Bayesian Statistics

Debdeep Pati
Florida State University

October 6, 2016

Collapsed / Marginal Gibbs sampler (Escober & West, 1995)

- For density estimation,consider the DP mixture (DPM)model

$$y_i \mid \mu_i, \tau_i \sim N(\mu_i, \tau_i^{-1}), \theta_i = (\mu_i, \tau_i) \sim P, P \sim \mathrm{DP}(\alpha P_0)(\cdot)$$

- Not immediate clear how to conduct posterior computation
- One strategy relies on marginalizing out P to obtain

$$(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0 + \sum_{j=1}^{i-1} \frac{1}{\alpha + i - 1} \delta_{\theta_j}$$

Computation of $(\theta_1, \dots, \theta_n) \mid \mathbf{y}$

- Computational methods like the Gibbs sampler will require the conditional distributions of $\theta_i \mid \mathbf{y}, \theta^{-i}$.
- conditional distribution of θ_i given $(\mathbf{y}, \theta^{-i})$ is proportional to

$$
\begin{aligned}
& N(y_i, \theta_i)\Big(\sum_{j \neq i} \delta_{\theta_j^{-i}}(d\theta_i) + \alpha G_0(d\theta_i)\Big) \\
= & \sum_{j \neq i} N(y_i; \theta_j^{-i})\delta_{\theta_j^{-i}}(d\theta_i) + \alpha N(y_i; \theta_i) G_0(\theta_i) \\
= & \sum_{j \neq i} N(y_i; \theta_j^{-i})\delta_{\theta_j^{-i}}(d\theta_i) + \alpha N(y_i, G_0) \frac{N(y_i; \theta_i) G_0(\theta_i)}{N(y_i, G_0)}
\end{aligned}
$$

where $N(y_i, G_0) = \int N(y_i; \theta_i) G_0(\theta_i) d\theta_i$.

- The normalizing constant is $\sum_{j \neq i} N(y_i; \theta_j^{-i}) + \alpha N(y_i; G_0)$ is available in closed form.

Computation of $f(y_{n+1} \mid \mathbf{y})$

- Let $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$.

$$\begin{aligned} f(y_{n+1} \mid \mathbf{y}) &= \int f(y_{n+1} \mid \mathbf{y}, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \\ &\approx \frac{1}{M} \sum_{t=1}^{M} f(y_{n+1} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)}) \end{aligned}$$

-

$$\begin{aligned} f(y_{n+1} \mid \mathbf{y}, \boldsymbol{\theta}) &= \int f(y_{n+1} \mid \mathbf{y}, \boldsymbol{\theta}, \theta_{n+1}) f(\theta_{n+1} \mid \boldsymbol{\theta}, \mathbf{y}) d\theta_{n+1} \\ &= \int f(y_{n+1} \mid \theta_{n+1}) f(\theta_{n+1} \mid \boldsymbol{\theta}) d\theta_{n+1} \\ &= \frac{\alpha}{n+\alpha} \int N(y_{n+1}; \theta_{n+1}) G_0(\theta_{n+1}) d\theta_{n+1} + \\ & \quad \frac{1}{n+\alpha} \sum_{j=1}^{n} N(y_{n+1}; \theta_j) \qquad (1) \end{aligned}$$

- Draw samples from the posterior $\boldsymbol{\theta} \mid \mathbf{y}$ and plug in (1) at each step of the Gibbs sampling.

Improved Collapsed Gibbs Sampler (Bush & MacEachern, 96)

- Let $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ denote the unique values of θ.
- Let $S_i = h$ if $\theta_i = \theta_h^*$ denote allocation of subject i to cluster h
- Let $k^{(-i)}$ is the number of unique values in $\theta^{(-i)}$ and $n_h^{(-i)}$ are the corresponding counts
- Gibbs sampler alternates between
  1. Update the allocation $S = (S_1, \ldots, S_n)'$ by sampling from multinomial with

$$P(S_i = h \mid -) \propto \begin{cases} n_h^{(-i)} N(y_i; \theta_h^*), h = 1, \ldots, k^{(-i)} \\ \alpha \int N(y_i; \theta) dP_0(\theta), h = k^{(-i)} + 1 \end{cases}$$

  2. Update the unique values of θ^* by sampling

$$(\mu_h^*, \tau_h^{*,-1} | -) = \mathrm{N}(\mu_h, \hat{\mu}_h, \hat{\kappa}_h \tau_h^{-1}) \mathrm{Ga}(\tau_h, \hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with parameters defined as in the finite mixture model case

Marginal Gibbs Sampler - Some Comments

- Only slightly more complicated the Gibbs sampling for finite mixture models
- Unless further collapsing is done, the chain might be "sticky" and prediction is more complicated
- \# mixture components k represented in the sample of n subjects is unknown
- From the MCMC samples, we can estimate posterior distribution of k
- As subjects are added k will increase stochastically
- To estimate the predictive density of y_{n+1} use

$$f(y) = \sum_{h=1}^{k} \frac{n_h}{n+\alpha} N(y; \theta_h^*) + \frac{\alpha}{n+\alpha} \int N(y; \theta) dP_0(\theta)$$

averaged over MCMC iterations after burn-in.

