Two-dimensional Dirichlet Process Mixture Model

Brian M. Brost

10 November 2015

Description

A Dirichlet process mixture model for cluster estimation of 2-dimensional, normally distributed data.

Implementation

The file dp.mixture.2d.sim.R simulates data according to the model statement presented below, and dp.mixture.2d.mcmc.R contains the MCMC algorithm for model fitting. Model implementation follows the blocked Gibbs sampler approach of Ishwaran and James (2001) and Gelman et al. (2014).

Model statement

Let $\mathbf{s}_t = (s_{1,t}, s_{2,t})'$, for $t = 1, \ldots, T$, be observations and $\boldsymbol{\mu}_h = (\mu_{1,h}, \mu_{2,h})'$, for $h = 1, \ldots, H$, be the locations of clusters. The parameter H denotes the maximum number of clusters allowed unde the truncation approximation of Dirichlet process mixture (Gelman et al. 2014). Define $\tilde{S}$ to be the uniform support of the Dirichlet process (i.e., all possible $\boldsymbol{\mu}_h$).

$$
\begin{aligned}
\mathbf{s}_t &\sim \mathcal{N}(\boldsymbol{\mu}_h, \sigma^2\mathbf{I}) \\
h_t &\sim \text{Cat}\left(\pi_1, \ldots, \pi_H\right) \\
\pi_h &\sim \text{Stick}(\theta) \\
\boldsymbol{\mu}_h &\sim \text{Unif}(\tilde{\mathcal{S}}) \\
\theta &\sim \text{Gamma}(r, q) \\
\sigma &\sim \text{Unif}(l, u)
\end{aligned}
$$

The concentration parameter θ affects the clustering in the Dirichlet process mixture: smaller values yield fewer clusters with more observations per cluster, whereas larger values yield more clusters with fewer observations per cluster. Note that the lines in this model statement pertaining to h_t, π_h, and $\boldsymbol{\mu}_h$ comprise the stick-breaking representation of the Dirichlet process mixture model, i.e.,

$$
\begin{aligned}
\boldsymbol{\mu}_h &\sim \mathbf{P} \\
\mathbf{P} &\sim \text{DP}(\theta, \mathbf{P}_0) \\
\mathbf{P}_0 &\sim \text{Unif}(\tilde{\mathcal{S}})
\end{aligned}
$$

Full conditional distributions

Cluster locations ($\boldsymbol{\mu}_h$):

$$
\begin{aligned}
[\boldsymbol{\mu}_h|\cdot] &\propto \prod_{t=1}^{T} [\mathbf{s}_t|\boldsymbol{\mu}_{h_t}, \sigma^2]^{1_{\{h_t=h\}}} [\boldsymbol{\mu}_h \mid \tilde{\mathcal{S}}] \\
&\propto \prod_{\{t:h_t=h\}} \mathcal{N}(\mathbf{s}_t|\boldsymbol{\mu}_{h_t}, \sigma^2) 1_{\{\boldsymbol{\mu}_h \in \tilde{\mathcal{S}}\}} \\
&\propto \prod_{\{t:h_t=h\}} \exp\left\{-\frac{1}{2}\left((\mathbf{s}_t - \boldsymbol{\mu}_{h_t})' \left(\sigma^2\mathbf{I}\right)^{-1} (\mathbf{s}_t - \boldsymbol{\mu}_{h_t})\right)\right\} 1_{\{\boldsymbol{\mu}_h \in \tilde{\mathcal{S}}\}}
\end{aligned}
$$

