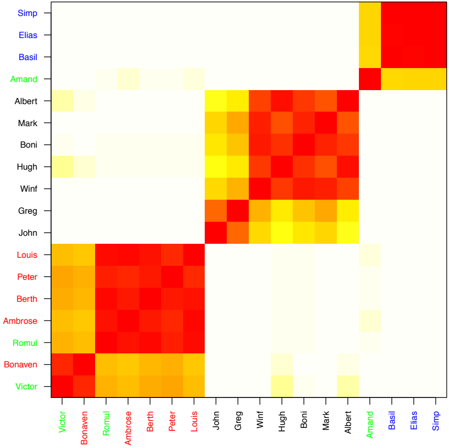
$(\omega_1^{(b)},\omega_2^{(b)},\ldots) \sim \mathrm{SB}(\alpha^{(b)}+n)$, and $N_\epsilon^{(b)}$ is such that it satisfies $\sum_{k=1}^{N_\epsilon^{(b)}}\omega_k^{*(b)} \geq 1-\epsilon$.

Clustering structure

- Summaries for $p(\{\xi_i\}|y) \Rightarrow$ Non-euclidean space.
- The vector $\{\xi_i\}$ can alternatively be represented by a $n \times n$ matrix T such that $T_{i,j} = 1$ iif $\xi_i = \xi_j$ and $T_{i,j} = 0$ otherwise ($T_{i,i} = 1$ by convention).
- $T \Rightarrow$ Incidence Matrix. Invariant to label switching (and hence, identifiable).
- $\hat{T} = \mathrm{E}(T|y_1, \dots, y_n) \Rightarrow \hat{T}_{i,j}$ is the marginal posterior probability that y_i and y_j are assigned to the same cluster.

Clustering structure

- T can be represented as an image plot.
- Gives an idea of both a point estimator and uncertainty.
- Reordering the observations to get a readable picture is important.



Point estimator

- A utility based approach to clustering (Lau & Green, 2007).
- Start with a utility function:

$$U(\hat{\xi}, \xi) = \sum_{i} \sum_{j<i} \left\{ a\mathbf{1}_{(\hat{\xi}_i \neq \hat{\xi}_i, \xi_i = \xi_j)} + b\mathbf{1}_{(\hat{\xi}_i = \hat{\xi}_i, \xi_i \neq \xi_j)} \right\}$$

- a and b are the costs of “Type I ” and “Type II” errors.
- Maximizing the expected utility is equivalent to maximizing

$$\hat{U}(\hat{\xi}) = \sum_{i} \sum_{j<i} \mathbf{1}_{(\hat{\xi}_i \neq \hat{\xi}_j)} \left\{ \hat{T}_{i,j} - \frac{b}{a+b} \right\}$$

- $b/(a+b) = 0 \Rightarrow$ one cluster, $b/(a+b) = 1 \Rightarrow n$ clusters.

The DPM as a clustering algorithm

- We originally motivated the DPM as a prior on random distributions.
- However, due to the a.s. nature of G, it allows for flexible clustering and automatic selection of the number of clusters.
- Some caveats:
  - For fixed K, the DP favors *a priori* partitions with uneven-sized clusters $\Rightarrow$ A few big clusters together with many very small clusters.
  - The "shape" of the clusters is determined by the kernel $\Rightarrow$ Multivariate normal kernels imply spherical clusters $\Rightarrow$ This problem is shared by all model-based clustering algorithms, but exacerbated for the DPM!

Homework

1. Modify the collapsed sampler you already implemented to work with a Poisson kernel $\psi(y_i|\theta) = \text{Poi}(y_i|\theta)$ and $H(\theta) = \text{Gam}(\nu, \theta_0)$. Also, use a Gamma hyperprior on θ_0 and a $\Gamma(1,1)$ prior on α.
2. What is the limiting behavior of this model when $\alpha \to \infty$?
3. To highlight possible problems with the DPM as a clustering mechanism, simulate an iid sample $y_1, \ldots, y_{50}$ with $y_i \sim \text{NegBin}(20, 1/3)$. Fit the Poisson mixture model to this data. How would you choose the parameters of the baseline measure? What would you expect to get?
4. Find the optimal clustering structure induced by this model if $a = b = 1$. How would you interpret these results? Can you describe a more appropriate model for this problem?
5. (This example was originally suggested by Mike Escobar).

Why other samplers?

