STA 610L: Module 4.11

Introduction to finite mixture models (continuous data)

Dr. Olanrewaju Michael Akande

Continuous data (univariate case)

- Suppose we have univariate continuous data $y_i \overset{iid}{\sim} f$, for $i, \ldots, n$, where f is an unknown density.
- Turns out that we can approximate "almost" any f with a mixture of normals. Usual choices are

1. Location mixture (multimodal):

$$f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu_k, \sigma^2\right)$$

2. Scale mixture (unimodal and symmetric about the mean, but fatter tails than a regular normal distribution):

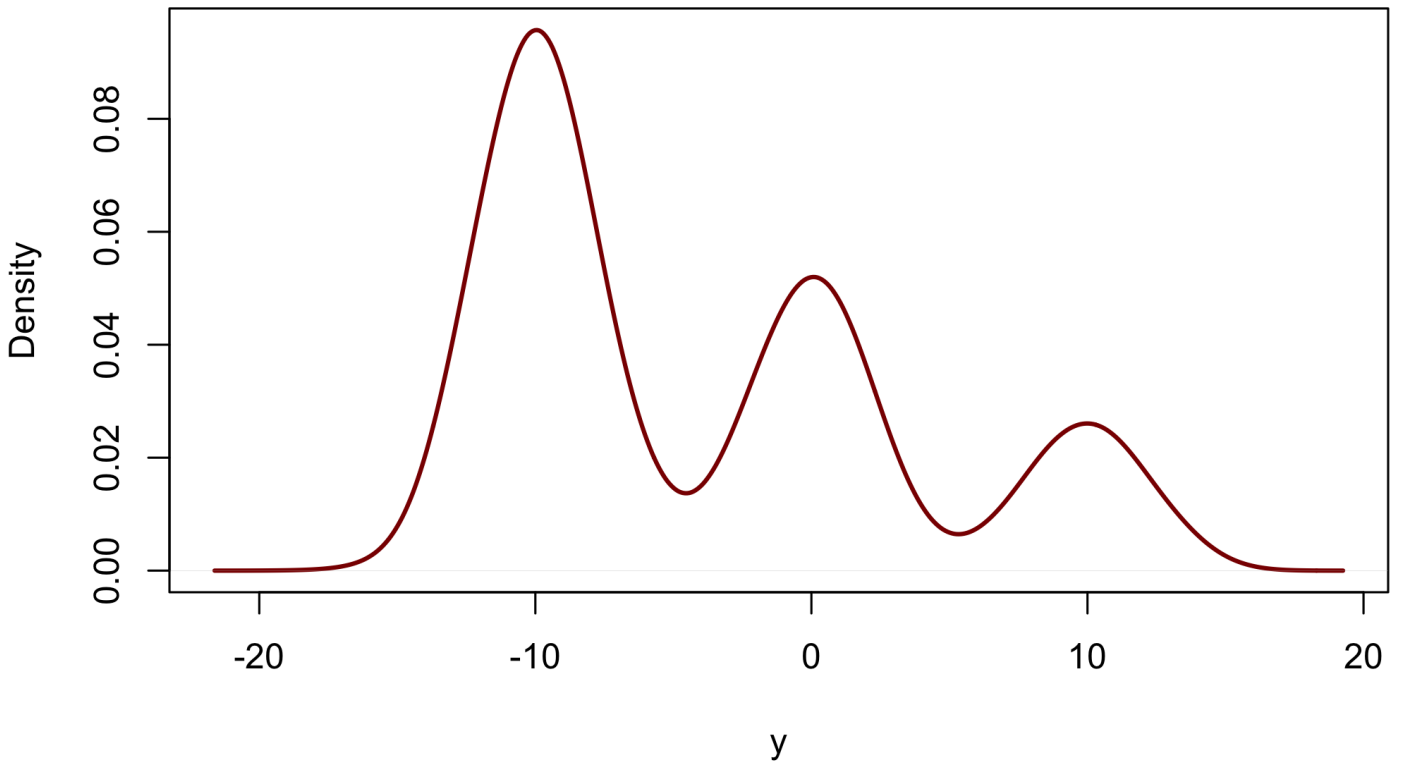
$$f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu, \sigma_k^2\right)$$

