C19 : Lecture 4 : A Gibbs Sampler for Gaussian Mixture Models

Frank Wood

University of Oxford

January, 2017

Figures and derivations from [Wood and Black, 2008]

Gaussian Mixture Model

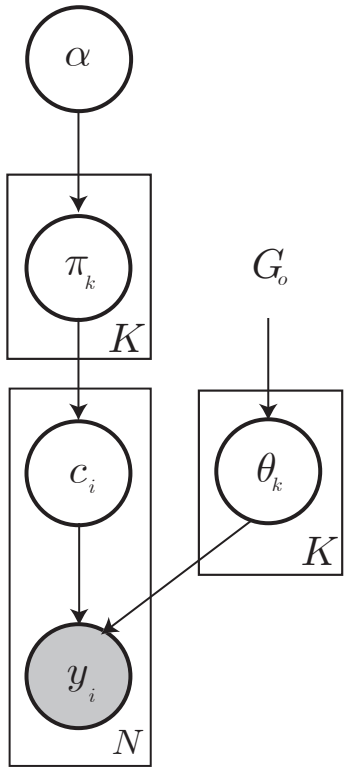

Figure: Bayesian GMM Graphical Model

$$
\begin{aligned}
c_i|\vec{\pi} &\sim \text{Discrete}(\vec{\pi}) \\
\vec{y}_i|c_i = k; \Theta &\sim \text{Gaussian}(\cdot|\theta_k). \\
\vec{\pi}|\alpha &\sim \text{Dirichlet}(\cdot|\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \\
\Theta &\sim \mathcal{G}_0
\end{aligned}
$$

Kinds of questions :

- What's the probability mass in this region?
- Is item i the same as item j?
- How many classes are there (somewhat dangerous).

Gibbs Sampler for GMM I

A Gaussian mixture model is density constructed by mixing Gaussians

$$P(\vec{y}_i) = \sum_{k=1}^{K} P(c_i = k) P(\vec{y}_i | \theta_k)$$

where K is the number of "classes," c_i is a class indicator variable (i.e. $c_i = k$ means that the ith observation came from class k), $P(c_i = k) = \pi_k$ represents the *a priori* probability that the observation i was generated by class k, and the likelihood $P(\vec{y}_i | \theta_k)$ is the generative model of data from class k parameterised by θ_k. The observation model is taken to be a multivariate Gaussian with parameters θ_k for each class k

Gibbs Sampler for GMM II

Notationally, $\mathcal{C} = \{c_i\}_{i=1}^{N}$ is the collection of all class indicator variables, $\Theta = \{\theta_k\}_{k=1}^{K}$, is the collection of all class parameters, $\theta_k = \{\vec{\mu}_k, \mathbf{\Sigma}_k\}$ are the mean and covariance for class k, and $\vec{\pi} = \{\pi_k\}_{k=1}^{K}, \pi_k = P(c_i = k)$ are the class prior probabilities.

Gibbs Sampler for GMM III

To estimate the posterior distribution we first have to specify a prior for all of the parameters of the model.

$$\begin{aligned} \vec{\pi}|\alpha &\sim \text{Dirichlet}(\cdot|\frac{\alpha}{K},\ldots,\frac{\alpha}{K}) \quad (1) \\ \Theta &\sim \mathcal{G}_0 \end{aligned}$$

where $\Theta \sim \mathcal{G}_0$ is shorthand for

$$\begin{aligned} \mathbf{\Sigma}_k &\sim \text{Inverse-Wishart}_{\upsilon_0}(\mathbf{\Lambda}_0^{-1}) \quad (2) \\ \vec{\mu}_k &\sim \text{Gaussian}(\vec{\mu}_0, \mathbf{\Sigma}_k/\kappa_0). \quad (3) \end{aligned}$$

These priors are chosen for mathematical convenience and interpretable expressiveness. They are conjugate priors which will allow us to analytically perform many of the marginalization steps (integrations) necessary to derive a Gibbs sampler for this model.

Gibbs Sampler for GMM IV

The parameters of the Inverse-Wishart prior, $\mathcal{H} = \{\mathbf{\Lambda}_0^{-1}, \upsilon_0, \vec{\mu}_0, \kappa_0\}$, are used to encode our prior beliefs regarding class observation distribution shape and variability. For instance $\vec{\mu}_0$ specifies our prior belief about what the mean of all classes should look like, where κ_0 is the number of pseudo-observations we are willing to ascribe to our belief (in a way similar to that described above for the Dirichlet prior). The hyper-parameters $\mathbf{\Lambda}_0^{-1}$ and υ_0 encode the ways in which individual observations are likely to vary from the mean and how confident we are in our prior beliefs about that.