- Sample ξ_is from full conditionals:
  - Collapsed samplers tend have high autocorrelations.
  - Slow in creating new components.
- The collapsed sampler we described works only for conjugate models.
- Some extensions of the DP do not have simple Pólya urn representations.
- We work with the basic mixture

$$y_i|\theta_i \sim \psi(y_i|\theta_i) \qquad \theta_i|G \sim G \qquad G \sim \mathrm{DP}(\alpha, H)$$

Split-Merge Metropolis Hastings

- Key references: Jain & Neal, 2004, 2007; Dahl, 2003.
- Motivation ⇒ Improve mixing by moving multiple observations at a time when creating new components.
- We focus only on conjugate models.
- Let $\mathcal{S}_{n,K} = \{S_1, \ldots, S_K\}$ be a partition of $\{1, \ldots, n\}$, i.e.,
  - $\cup_{k=1}^{K} S_k = \{1, \ldots, n\}$.
  - $S_k \cap S_j = \emptyset$ for $k \neq j$.
- You can recover $\mathcal{S}_{n,K}$ from $\{\xi_i\}$ (but, because of label switching, there are many sets $\{\xi_i\}$ that lead to the same $\mathcal{S}_{n,K}$).

Split-Merge Metropolis Hastings

- The prior on partitions implied by the CRP is:

$$\Pr(\mathcal{S}_{n,K}) = \frac{\alpha^K}{\prod_{i=1}^{n}(\alpha + i - 1)} \prod_{k=1}^{K} (|S_k| - 1)!$$

- With a conjugate model, we can integrate out the θ_is, so that

$$p(y_1, \ldots, y_n | \mathcal{S}_{n,K}) = \prod_{k=1}^{K} \int \left\{ \prod_{j \in S_k} \psi(y_j|\theta) \right\} dH(\theta)$$

- Hence, the posterior is

$$\Pr(\mathcal{S}_{n,K}|y_1, \ldots, y_n) \propto \left\{ \prod_{k=1}^{K} \int \left[\prod_{j \in S_k} \psi(y_j|\theta) \right] dH(\theta) \right\} \alpha^{K-1} \prod_{k=1}^{K} (|S_k| - 1)!$$

Split-Merge Metropolis Hastings

- This posterior is defined on a HUGE discrete space. How can we explore it? ⇒ Metropolis-Hastings: propose a change in the partition, and accept or reject it.
  - Move one observation at a time ⇒ We only have $K + 1$ choices, so we can compute the probabilities associated with each new placement ⇒ Collapsed Gibbs sampler!!!!
  - More than one observation at atime: Split-Merge moves ⇒ Either take an existing group and split it, or take two an merge them.
  - There are MANY ways in which this can be done, from the very naive, to the smart. We focus on Dahl, 2003.

Sequentially allocated split-mergue

- Uniformly select a pair of indices i and j.
- If i, j are in the same component of $\mathcal{S}_{n,K}$ (say S_k) then split
  - Remove indices i and j from S and form $S_{k'} = \{i\}$ and $S_{k''} = \{j\}$.
  - Do random permutation of the indexes remaining in S_k.
  - Sequentially add index $r \in S_k$ to sets $S_{k'}$ or $S_{k''}$ with probabilities

$$\Pr(r \in S_{k'} | S_{k'}, S_{k''}, y) \propto |S_{k'}| \int \psi(y_r|\theta) p(\theta | y_{S_{k'}}) d\theta$$

$$\Pr(r \in S_{k''} | S_{k'}, S_{k''}, y) \propto |S_{k''}| \int \psi(y_r|\theta) p(\theta | y_{S_{k''}}) d\theta$$

  - Eliminate component S_k.

Sequentially allocated split-mergue

- If i and j are in different components of $\mathcal{S}_{n,K}$ (say $S_{k'}$ and $S_{k''}$), mergue:
  - Form a merged component $S_k = S_{k'} \cup S_{k''}$ and eliminate $S_{k'}$ and $S_{k''}$.
- To compute the acceptance ratio, note that the split and the merge steps are the corresponding reversible steps.
- The probability of proposing the mergue step is proportional to 1, the probability of proposing a split is proportional to the product of the proposal probabilities corresponding to the sequential allocations.

Collapsed samplers for non-conjugate models

- Neal (2000) provides an excellent review.
- Use same ideas as for split-mergue algorithms. If the ϑ_ks are not integrated out of the model the posterior looks like

$$p(\{\xi_i\}, \{\vartheta_k\}|\mathbf{y}) \propto p(\xi_1, \ldots, \xi_n) \prod_{k=1}^{K} G_0(\vartheta_k) \prod_{i=1}^{n} \psi(y_i|\vartheta_{\xi_i})$$

  where $p(\xi_1, \ldots, \xi_n) \propto \alpha^{K-1} \prod_{k=1}^{K} (m_k - 1)!$ just as before.
- A variety of samplers can be obtained by using different proposals that simultaneously change $\{\xi_i\}$ and $\{\vartheta_k\}$.

