Finite mixture models cont'd; course wrap-up and brief review session

Dr. Olanrewaju Michael Akande

April 15, 2020

Announcements

- Final class today!!!
- Final reminder: let the instructor know if you plan to request a letter grade.
- Complete course evaluations.

Outline

- Finite mixture models
  - Continuous data – univariate case
  - Illustration
  - Categorical data – bivariate case

Continuous data

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):

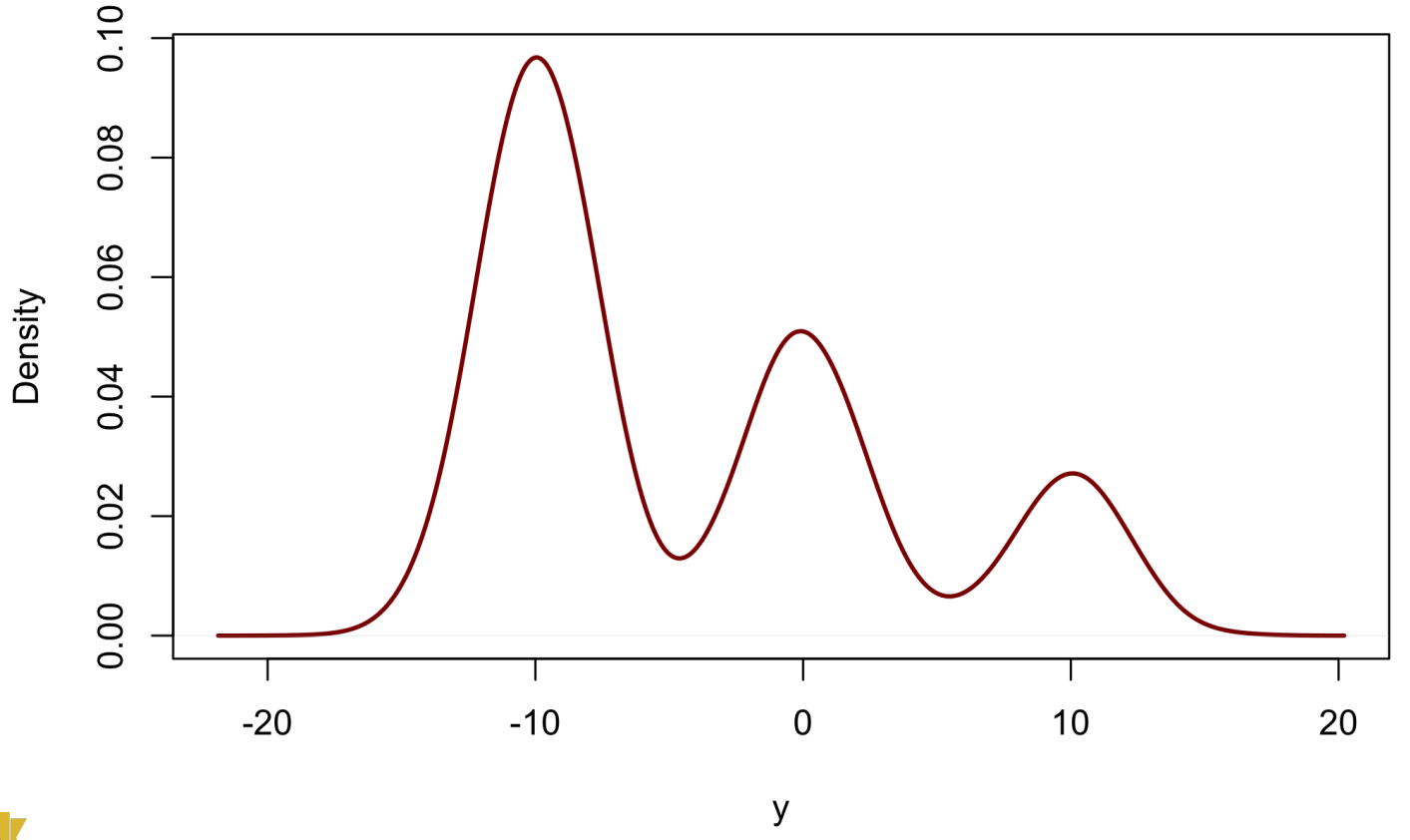
$$f(y) = \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\mu, \sigma_k^2\right)$$