Gibbs Sampler for GMM V

The joint distribution (from the graphical model) for a Gaussian mixture mode is

$$
\begin{aligned}
& P(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha; \mathcal{H}) \\
&= \left(\prod_{j=1}^{K} P(\theta_j; \mathcal{H}) \right) \left(\prod_{i=1}^{N} P(\vec{y}_i | c_i, \theta_{c_i}) P(c_i | \vec{\pi}) \right) P(\vec{\pi} | \alpha) P(\alpha).
\end{aligned}
$$

Gibbs Sampler for GMM VI

Applying Bayes rule and conditioning on the observed data we see that posterior distribution is simply proportional to the joint (rewritten slightly)

$$
\begin{aligned}
& P(\mathcal{C}, \Theta, \vec{\pi}, \alpha | \mathcal{Y}; \mathcal{H}) \\
& \propto \quad P(\mathcal{Y}|\mathcal{C}, \Theta) P(\Theta; \mathcal{H}) \left(\prod_{i=1}^{N} P(c_i|\vec{\pi}) \right) P(\vec{\pi}\,|\alpha) P(\alpha).
\end{aligned}
$$

where

$$
P(\mathcal{Y}|\mathcal{C}, \Theta) = \prod_{i=1}^{N} P(\vec{y}_i | c_i, \theta_{c_i})
$$

and

$$
P(\Theta; \mathcal{H}) = \prod_{j=1}^{K} P(\theta_j; \mathcal{H}).
$$

Gibbs Sampler for GMM VII

Gibbs sampling, as developed in general by, is possible in this model. Deriving Gibbs sampler for this model requires deriving an expression for the conditional distribution of every latent variable conditioned on all of the others.

To start note that $\vec{\pi}$ can be analytically marginalised out

$$\begin{aligned} P(\mathcal{C}|\alpha) &= \int d\vec{\pi} \prod_{i=1}^{N} P(c_i|\vec{\pi}) P(\vec{\pi}\,|\alpha) \\ &= \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}. \end{aligned} \tag{4}$$

Gibbs Sampler for GMM VIII

Continuing along this line, it is also possible to analytically marginalize out Θ. In the following expression the joint over the data is considered and broken into K parts, each corresponding to one class. We use $\mathcal{H}$ to denote all hyper parameters.

$$
\begin{aligned}
P(\mathcal{C}, \mathcal{Y}; \mathcal{H}) &= \int d\Theta P(\mathcal{C}, \Theta, \mathcal{Y}; \mathcal{H}) \\
&\propto P(\mathcal{C}; \mathcal{H}) \int \cdots \int d\theta_1 \cdots d\theta_K \left(\prod_{j=1}^{K} P(\theta_j; \mathcal{H}) \right) \prod_{i=1}^{N} P(\vec{y}_i | c_i = j, \theta_j) \\
&\propto P(\mathcal{C}; \mathcal{H}) \int \cdots \int d\theta_1 \cdots d\theta_K \prod_{j=1}^{K} \left((\prod_{i=1}^{N} P(\vec{y}_i | c_i = j, \theta_j)^{\mathbb{I}(c_i = j)}) P(\theta_j; \mathcal{H}) \right) \\
&\propto P(\mathcal{C}; \mathcal{H}) \prod_{j=1}^{K} \int d\theta_j (\prod_{i=1}^{N} P(\vec{y}_i | c_i = j, \theta_j)^{\mathbb{I}(c_i = j)}) P(\theta_j; \mathcal{H}) \qquad (5)
\end{aligned}
$$

Gibbs Sampler for GMM IX

Focusing on a single cluster and dropping the class index j for now we see that

$$P(\mathcal{Y}|\mathcal{H}) = \int d\theta \prod_{i=1}^{N} P(\vec{y}_i|\theta) P(\theta; \mathcal{H}) \tag{6}$$

is the standard likelihood conjugate prior integral for a MVN-IW where, remembering that $\theta = \{\vec{\mu}, \mathbf{\Sigma}\}$ the expression for the MVN likelihood term, $P(\vec{y}_i|\theta)$ expands to the familiar MVN normal joint distribution for N i.i.d. observations

$$\prod_{i=1}^{N} P(\vec{y}_i|\theta) = (2\pi)^{-\frac{Nd}{2}} |\mathbf{\Sigma}|^{-\frac{N}{2}} e^{-\frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}_0)} \tag{7}$$

where $\mathbf{S}_0 = \sum_{i=1}^{N} (\vec{y}_i - \vec{\mu})(\vec{y}_i - \vec{\mu})^T$. Here, following convention, $|\mathbf{X}|$ means the matrix determinant of $\mathbf{X}$.