$$
\begin{aligned}
&\propto \prod_{\{t:h_t=h\}} \exp\left\{-\frac{1}{2}\left(\mathbf{s}_t'\left(\sigma^2\mathbf{I}\right)^{-1}\mathbf{s}_t - 2\mathbf{s}_t'\left(\sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{h_t} + \boldsymbol{\mu}_{h_t}'\left(\sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{h_t}\right)\right\} 1_{\{\boldsymbol{\mu}_h\in\tilde{\mathcal{S}}\}} \\
&\propto \prod_{\{t:h_t=h\}} \exp\left\{-\frac{1}{2}\left(-2\mathbf{s}_t'\left(\sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{h_t} + \boldsymbol{\mu}_{h_t}'\left(\sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{h_t}\right)\right\} 1_{\{\boldsymbol{\mu}_h\in\tilde{\mathcal{S}}\}} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\sum_{\{t:h_t=h\}}\mathbf{s}_t'\left(\sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{h_t} + \boldsymbol{\mu}_{h_t}'\left(n_h\left(\sigma^2\mathbf{I}\right)^{-1}\right)\boldsymbol{\mu}_{h_t}\right)\right\} 1_{\{\boldsymbol{\mu}_h\in\tilde{\mathcal{S}}\}} \\
&= \mathcal{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}) 1_{\{\boldsymbol{\mu}_h\in\tilde{\mathcal{S}}\}}
\end{aligned}
$$

where $\mathbf{A} = n_h\left(\sigma^2\mathbf{I}\right)^{-1}$ and $\mathbf{b} = \sum_{\{t:h_t=h\}}\mathbf{s}_t'\left(\sigma^2\mathbf{I}\right)^{-1}$; therefore, $[\boldsymbol{\mu}_h|\cdot] = \mathcal{N}\left(\frac{1}{n_h}\sum_{\{t:h_t=h\}}\mathbf{s}_t, \frac{\sigma^2}{n_h}\mathbf{I}\right)$. Note that the product (or summation) is over all $\mathbf{s}_t$ that belong to cluster h (h_t is a latent class status that indicates membership of observation $\mathbf{s}_t$ to cluster h), and n_h is the number of observations allocated to h. Also note that 'proposed' values for $\boldsymbol{\mu}_h$ not in $\tilde{\mathcal{S}}$ are rejected, i.e., $[\boldsymbol{\mu}_h|\cdot] = \mathcal{TN}(\mathbf{s}_t, \frac{\sigma^2}{n_h}\mathrm{I})_{\tilde{\mathcal{S}}}$.

Probability mass for cluster location $\boldsymbol{\mu}_h$ (π_h):

The stick-breaking representation of a Dirichlet process mixture consists of two components, namely a cluster weight and a cluster probability. Let η_h denote the weight assigned to cluster h, where $\eta_h \sim \mathrm{Beta}(1,\theta)$. The associated full-conditional is

$$
[\eta_h|\cdot] \quad \sim \quad \mathrm{Beta}\left(1 + n_h, \theta + \sum_{\tilde{h}=h+1}^{H} n_{\tilde{h}}\right), \text{ for } h = 1, \ldots, H-1,
$$

and $\eta_H = 1$. The parameter n_h denotes the number of observations allocated to cluster h. Note that η_h is sampled in order of decreasing n_h, i.e., n_h is sorted largest to smallest and η_h is sampled in sequence. The cluster probabilities (π_h) are deterministic and calculated as

$$
\pi_h = \eta_h \prod_{\tilde{h}<h}(1-\eta_{\tilde{h}}).
$$

The probabilities π_h are also calculated in order of decreasing n_h. The derivation of this full-conditional is as follows:

$$
\begin{aligned}
[\eta_h|\cdot] &\propto \prod_{t=1}^{T}[h_t \mid \pi_h]^{1_{\{h_t=h\}}} \prod_{\tilde{h}=h+1}^{H}\prod_{t=1}^{T}\left[h_t \mid \pi_{\tilde{h}}\right]^{1_{\{h_t=\tilde{h}\}}} [\eta_h|\theta_1,\theta_2] \\
&\propto \prod_{t=1}^{T}\pi_h^{1_{\{h_t=h\}}} \prod_{\tilde{h}=h+1}^{H}\prod_{t=1}^{T}\pi_{\tilde{h}}^{1_{\{h_t=\tilde{h}\}}} [\eta_h|\theta_1,\theta_2] \\
&\propto \pi_h^{\sum_t 1_{\{h_t=h\}}} \prod_{\tilde{h}=h+1}^{H}\pi_{\tilde{h}}^{\sum_t 1_{\{h_t=\tilde{h}\}}} [\eta_h|\theta_1,\theta_2] \\
&\propto \left(\eta_h\prod_{l<h}(1-\eta_l)\right)^{\sum_t 1_{\{h_t=h\}}} \prod_{\tilde{h}=h+1}^{H}\left(\eta_{\tilde{h}}\prod_{l<\tilde{h}}(1-\eta_l)\right)^{\sum_t 1_{\{h_t=\tilde{h}\}}} \eta_h^{\theta_1-1}(1-\eta_h)^{\theta_2-1} \\
&\propto \eta_h^{\sum_t 1_{\{h_t=h\}}} \prod_{\tilde{h}=h+1}^{H}\left(\prod_{l<\tilde{h}}(1-\eta_l)\right)^{\sum_t 1_{\{h_t=\tilde{h}\}}} \eta_h^{\theta_1-1}(1-\eta_h)^{\theta_2-1} \\
&\propto \eta_h^{\sum_t 1_{\{h_t=h\}}}(1-\eta_h)^{\sum_{\tilde{h}=h+1}^{H}\sum_t 1_{\{h_t=\tilde{h}\}}}\eta_h^{\theta_1-1}(1-\eta_h)^{\theta_2-1} \\
&\propto \eta_h^{\sum_t 1_{\{h_t=h\}}+\theta_1-1}(1-\eta_h)^{\sum_{\tilde{h}=h+1}^{H}\sum_t 1_{\{h_t=\tilde{h}\}}+\theta_2-1}
\end{aligned}
$$

$$= \quad \text{Beta}\left(\sum_t 1_{\{h_t=h\}} + \theta_1, \sum_{\tilde{h}=h+1}^{H} \sum_t 1_{\{h_t=\tilde{h}\}} + \theta_2\right)$$

See page 553 in Gelman et al. (2014) and Section 5.2 in Ishwaran and James (2001).

Dirichlet process concentration parameter (θ):

$$[\theta|\cdot] \quad \propto \quad \text{Gamma}(r + H - 1, q - \sum_{h=1}^{H-1} \log(1-\eta_h)).$$

See page 553 in Gelman et al. (2014). Also see Escobar and West (1995) and West (1997?, white paper) for alternative full-conditionals for θ. Also see Ishwaran and Zarepour (2000) for derivation.

Latent cluster classification variable (h_t):

$$\begin{aligned}
[h_t|\cdot] \quad &\sim \quad \left[\mathbf{s}_t \mid \boldsymbol{\mu}_{h_t}, \sigma^2\right][h_t \mid \pi_h] \\
&\sim \quad \text{Cat}\left(\frac{\pi_h\left[\mathbf{s}_t \mid \boldsymbol{\mu}_{h_t}, \sigma^2\right]}{\sum_{\tilde{h}=1}^{H} \pi_{\tilde{h}}\left[\mathbf{s}_t \mid \boldsymbol{\mu}_{\tilde{h}}, \sigma^2\right]}\right) \\
&\sim \quad \text{Cat}\left(\frac{\pi_h\left(\mathcal{N}\left(\mathbf{s}_t \mid \boldsymbol{\mu}_{h_t}, \sigma^2\mathbf{I}\right)\right)}{\sum_{\tilde{h}=1}^{H} \pi_{\tilde{h}}\left(\mathcal{N}\left(\mathbf{s}_t \mid \boldsymbol{\mu}_{\tilde{h}}, \sigma^2\mathbf{I}\right)\right)}\right)
\end{aligned}$$

This update proceeds just as in multinomial sampling; see page 552 in Gelman et al. (2014).

Error in the observation process (σ):

$$[\sigma|\cdot] \quad \propto \quad \prod_{t=1}^{T}[\mathbf{s}_t|\boldsymbol{\mu}_h, \sigma^2][\sigma]$$

The update for σ proceeds using Metropolis-Hastings.

References

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2014. Bayesian data analysis. CRC Press.

Ishwaran, H., and L.F. James. 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96: 161–173.