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

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

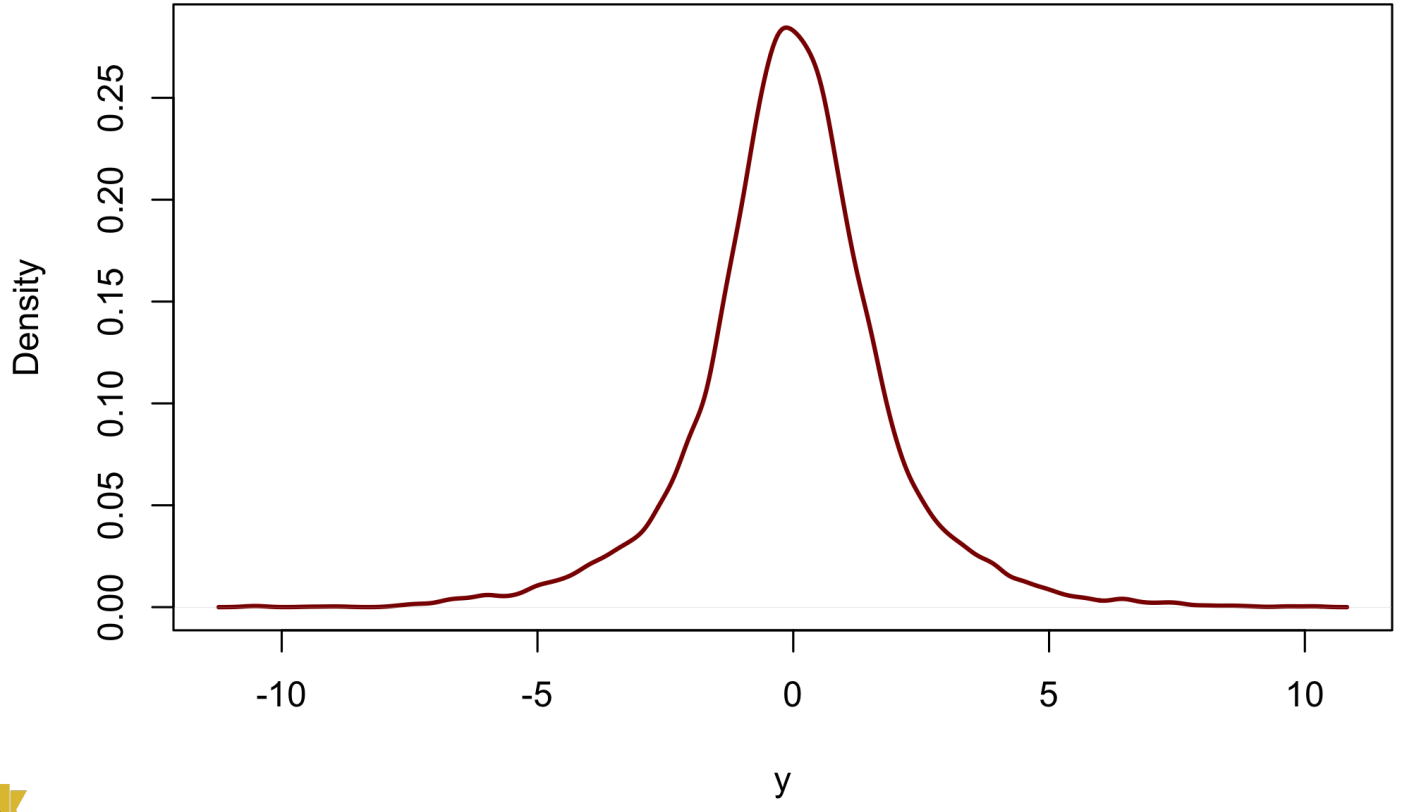
Location mixture example

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



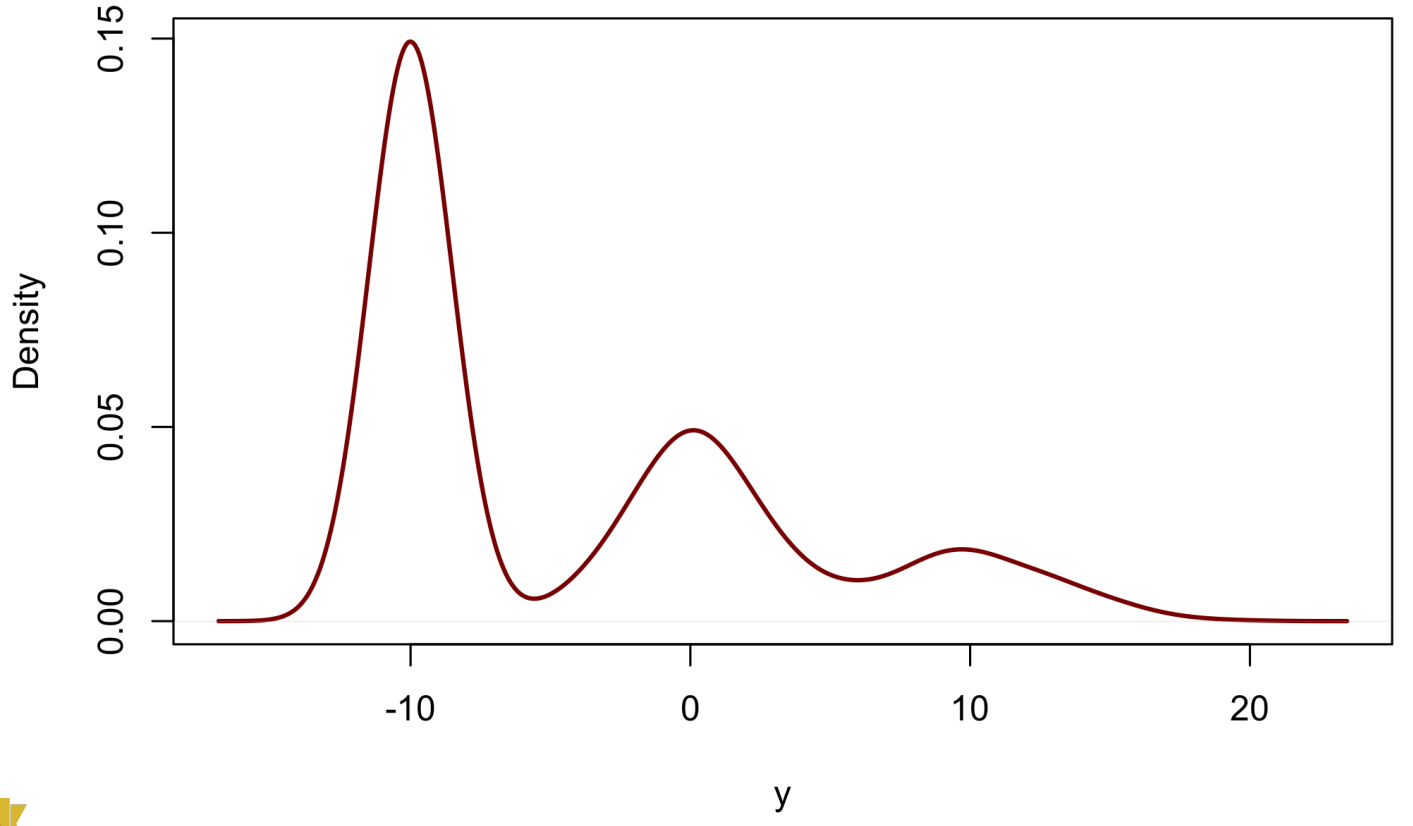
Scale mixture example

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



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 brings us the to concept of data augmentation, which we actually already used in the mixture of multinomials.
- 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

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

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

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

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

Full conditionals

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

$$\pi[\boldsymbol{\lambda}|\ldots] \equiv \text{Dirichlet}\left(a_1 + n_1, \ldots, a_d + n_d\right),$$

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

- As we will see in the illustration very soon, the 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.

In-class analysis: move to the R script here.

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.
  - Go Bayesian non-parametric: Dirichlet processes!