3. Location-scale mixture (multimodal with potentially fat tails):

$$f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

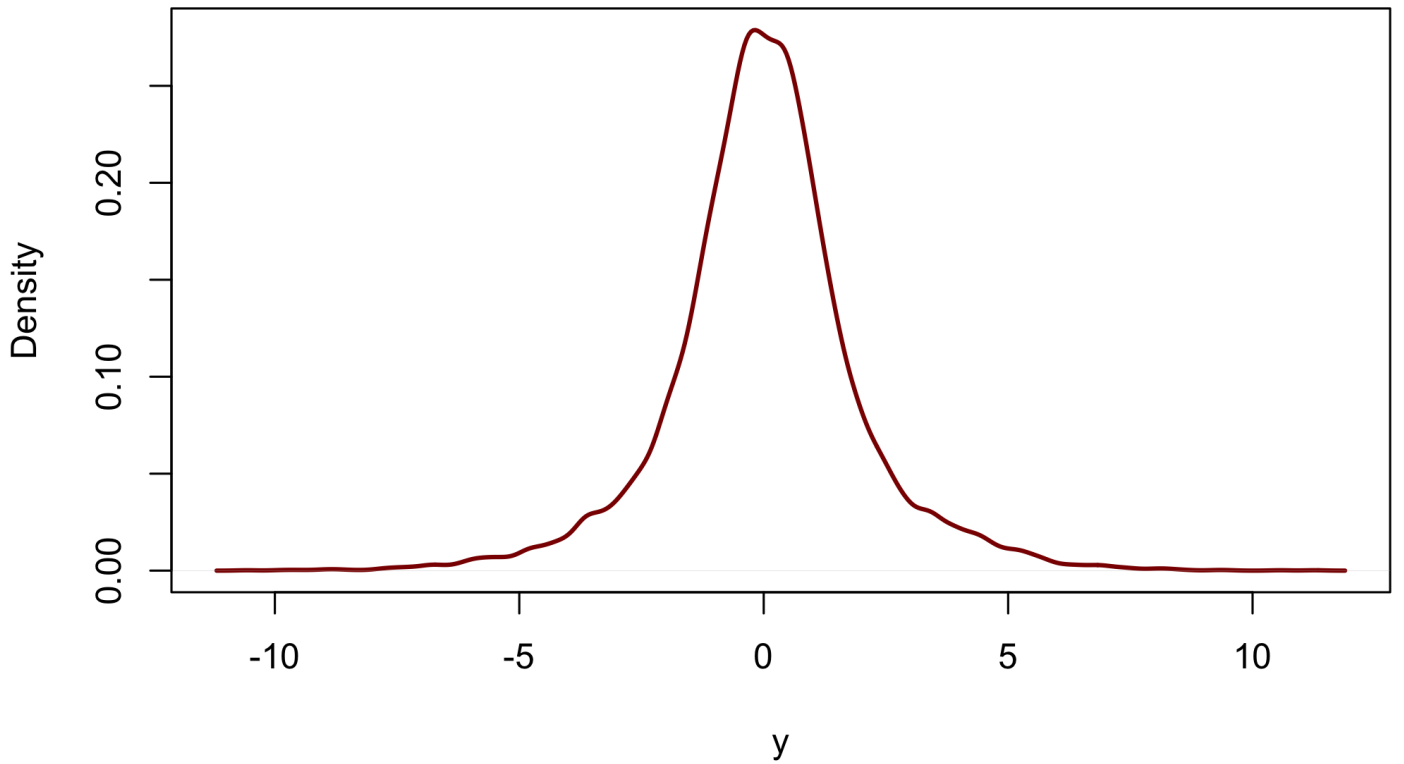
Location mixture example

$$f(y) = 0.55\mathcal{N}(-10, 4) + 0.30\mathcal{N}(0, 4) + 0.15\mathcal{N}(10, 4)$$