Clustering & Label Ambiguity via the Dirichlet Process

- Clustering via the Dirichlet Process
- If we let $y_i \sim f$,with f assigned the prior described above, then

$$y_i \sim \mathrm{N}(\mu_{S_i}, \tau_{S_i}^{-1}), S_i \sim \sum_{h=1}^{k} \pi_h \delta_h$$

  where S_i is a cluster index for subject i and $(\mu_h, \tau_h) \sim P_0$ independently.
- $(\pi_1, \dots, \pi_k) \sim \mathrm{Dir}(\alpha/k, \dots, \alpha/k)$
- $\{\theta_h = (\mu_h, \tau_h), h = 1, \dots, k\}$ are component specific parameters.

Estimating component specific parameters and Label Ambiguity

- Note that the labels $\{1, \ldots, k\}$ are treated as exchangeable in the above mixture model
- There is nothing in the prior or likelihood distinguishing mixture component (cluster) h and cluster h'
- Hence, the true marginal posterior distribution of θ_h (the parameters specific to component/cluster h) will be identical for all $h \in \{1, \ldots, k\}$.
- Each of these marginals can be expected to be multi-modal

Problems with Label Ambiguity & MCMC

- Due to the multi-modality of the posterior distributions, the Gibbs sampler described above will have a tendency to get stuck for long intervals in local modes
- This "stickiness" depends strongly on the separation between the different components
- If the components are widely separated, then one may obtain an apparently unimodal posterior for each θ_h and the Gibbs sampling trace plots may seem well behaved
- For example, if $k = 2$ with one component close to $\mu = -1$ and one close to $\mu = 1$, the samples of μ_1 may remain close to -1 while the samples of μ_2 remain close to 1
- Is this evidence of convergence? Are we happy with this?

Label switching & MCMC

- No! We know in advance that the marginal posteriors of every θ_h are identical
- Hence, if we observe MCMC chains that do not converge to the same stationary distribution, then we know these chains haven't converged
- Is this a problem if our focus is on estimating the density & not on inferences on component-specific parameters? Seemingly not, as the modes corresponding to permutations of the label indices all correspond to the same posterior on the induced density.
- However, what about if we are interested in mixture-component specific inferences? i.e., we like to know where the different components are located and report this.

Dealing with Label Switching

- It is very common to simply apply standard methods of summarizing the component-specific parameters - e.g., take posterior means & 95% credible intervals for each μ_h - is this a good idea?
- No! This is a very bad idea, because unless weve gotten "lucky" and are stuck in one local mode/configuration of the cluster indices, then posterior summaries are completely meaningless
- In fact, if we had a large number of perfect samples from the true joint posterior, then posterior summaries of μ_h would be identical to those for μ'_h
- One possibility is to relabel the mixture indices after running the MCMC algorithm in a post-processing step (Stephens, 2000; Jasra et al., 2005)

What about putting in order restrictions?

- To deal with label ambiguity, another very common strategy is to put on some identifying restriction to avoid a priori exchangeability
- For example, we could let $\mu_1 < \mu_2 < \ldots < \mu_k$ - any problems with this approach?
- When θ_h has dimension greater than one, it is typically not clear how to define an appropriate constraint
- For example, it may be the case that the means are the same for different components but only the variances differ
- Difficult to implement in general

Approaches to clustering

- There is commonly interest in clustering observations into groups
- Suppose we have $y_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, we may want to group subjects that have similar y values
- There is a very rich literature on clustering via distance-based methods without a likelihood specification
- From a Bayes perspective, "model-based" clustering is more natural (Banfield & Raftery, 93; Fraley & Raftery, 98)