Back to categorical data again

Categorical data: bivariate case

- Suppose we have data (y_{i1}, y_{i2}), for $i = 1, \ldots, n$, where
  - $y_{i1} \in \{1, \ldots, D_1\}$
  - $y_{i2} \in \{1, \ldots, D_2\}$.
- This is just a two-way contingency table, so that we are interested in estimating the probabilities $\Pr(y_{i1} = d_1, y_{i2} = d_2) = \theta_{d_1 d_2}$.
- Write $\boldsymbol{\theta} = \{\theta_{d_1 d_2}\}$, which is a $D_1 \times D_2$ matrix of all the probabilities.
- The likelihood is therefore

$$L[Y|\boldsymbol{\theta}] = \prod_{i=1}^{n} \prod_{d_2=1}^{D_2} \prod_{d_1=1}^{D_1} \theta_{d_1 d_2}^{1[y_{i1}=d_1, y_{i2}=d_2]} = \prod_{d_2=1}^{D_2} \prod_{d_1=1}^{D_1} \theta_{d_1 d_2}^{\sum_{i=1}^{n} 1[y_{i1}=d_1, y_{i2}=d_2]} = \prod_{d_2=1}^{D_2} \prod_{d_1=1}^{D_1} \theta_{d_1 d_2}^{n_{d_1 d_2}}$$

  where $n_{d_1 d_2} = \sum_{i=1}^{n} 1[y_{i1} = d_1, y_{i2} = d_2]$ is just the number of observations in cell (d_1, d_2) of the contingency table.

Categorical data: bivariate case

- How can we do Bayesian inference? Several options! Most common are:
- Option 1: Follow the univariate approach.
  - rewrite the bivariate data as univariate data, that is, $y_i \in \{1, \ldots, D_1 D_2\}$;
  - write $\Pr(y_i = d) = \nu_d$ for each $d = 1, \ldots, D_1 D_2$;
  - specify Dirichlet prior as $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{D_1 D_2}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{D_1 D_2})$.
- Option 2: Assume independence, then follow the univariate approach.
  - write $\Pr(y_{i1} = d_1, y_{i2} = d_2) = \Pr(y_{i1} = d_1) \Pr(y_{i2} = d_2)$, so that $\theta_{d_1 d_2} = \lambda_{d_1} \psi_{d_2}$;
  - specify independent Dirichlet priors on λ_{d_1} and ψ_{d_2}, that is;
  - reduces number of parameters from $D_1 D_2 - 1$ to $D_1 + D_2 - 2$.

Categorical data: bivariate case

- Option 3: Log-linear model
  - $\theta_{d_1 d_2} = \dfrac{e^{\alpha_{d_1}+\beta_{d_2}+\gamma_{d_1 d_2}}}{\sum_{d_2}\sum_{d_1} e^{\alpha_{d_1}+\beta_{d_2}+\gamma_{d_1 d_2}}}$;
  - Specify priors (perhaps normal) on the parameters.
- Option 4: Latent structure model
  - Assume conditional independence given a latent variable;
  - That is, write

$$\begin{aligned}\theta_{d_1 d_2} &= \Pr(y_{i1}=d_1, y_{i2}=d_2) = \sum_{k=1}^{K} \Pr(y_{i1}=d_1, y_{i2}=d_2|z_i=k)\cdot\Pr(z_i=k)\\ &= \sum_{k=1}^{K} \Pr(y_{i1}=d_2|z_i=k)\cdot\Pr(y_{i2}=d_2|z_i=k)\cdot\Pr(z_i=k)\\ &= \sum_{k=1}^{K} \lambda_{k,d_1}\psi_{k,d_2}\cdot\omega_k.\end{aligned}$$

  - This is a finite mixture of multinomial distributions;

Categorical data: extensions

- For categorical data with more than two categorical variables, it is relatively easy to extend the framework for latent structure models.
- Clearly, there will be many more parameters (vectors and matrices) to keep track of, depending on the number of clusters and number of variables!
- If interested, read up on finite mixture of products of multinomials.
- Happy to provide resources for those interested!

Final remarks

- Unfortunately, this is as much as we can cover in this course.
- I hope you learned a lot about Bayesian inference, even with the many adjustments we had to make mid-semester due to the current state of the world.
- Now, just the final exam to look forward to!
- It has been a pleasure having you all in my class...
- For those who haven't, remember to complete the course evaluations!