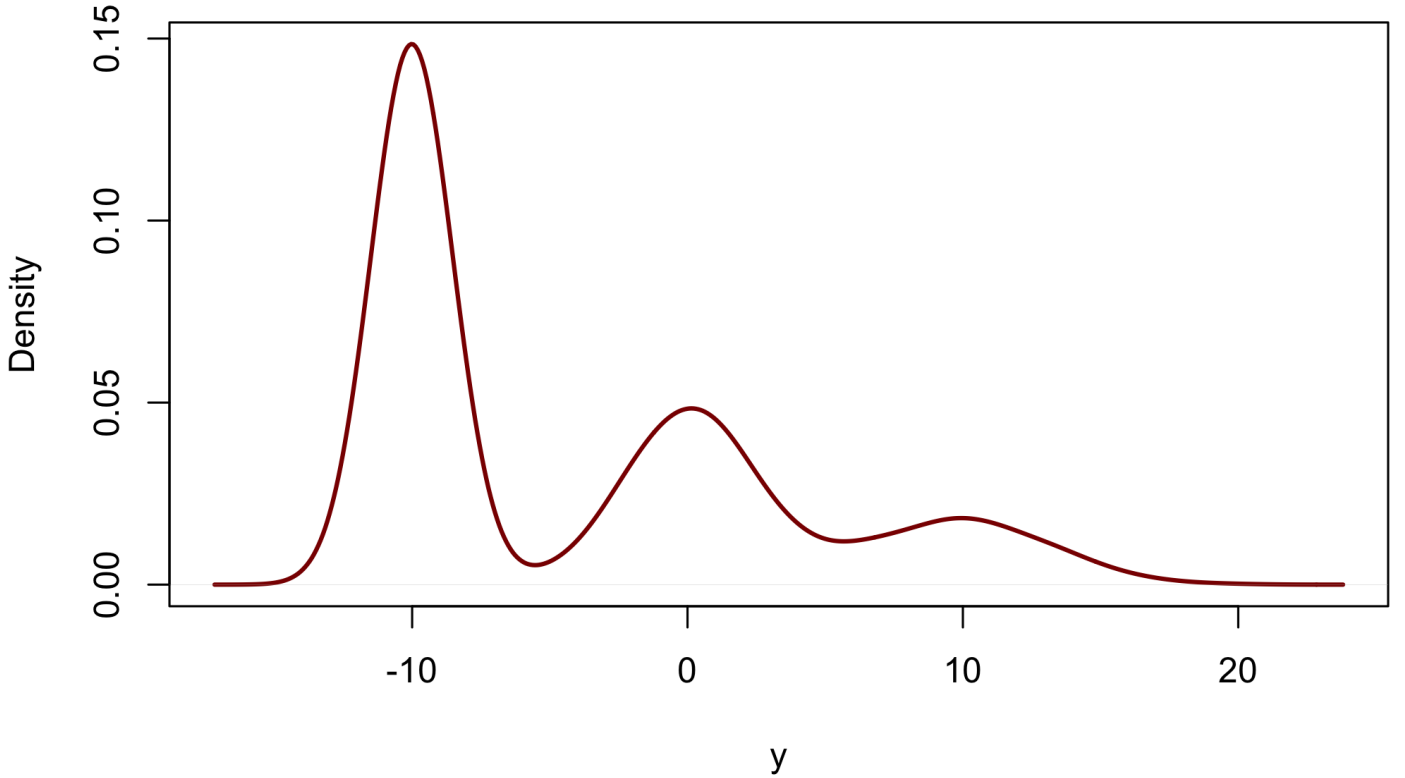
Scale mixture example

$$f(y) = 0.55\mathcal{N}(0,1) + 0.30\mathcal{N}(0,5) + 0.15\mathcal{N}(0,10)$$



Location-scale mixture example

$$f(y) = 0.55\mathcal{N}\left(-10, 1\right) + 0.30\mathcal{N}\left(0, 5\right) + 0.15\mathcal{N}\left(10, 10\right)$$



Location mixture of normals

- Consider the location mixture $f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu_k, \sigma^2\right)$. How can we do inference?
- Right now, we only have three unknowns: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and σ^2.
- For priors, the most obvious choices are
  - $\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$,
  - $\mu_k \sim \mathcal{N}(\mu_0, \gamma_0^2)$, for each $k = 1, \ldots, K$, and
  - $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$.
- However, we do not want to use the likelihood with the sum in the mixture. We prefer products!

Data augmentation

- This once again brings us to the concept of data augmentation, which we can now discuss in a bit more detail.
- Data augmentation is a commonly-used technique for designing MCMC samplers using auxiliary/latent/hidden variables. Again, we have already seen this.
- **Idea**: introduce variable Z that depends on the distribution of the existing variables in such a way that the resulting conditional distributions, with Z included, are easier to sample from and/or result in better mixing.
- Z's are just latent/hidden variables that are introduced for the purpose of simplifying/improving the sampler.

Data augmentation

- For example, suppose we want to sample from $p(x, y)$, but $p(x|y)$ and/or $p(y|x)$ are complicated.
- Choose $p(z|x, y)$ such that $p(x|y, z)$, $p(y|x, z)$, and $p(z|x, y)$ are easy to sample from. Note that we have $p(x, y, z) = p(z|x, y)p(x, y)$.
- Alternatively, rewrite the model as $p(x, y|z)$ and specify $p(z)$ such that

$$p(x, y) = \int p(x, y|z)p(z)\mathrm{d}z,$$

  where the resulting $p(x|y, z)$, $p(y|x, z)$, and $p(z|x, y)$ from the joint $p(x, y, z)$ are again easy to sample from.
- Next, construct a Gibbs sampler to sample all three variables (X, Y, Z) from $p(x, y, z)$.
- Finally, throw away the sampled Z's and from what we know about Gibbs sampling, the samples (X, Y) are from the desired $p(x, y)$.