Gibbs Sampler for GMM X

For ease of reference the MVN-IW prior $P(\Theta; \mathcal{H})$ is

$$
\begin{aligned}
P(\theta; \mathcal{H}) &= P(\vec{\mu}, \mathbf{\Sigma} | \vec{\mu}_0, \mathbf{\Lambda}_0, \nu_0, \kappa_0) \\
&= \frac{(\frac{2\pi}{\kappa_0})^{-\frac{d}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} e^{-\frac{\kappa_0}{2}(\vec{\mu}-\vec{\mu}_0)^T \mathbf{\Sigma}^{-1}(\vec{\mu}-\vec{\mu}_0)}}{2^{\frac{\nu_0 d}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma(\frac{\nu_o+1-j}{2})} |\mathbf{\Lambda}_0|^{\frac{\nu_0}{2}} |\mathbf{\Sigma}|^{-\frac{\nu_0+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{\Lambda}_0 \mathbf{\Sigma}^{-1})}
\end{aligned} \tag{8}
$$

Gibbs Sampler for GMM XI

Now, the choice earlier in this section of a conjugate prior for the MVN class parameters helps tremendously. This seemingly daunting integral has a simple analytical solution thanks to conjugacy. Following [Gelman et al., 1995] pg. 87 in making the following variable substitutions

$$\begin{aligned}
\vec{\mu}_n &= \frac{\kappa_0}{\kappa_0 + N}\vec{\mu}_0 + \frac{N}{\kappa_0 + N}\bar{y} \\
\kappa_n &= \kappa_0 + N \\
\nu_n &= \nu_0 + N \\
\mathbf{\Lambda}_n &= \mathbf{\Lambda}_0 + \mathbf{S} + \frac{\kappa_0 n}{\kappa_0 + N}(\bar{y} - \vec{\mu}_0)(\bar{y} - \vec{\mu}_0)^T
\end{aligned}$$

where

$$\mathbf{S} = \sum_{i=1}^{N}(\vec{y}_i - \bar{y})(\vec{y}_i - \bar{y})^T$$

Gibbs Sampler for GMM XII

Now

$$P(\mathcal{Y};\mathcal{H}) = \frac{1}{\mathcal{Z}_0}\int\int d\vec{\mu}d\mathbf{\Sigma}|\mathbf{\Sigma}|^{-(\frac{\nu_n+d}{2}+1)}e^{-\frac{1}{2}\text{tr}(\mathbf{\Lambda}_n\mathbf{\Sigma}^{-1})-\frac{\kappa_n}{2}(\vec{\mu}-\vec{\mu}_n)^T\mathbf{\Sigma}^{-1}(\vec{\mu}-\vec{\mu}_n)}$$

can be solved immediately by realizing that this is itself a MVN-IW distribution and the integral is simply the inverse of its normalization constant $\mathcal{Z}_0$

$$\mathcal{Z}_0 = (2\pi)^{\frac{Nd}{2}}(\frac{2\pi}{\kappa_0})^{\frac{d}{2}}2^{\frac{\nu_0 d}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{j=1}^{d}\Gamma(\frac{\nu_0+1-j}{2})|\mathbf{\Lambda}_0|^{-\frac{\nu_0}{2}}. \quad (9)$$

with the same variable substitutions applied, i.e.,

Gibbs Sampler for GMM XIII

$$\begin{aligned}\mathcal{Z}_n &= \int\int d\vec{\mu} d\mathbf{\Sigma} |\mathbf{\Sigma}|^{-(\frac{\nu_n+d}{2}+1)} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{\Lambda}_n\mathbf{\Sigma}^{-1})-\frac{\kappa_n}{2}(\vec{\mu}-\vec{\mu}_n)^T\mathbf{\Sigma}^{-1}(\vec{\mu}-\vec{\mu}_n)} \\ &= (2\pi)^{\frac{Nd}{2}}(\frac{2\pi}{\kappa_n})^{\frac{d}{2}} 2^{\frac{\nu_n d}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma(\frac{\nu_n+1-j}{2}) |\mathbf{\Lambda}_n|^{-\frac{\nu_n}{2}}\end{aligned}$$

which yields

$$P(\mathcal{Y};\mathcal{H}) = \frac{\mathcal{Z}_n}{\mathcal{Z}_0} \tag{10}$$

$$= (\frac{\kappa_0}{\kappa_n})^{\frac{d}{2}} 2^{\frac{d}{2}(\nu_n-\nu_0)} \frac{|\mathbf{\Lambda}_0|^{\frac{\nu_0}{2}} \prod_{j=1}^{d} \Gamma(\frac{\nu_n+1-j}{2})}{|\mathbf{\Lambda}_n|^{\frac{\nu_n}{2}} \prod_{j=1}^{d} \Gamma(\frac{\nu_0+1-j}{2})} \tag{11}$$

Gibbs Sampler for GMM XIV

Remembering that our derivation of $P(\mathcal{Y};\mathcal{H})$ was for a *single class*, we now have an analytic expression for

$$P(\mathcal{C},\mathcal{Y};\mathcal{H}) = \prod_{j=1}^{K} P(\mathcal{Y}^{(j)}|\mathcal{C};\mathcal{H})P(\mathcal{C}|\mathcal{H})$$

From which we can MH sample by modifying each c_i and recomputing the joint.