Collapsed samplers for non-conjugate models

- A simple example: For each i, propose $\vartheta_k^{(p)} = \vartheta_k^{(c)}$ for $k \leq K^{(c)}$, $\vartheta_{K^{(c)}+1}^{(p)} \sim H$, and $\xi_i^{(p)}$ from

$$\Pr(\xi_i^{(p)} = k | \xi_i^{(c)} = k', \{\vartheta_k^{(p)}\}) \propto \begin{cases} m_k \psi(y_i|\vartheta_k) & k \neq k', k \leq K \\ (m_{k'} - 1)\psi(y_i|\vartheta_{k'}) & k = k', k \leq K \\ \alpha\psi(y_i|\vartheta_{K+1}^{(p)}) & k = K + 1 \end{cases}$$

- The acceptance probability is 1!!!!.
- This looks a lot like the collapsed sampler for the conjugate case, but we use the likelihood evaluated on samples of the θ_ks (rather than the marginal likelihoods) to construct the probability of each component.
- Better mixing if multiple new components are used!

Blocked Gibbs samplers

- Introduced by Ishwaran & James, 2001.
- Approximate $G = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_k}$ with $G^N = \sum_{k=1}^{N} \omega_k \delta_{\vartheta_k}$ for large enough N.
- The weights for G^N are constructed just like G, but letting $z_N = 1$.
- For a sample $y = (y_1, \ldots, y_n)$, we have

$$\int \left| \int \psi(y|\theta) dG(\theta) - \int \psi(y|\theta) dG^N(\theta) \right| dy \leq \int |dG(\theta) - dG^N(\theta)| \leq 4 \left\{ 1 - \left[1 - \left(\frac{\alpha}{\alpha + 1} \right)^{N-1} \right]^n \right\}$$

Blocked Gibbs samplers

- The prior on $(\omega_1, \ldots, \omega_N)$ implied by the truncation is

$$p(\omega_1, \ldots, \omega_N | \alpha) = \alpha^{N-1} \omega_N^{\alpha-1} (1 - \omega_1)^{-1} \\ (1 - \{\omega_1 + \omega_2\})^{-1} \cdots \left(1 - \sum_{k=1}^{N-2} \omega_k\right)^{-1}$$

- Finite mixture model with a Generalized Dirichlet prior on the weights $\Rightarrow$ We can use samplers for finite mixture models.
- By conditioning on the θ_is, the ξ_is become conditionally independent $\Rightarrow$ Better mixing (!?)
- Easy(!?) to sample when ψ and H are not conjugate $\rightarrow$ Gibbs/Metropolis steps (no need for a direct sampler for H).
- Much simpler to implement, and no need to do anything fancy for inferences on G!!!

Blocked Gibbs samplers

- Sample ξ_i from

$$\Pr(\xi_i = k \mid \cdots) \propto \omega_k \psi(y_i \mid \vartheta_k) \qquad\qquad k = 1, \ldots, N$$

- Sample ϑ_k from

$$p(\vartheta_k \mid \cdots) \propto \left\{ \prod_{\{i:\xi_i = k\}} \psi(y_i \mid \vartheta_k) \right\} H(\vartheta)$$

- Sample $(\omega_1, \ldots, \omega_N)$ by first sampling $\{z_k\}$

$$z_k \mid \cdots \sim \text{beta}\left(1 + m_k, \alpha + \sum_{l=k+1}^{N} m_l\right) \qquad\qquad m_k = \sum_{i=1}^{n} \mathbf{1}_{(\xi_i = k)}$$

  and setting $\omega_k = z_k \prod_{l<k} \{1 - z_l\}$.

Slice samplers

- Introduced in Walker, (2007).
- Start with the representation of the DPM as an uncountable mixture

$$y_i | \{\omega_k\}, \{\vartheta_k\} \sim_{iid} \sum_{k=1}^{\infty} \omega_k \psi(y_i | \vartheta_k)$$

- Data augmentation ⇒ Introduce uniform random variables $u_1, \ldots, u_n$ and define

$$y_i, u_i | \{\omega_k\}, \{\vartheta_k\} \sim_{iid} \sum_{k=1}^{\infty} \mathbf{1}_{(u_i \leq \omega_k)} \psi(y_i | \vartheta_k)$$

If you marginalize u_i you recover the first expression!!

Slice samplers

- Data augmentation (again) ⇒ Introduce indicators $\xi_1, \ldots, \xi_n$ and define

  $$y_i, u_i, \xi_i | \{\omega_k\}, \{\vartheta_k\} \sim_{iid} \mathbf{1}_{(u_i \leq \omega_{\xi_i})} \psi(y_i | \vartheta_{\xi_i})$$

  If you marginalize both ξ_i and u_i you recover the countable mixture representation.
- Joint distribution

  $$p(\{y_i\}, \{u_i\}, \{\xi_i\} | \{\omega_k\}, \{\vartheta_k\}) = \prod_{i=1}^{n} \left\{ \mathbf{1}_{(u_i \leq \omega_{\xi_i})} \psi(y_i | \vartheta_{\xi_i}) \right\}$$
- Full conditionals on this extended model are easy to obtain.