Location mixture of normals

- Back to location mixture $f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu_k, \sigma^2\right)$.
- Introduce latent variable $z_i \in \{1, \ldots, K\}$.
- Then, we have
  - $y_i | z_i \sim \mathcal{N}\left(\mu_{z_i}, \sigma^2\right)$, and
  - $\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}$.
- How does that help? Well, the observed data likelihood is now

$$p\left[Y = (y_1, \ldots, y_n) | Z = (z_1, \ldots, z_n), \boldsymbol{\lambda}, \boldsymbol{\mu}, \sigma^2\right] = \prod_{i=1}^{n} p\left(y_i | z_i, \mu_{z_i}, \sigma^2\right)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{z_i})^2\right\}$$

which is much easier to work with.

Posterior inference

- The joint posterior is

$$
\begin{aligned}
\pi\left(Z, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\lambda} | Y\right) &\propto \left[\prod_{i=1}^{n} p\left(y_i | z_i, \mu_{z_i}, \sigma^2\right)\right] \cdot \Pr(Z | \boldsymbol{\mu}, \sigma^2, \boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\lambda}) \\
&\propto \left[\prod_{i=1}^{n} p\left(y_i | z_i, \mu_{z_i}, \sigma^2\right)\right] \cdot \Pr(Z | \boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\mu}) \cdot \pi(\sigma^2) \\
&\propto \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{z_i})^2\right\}\right] \\
&\quad \times \left[\prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}\right] \\
&\quad \times \left[\prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}\right] \cdot \\
&\quad \times \left[\prod_{k=1}^{K} \mathcal{N}(\mu_k; \mu_0, \gamma_0^2)\right] \\
&\quad \times \left[\mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)\right].
\end{aligned}
$$

Full conditionals

- For $i = 1, \ldots, n$, sample $z_i \in \{1, \ldots, K\}$ from a categorical distribution (multinomial distribution with sample size one) with probabilities

$$\begin{aligned}
\Pr[z_i = k | \ldots] &= \frac{\Pr[y_i, z_i = k | \mu_k, \sigma^2, \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i, z_i = l | \mu_l, \sigma^2, \lambda_l]} \\
&= \frac{\Pr[y_i | z_i = k, \mu_k, \sigma^2] \cdot \Pr[z_i = k | \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i | z_i = l, \mu_l, \sigma^2] \cdot \Pr[z_i = l | \lambda_l]} \\
&= \frac{\lambda_k \cdot \mathcal{N}\left(y_i; \mu_k, \sigma^2\right)}{\sum_{l=1}^{K} \lambda_l \cdot \mathcal{N}\left(y_i; \mu_l, \sigma^2\right)}.
\end{aligned}$$

- Note that $\mathcal{N}\left(y_i; \mu_k, \sigma^2\right)$ just means evaluating the density $\mathcal{N}\left(\mu_k, \sigma^2\right)$ at the value y_i.

Full conditionals

- Next, sample $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ from

$$\pi[\boldsymbol{\lambda}|\ldots] \equiv \text{Dirichlet}\left(\alpha_1 + n_1, \ldots, \alpha_K + n_K\right),$$

where $n_k = \sum_{i=1}^{n} 1[z_i = k]$, the number of individuals assigned to cluster k.

- Sample the mean μ_k for each cluster from

$$\pi[\mu_k|\ldots] \equiv \mathcal{N}(\mu_{k,n}, \gamma^2_{k,n});$$

$$\gamma^2_{k,n} = \frac{1}{\frac{n_k}{\sigma^2} + \frac{1}{\gamma_0^2}}; \qquad \mu_{k,n} = \gamma^2_{k,n}\left[\frac{n_k}{\sigma^2}\bar{y}_k + \frac{1}{\gamma_0^2}\mu_0\right],$$

- Finally, sample σ^2 from

$$\pi(\sigma^2|\ldots) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right).$$

$$\nu_n = \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu_{z_i})^2\right].$$

Practical considerations

- The standard Gibbs sampler for this model can suffer from label switching.
- For example, suppose our groups are men and women. Then, if we run the sampler multiple times (starting from the same initial values), sometimes it will settle on females as the first group, and sometimes on females are the second group.
- Specifically, MCMC on mixture models in general can suffer from label switching.
- Fortunately, results are still valid if we interpret them correctly.
- Specifically, we should focus on quantities and estimands that are invariant to permutations of the clusters. For example, look at marginal quantities, instead of conditional ones.