Gibbs Sampler for GMM XV

In this model we can do better by deriving a Gibbs update for each class indicator variable c_i.

$$\begin{aligned} P(c_i = j|\mathcal{C}_{-i}, \mathcal{Y}, \alpha; \mathcal{H}) &\propto P(\mathcal{Y}|\mathcal{C}; \mathcal{H})P(\mathcal{C}|\alpha) \\ &\propto \prod_{j=1}^{K} P(\mathcal{Y}^{(j)}; \mathcal{H})P(c_i = j|\mathcal{C}_{-i}, \alpha) \\ &\propto P(\mathcal{Y}^{(j)}; \mathcal{H})P(c_i = j|\mathcal{C}_{-i}, \alpha) \\ &\propto P(y_i|\mathcal{Y}^{(j)} \backslash y_i; \mathcal{H})P(c_i = j|\mathcal{C}_{-i}, \alpha) \quad (12) \end{aligned}$$

where $\mathcal{Y}^{(j)} \backslash y_i$ is the set of observations currently assigned to class j except y_i (y_i is "removed" from the class to which it belongs when sampling).

Gibbs Sampler for GMM XVI

Because of our choice of conjugate prior we know that we know that $P(y_i|\mathcal{Y}^{(j)}\backslash y_i; \mathcal{H})$ is multivariate Student-t (Gelman et al. [1995] pg. 88)

$$y_i|\mathcal{Y}^{(j)}\backslash y_i; \mathcal{H} \sim t_{\nu_n - D+1}(\vec{\mu}_n, \mathbf{\Lambda}_n(\kappa_n + 1)/(\kappa_n(\nu_n - D + 1))) \tag{13}$$

where

$$\begin{aligned}
\vec{\mu}_n &= \frac{\kappa_0}{\kappa_0 + N}\vec{\mu}_0 + \frac{N}{\kappa_0 + N}\bar{y} \\
\kappa_n &= \kappa_0 + N \\
\nu_n &= \nu_0 + N \\
\mathbf{\Lambda}_n &= \mathbf{\Lambda}_0 + \mathbf{S} + \frac{\kappa_0 n}{\kappa_0 + N}(\bar{y} - \vec{\mu}_0)(\bar{y} - \vec{\mu}_0)^T
\end{aligned}$$

and D is the dimensionality of y_i. Note that $N, \bar{y}, \vec{\mu}_n, \kappa_n, \nu_n, \mathbf{\Lambda}_n$ must all be computing *excluding* y_i.

Gibbs Sampler for GMM XVII

Deriving $P(c_i = j|\mathcal{C}_{-i}, \alpha)$ was covered in the first lecture and involves simplifying a ratio of two Dirichlet normalising constants.

Its simplified form is

$$P(c_i = j|\mathcal{C}_{-i}, \alpha) = \frac{m_j + \frac{\alpha}{K}}{N - 1 + \alpha}$$

Gibbs Sampler for GMM XVIII

Given

$$\begin{aligned} & P(c_i = j | \mathcal{C}_{-i}, \mathcal{Y}, \alpha; \mathcal{H}) \\ \propto \quad & P(y_i | \mathcal{Y}^{(j)} \backslash y_i; \mathcal{H}) P(c_i = j | \mathcal{C}_{-i}, \alpha) \\ = \quad & P(y_i | \mathcal{Y}^{(j)} \backslash y_i; \mathcal{H}) \left(\frac{m_j + \frac{\alpha}{K}}{N - 1 + \alpha} \right) \end{aligned}$$

we can enumerate all K values c_i could take and normalise then sample from this discrete distribution.

Study Suggestions

- Implement this sampler
- Write the following question as an integral : what's the probability of a datapoint falling in a particular region of space?
- Given the Gibbs sampler described, how would you answer this question efficiently?
- Derive and implement a Gibbs sampler for LDA.
- Derive and implement a sampler for PPCA.
- How would you answer the question, how many classes are there?
- Is it safe, in this model, to ask questions about the characteristics of class i?

Bibliography I

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall, New York, 1995.

F. Wood and M. J. Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, page to appear, 2008.