Slice samplers

- The samplers for $\{\omega_k\}$ and $\{\theta_k\}$ are the same as for the blocked Gibbs sampler.

$$p(\vartheta_k | \cdots) \propto \left\{ \prod_{\{i:\xi_i=k\}} \psi(y_i|\vartheta_k) \right\} H(\vartheta)$$

$$z_k | \cdots \sim \text{beta}\left(1 + m_k, \alpha + \sum_{l=k+1}^{N} m_l\right) \qquad m_k = \sum_{i=1}^{n} \mathbf{1}_{(\xi_i=k)}$$

with $\omega_k = z_k \prod_{l<k}\{1 - z_l\}$.

- For the "slice" variables $u_i | \cdots \sim \text{Uni}[0, \omega_{\xi_i}]$

Slice samplers

- For the indicator variables $\{\xi_i\}$.

$$\Pr(\xi_i = k | \cdots) \propto \mathbf{1}_{w_k > u_i} \psi(y_i | \vartheta_k)$$

- In principle, this implies an infinite number of terms. However, for each i, only a finite number of the ω_ks are such that $\omega_k > u_i$. Hence,

$$\Pr(\xi_i = k | \cdots) = \frac{\{\mathbf{1}_{w_k > u_i}\} \psi(y_i | \vartheta_k)}{\sum_{\{l : \omega_l > u_i\}} \psi(y_i | \vartheta_l)}$$

- In general, we need to represent explicitly only a finite number of components N such that $1 - \sum_{k=1}^{N} \omega_k < \min\{u_i\} \Rightarrow$ Adaptive truncation of the mixture.

Comparison among samplers

	$\Pr(\xi_i = k \mid \cdots) \propto$	$p(\vartheta_k \mid \cdots) \propto$	$p(\omega \mid \cdots)$
Trunc	(Fixed size) $\omega_k \psi(y_i \mid \vartheta_k)$	$\left\{ \prod_{\{j:\xi_j=k\}} \psi(y_j \mid \vartheta_k, \phi) \right\} h_\eta(\vartheta_k)$	$\omega_k = z_k \prod_{l<k} \{1 - z_l\}$ $z_k \sim \text{beta}(1 + m_k, \alpha + s_{k+1})$
Slice	(Variable size) $\mathbf{1}_{\omega_k > u_i} \psi(y_i \mid \vartheta_k)$	Same	Same
Colla	(Variable size) $\begin{cases} m_k^{-i} \psi(y_i \mid \vartheta_k^{-i}) & k \leq K^{-i} \\ \alpha p(y_i \mid \eta) & k = K^{-i} + 1 \end{cases}$	Same	Not needed (integrated out)
Sp-Me	Sampled in block (varies)	Same	Not needed (integrated out)

Comparison among samplers

	A.S. Trunc	Slice	Collapsed	Split-Merge
Easy to code	Easy	Easy to Moder	Moder	Moder to Hard
Mixing	Moder to Good	Moder to Good	Moder	Good
Inference on G	Easy	Easy	Moder	Moder
Approx (beyond MC)	Yes	No	No	No
Memory requir	Large	Large	Moder	Moder

Some general comments

- A sceptic's view ⇒ The truncated version of the MDP suggests that there is nothing really new about DPs that had not been discovered with finite mixtures.
- This is not quite fair
    - The link highlights that we need to be careful with how we pick the prior on the ω_ks if we want the model to automatically select the number of mixture components.
    - What if you want a large number of components? (Poisson / Neg Binomial example).
- Also there is a difference between the theoretical properties of an infinite model which is only truncated for computational reasons, and one that is truncated from the start.
- For the algorithms that truncate the model (either almost surely or adaptively), it is better if ϕ, α and η are sampled as if the parameters from the occupied components came from a collapsed Gibbs sampler.

Other options

- Retrospective samplers (Roberts and Papaspiliopoulos, 2008) $\Rightarrow$ Another form of adaptive truncation.
- Variational algorithms (Blei and Jordan, 2006) $\Rightarrow$ Replace the intractable posterior by a simpler form that is tractable, and optimize its parameters to minimize KL distance.
- Sequential Monte Carlo and particle filters (MacEachern et al, 1999; Carvalho et al, 2009).

Homework

1. Divide the class en three groups.
2. For the location mixture of normals that has been the running example, have each group implement one of the following:
    1. Blocked Gibbs sampler.
    2. Slice Sampler.
    3. Non-conjugate collapsed sampler (work as if ψ and H were not conjugate).
3. Compare the performance of the algorithms against that of the collapsed Gibbs sampler on the `galaxy` dataset (available as part of `DPpackage`).