Other practical considerations

- So far we have assumed that the number of clusters K is known.
- What if we don't know K?
  - Compare marginal likelihood for different choices of K and select K with best performance.
  - Can also use other metrics, such as MSE, and so on.
  - Maybe a prior on K?
  - Go Bayesian non-parametric: Dirichlet processes!

See the R script here for sample code.

Finite mixture of multivariate normals

- It is relatively easy to extend this to the multivariate case.
- As with the univariate case, given a sufficiently large number of mixture components, a scale-location multivariate normal mixture model can be used to approximate any multivariate density.
- We have

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \lambda_k \cdot \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k)$$

- Or equivalently,

$$\mathbf{y}_i | z_i, \boldsymbol{\mu}_{z_i}, \Sigma_{z_i} \sim \mathcal{N}_p(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i})$$

$$\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}$$

Posterior inference

- We can then specify priors as

$$\pi(\boldsymbol{\mu}_k) = \mathcal{N}_p\left(\boldsymbol{\mu}_0, \Lambda_0\right) \quad \text{for } k = 1, \ldots, K;$$

$$\pi(\Sigma_k) = \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for } k = 1, \ldots, K;$$

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(a_1, \ldots, a_K).$$

- We can also just use the conjugate option for $\pi(\boldsymbol{\mu}_k, \Sigma_k)$ to avoid specifying Λ_0, so that we have

$$\begin{aligned}\pi(\boldsymbol{\mu}_k, \Sigma_k) &= \pi(\boldsymbol{\mu}_k | \Sigma_k) \cdot \pi(\Sigma_k) \\ &= \mathcal{N}_p\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k\right) \cdot \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for } k = 1, \ldots, K;\end{aligned}$$

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(a_1, \ldots, a_K).$$

- Gibbs sampler for both options follow directly from STA 360/601/602 and what we have covered so far.

Label switching again

- To avoid label switching when fitting the model, we can constrain the order of the $\boldsymbol{\mu}_k$'s.
- Here are three of many approaches:

1. Constrain the prior on the $\boldsymbol{\mu}_k$'s to be

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k) \quad \boldsymbol{\mu}_{k-1} < \boldsymbol{\mu}_k < \boldsymbol{\mu}_{k+1},$$

   which does not always seem reasonable.

2. Relax option 1 above to only the first component of the mean vectors

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma_k) \quad \mu_{1,k-1} < \mu_{1,k} < \mu_{1,k+1}.$$

3. Try an ad-hoc fix. After sampling the $\boldsymbol{\mu}_k$'s, rearrange the labels to satisfy $\mu_{1,k-1} < \mu_{1,k} < \mu_{1,k+1}$ and reassign the labels on $\boldsymbol{\Sigma}_k$ accordingly.

DP MIXTURE OF NORMALS (TEASER)

- To avoid setting K apriori, we can extend this finite mixture of normals to a Dirichlet process (DP) mixture of normals.
- The first level of the model remains the same. That is,

$$\mathbf{y}_i | z_i, \boldsymbol{\mu}_{z_i}, \Sigma_{z_i} \sim \mathcal{N}_p(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) \quad \text{for each } i;$$

$$\pi(\boldsymbol{\mu}_k, \Sigma_k) = \pi(\boldsymbol{\mu}_k | \Sigma_k) \cdot \pi(\Sigma_k)$$

$$= \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{\kappa_0}\Sigma_k\right) \cdot \mathcal{IW}_p\left(\nu_0, S_0\right) \quad \text{for each } k.$$

DP MIXTURE OF NORMALS (TEASER)

- For the prior on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, use the following stick breaking representation of the Dirichlet process.

$$
\begin{aligned}
P(z_i = k) &= \lambda_k; \\
\lambda_k &= V_k \prod_{l<k} (1 - V_l) \;\text{ for }\; k = 1, \ldots, \infty; \\
V_k &\overset{iid}{\sim} \text{Beta}(1, \alpha); \\
\alpha &\sim \text{Gamma}(a, b).
\end{aligned}
$$

- As an approximation, use $\lambda_k = V_k \prod_{l<k} (1 - V_l) \;\text{ for }\; k = 1, \ldots, K^\star$ with $K^\star$ set to be as large as possible!
- This specification forces the model to only use as many components as needed, and usually, no more. Also, the Gibbs sampler is relatively straightforward.
- Other details are beyond the scope of this course, but I am happy to provide resources for those interested!

What's next?

Move on to the readings for the next module!