

STA 360/602L: MODULE 8.5

FINITE MIXTURE MODELS: MULTIVARIATE CATEGORICAL DATA

DR. OLANREWAJU MICHAEL AKANDE

CATEGORICAL DATA: BIVARIATE CASE

- Suppose we have data (y_{i1}, y_{i2}) , for $i = 1, \dots, n$, where
 - $y_{i1} \in \{1, \dots, D_1\}$
 - $y_{i2} \in \{1, \dots, D_2\}$.
- 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 $\theta = \{\theta_{d_1 d_2}\}$, which is a $D_1 \times D_2$ matrix of all the probabilities.

CATEGORICAL DATA: BIVARIATE CASE

- The likelihood is therefore

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

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.

POSTERIOR INFERENCE

- 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, $y_i \in \{1, \dots, D_1 D_2\}$.
 - Write $\Pr(y_i = d) = \nu_d$ for each $d = 1, \dots, D_1 D_2$.
 - Specify Dirichlet prior as $\boldsymbol{\nu} = (\nu_1, \dots, \nu_{D_1 D_2}) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_{D_1 D_2})$.
 - Then, posterior is also Dirichlet with parameters updated with the number in each cell of the contingency table.

POSTERIOR INFERENCE

- **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 $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{D_1})$ and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_{D_2})$.
 - That is,
 - $\boldsymbol{\lambda} \sim \text{Dirichlet}(a_1, \dots, a_{D_1})$
 - $\boldsymbol{\psi} \sim \text{Dirichlet}(b_1, \dots, b_{D_2})$.
 - This reduces the number of parameters from $D_1 D_2 - 1$ to $D_1 + D_2 - 2$.

POSTERIOR INFERENCE

- **Option 3: Log-linear model**

- $$\theta_{d_1 d_2} = \frac{e^{\alpha_{d_1} + \beta_{d_2} + \gamma_{d_1 d_2}}}{\sum_{d_2=1}^{D_2} \sum_{d_1=1}^{D_1} e^{\alpha_{d_1} + \beta_{d_2} + \gamma_{d_1 d_2}}};$$

- Specify priors (perhaps normal) on the parameters.

POSTERIOR INFERENCE

- **Option 4:** Latent structure model
 - Assume conditional independence given a **latent variable**;
 - That is, write

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

- This is once again, 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](#).
- Can also go full Bayesian nonparametrics with a [Dirichlet process mixture of products of multinomials](#).
- Happy to provide resources for those interested!

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!