Model based clustering

- Let $y_i \sim \sum_{h=1}^{k} \pi_h \mathcal{K}(y; \theta_h)$, for some parametric kernel $\mathcal{K}$ (typically Gaussian), for $i = 1, \ldots, n$.
- The n subjects allocated to at most k clusters, with each mixture component corresponding to a different cluster
- Suppose we fit the finite mixture model using the EM algorithm to obtain an MLE $\hat{\pi}_h, \hat{\theta}_h, h = 1, \ldots, k$, with k the number of components estimated using BIC
- Conditionally on the estimated parameters, we obtain

$$P(S_i = h \mid y_i, \hat{\pi}, \hat{\theta}) = \frac{\hat{\pi}_h \mathcal{K}(y_i; \hat{\theta}_h)}{\sum_{l=1}^{k} \hat{\pi}_l \mathcal{K}(y_i; \hat{\theta}_l)}$$

with the optimal allocation corresponding to the h that maximizes these probs

Clustering - Comments

- Allocating all the subjects to clusters in this manner, we obtain a partition of $\{1, \ldots, n\}$ into $k_n \leq k$ clusters
- The index on the different clusters is not important - the grouping of the subjects is the focus
- Note that the choice of kernel $\mathcal{K}$ can have a big impact on the estimated number of clusters & the allocation to clusters
- In fact, the definition of a "cluster" is inherently determined entirely by the kernel - if we have a flexible enough kernel, then subjects can always be allocated to a single cluster

Pitfalls & Limitations of Clustering

- From a statistical perspective, new clusters are introduced to accommodate lack of fit in the parametric model $\mathcal{K}(\cdot)$.
- Clearly this is hugely sensitive to $\mathcal{K}$ & it is not clear that clusters obtained from a statistical procedure correspond to scientifically meaningful clusters
- Scientifically, "clusters" are often viewed as corresponding to different mode in a multi-modal distribution, with clusters well defined if these modes are well separated
- Each mixture component does not correspond to a different mode - the relationship between the number of components, the component-specific parameters & the number of modes is complex even for multivariate normal distributions (Ray & Lindsay, 05)

Robust Clustering

- Even focusing on multivariate normals, the clusters can be sensitive to parameterization of the covariance
- Clustering based on normals with diagonal covariance may lead to too many clusters - from the viewpoint of sparsity of modeling & scientific interpretability of the clusters
- Li, Ray & Lindsay (07, JMLR) propose an approach for clustering via mode identification using kernel density estimation & a modal EM algorithm
- Would be interesting to develop a np Bayes version of their approach - e.g., modeling $\mathcal{K}_h$ (the kernel specific to component h) as an unknown unimodal density

How to Estimate Clusters from the MCMC Draws?

- Medvedovic & Sivaganesan (2002) propose to apply standard clustering methods (e.g., hierarchical agglomerative clustering) to a distance matrix obtained using the posterior probabilities of pairwise clustering
- Dahl (2006) proposes a simple approach to obtain a clustering estimate based on the MCMC output using least squares distances from the posterior probability that two subjects are clustered

Modal Clustering

- Note that each MCMC iteration produces one clustering
- One possibility is to estimate the clustering probabilities as the proportion of samples in which that clustering is drawn, and then use the MAP as the optimal clustering under 0-1 loss
- $\#$ possible clusterings in n subjects grows exponentially via Bell number (e.g., > 10275 for $n = 200$)
- Hence, it is very difficult to get accurate estimates of the posterior clustering probabilities & the MAP will have a low posterior probability anyway

Dahl (2006) Cluster Estimation Method

- Dahl (2006) proposed a useful alternative to ad hoc clustering based on the MCMC results & MAP
- Let $\hat{\pi} = \{\hat{\pi}_{ij}\}$ denote the $n \times n$ matrix with elements corresponding to the estimated pairwise posterior probabilities of clustering subjects i and j
- Dahl proposes to choose the least-squares clustering c_{LS}

$$c_{LS} = \text{argmin}_{c \in \{c_1, \ldots, c_B\}} \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_{ij}(c) - \hat{\pi}_{ij})^2$$

  where $\delta_{ij}(c) = 1$ if subjects i and j are in the same cluster under clustering c & 0 otherwise
- We just calculate the least squares distance for each MCMC iteration & choose the best of these iterations

Zhang et al (2014+) Cluster Estimation Method

- Let $\mathcal{F}_B$ denote the space of all membership matrices, as a subset of symmetric $n \times n$ matrices with restrictions: (1) $B(i,j) = \{0,1\}$ for all $i,j = 1, ..., n$; (2) $B(i,\cdot) = B(j,\cdot)$ and $B(\cdot,i) = B(\cdot,j)$ if i-th observation and j-th observation are in the same cluster.
- Obtain posterior samples $\{B^{(i)}, i = 1, \ldots, M\}$
- The final matrix B^* is obtained by calculating the extrinsic mean of the posterior samples defined as follows:
- Find the mode of the number of clusters k_0 based on the samples $B^{(1)}, \ldots, B^{(M)}$.
- Calculate the Euclidean mean and project it onto the membership matrix space:
  1. Euclidean mean: let $\bar{B} = \frac{1}{M} \sum_{t=1}^{M} B^{(t)}$.
  2. Projection: Project the Euclidean mean onto the space of membership matrix by a thresholding operation $B^* = \text{threshold}(\bar{B}, t^*)$ where t^* is the largest threshold such that B^* has k_0 clusters.