STA 610L: Module 4.5

Introduction to finite mixture models (categorical data)

Dr. Olanrewaju Michael Akande

STA 610L

Categorical data (univariate)

- Suppose
  - $Y \in \{1, \ldots, D\}$;
  - $\Pr(y = d) = \theta_d$ for each $d = 1, \ldots, D$; and
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$.
- Then the pmf of Y is

$$\Pr[y = d | \boldsymbol{\theta}] = \prod_{d=1}^{D} \theta_d^{1[y=d]}.$$

- We say Y has a multinomial distribution with sample size 1, or a categorical distribution.
- Write as $Y|\boldsymbol{\theta} \sim \text{Multinomial}(1, \boldsymbol{\theta})$ or $Y|\boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Clearly, this is just an extension of the Bernoulli distribution.

STA 610L

Dirichlet distribution

- Since the elements of the probability vector $\boldsymbol{\theta}$ must always sum to one, that is, its support is the $D-1$ simplex.
- A conjugate prior for categorical/multinomial data is the Dirichlet distribution.
- A random variable $\boldsymbol{\theta}$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, if

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D}\theta_d^{\alpha_d-1}, \quad \alpha_d > 0 \text{ for all } d=1,\ldots,D.$$

where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_D)$, and

$$\sum_{d=1}^{D}\theta_d = 1, \;\; \theta_d \geq 0 \text{ for all } d=1,\ldots,D.$$

- We write this as $\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}) = \mathrm{Dirichlet}(\alpha_1,\ldots,\alpha_D)$.
- The Dirichlet distribution is a multivariate generalization of the beta distribution.

STA 610L

Dirichlet distribution

- Write

$$\alpha_0 = \sum_{d=1}^{D} \alpha_d \quad \text{and} \quad \alpha_d^\star = \frac{\alpha_d}{\alpha_0}.$$

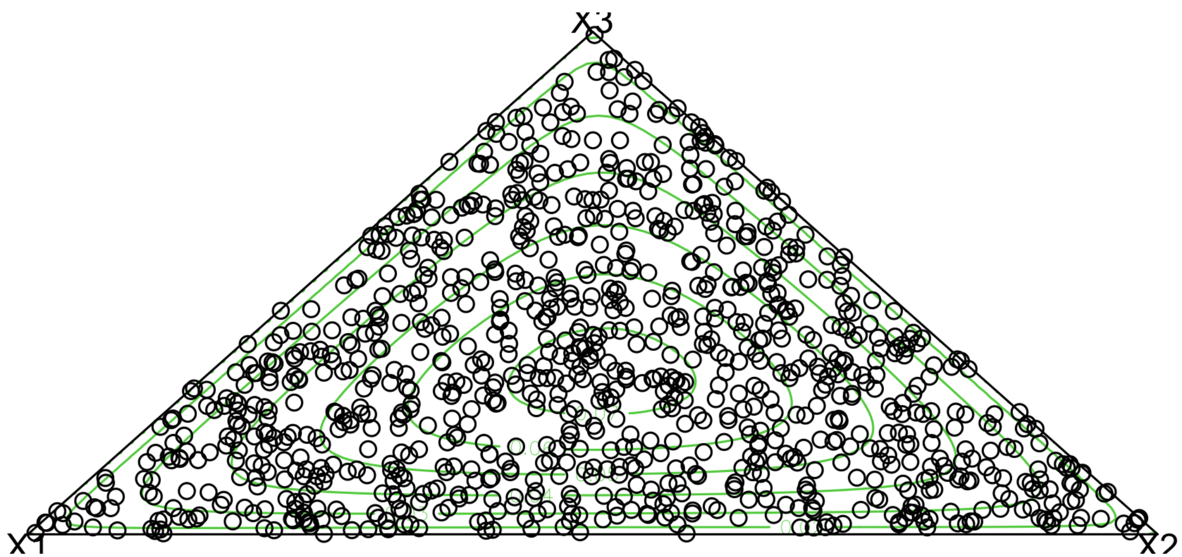
- Then we can re-write the pdf as

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma\left(\alpha_0\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} \theta_d^{\alpha_d - 1}, \quad \alpha_d > 0 \ \text{ for all } \ d = 1, \ldots, D.$$

- Properties:
  - $$\mathbb{E}[\theta_d] = \alpha_d^\star;$$
  - $$\text{Mode}[\theta_d] = \frac{\alpha_d - 1}{\alpha_0 - d};$$
  - $$\mathbb{V}\text{ar}[\theta_d] = \frac{\alpha_d^\star(1 - \alpha_d^\star)}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_d](1 - \mathbb{E}[\theta_d])}{\alpha_0 + 1};$$
  - $$\mathbb{C}\text{ov}[\theta_d, \theta_k] = \frac{\alpha_d^\star \alpha_k^\star}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_d]\mathbb{E}[\theta_k]}{\alpha_0 + 1}.$$

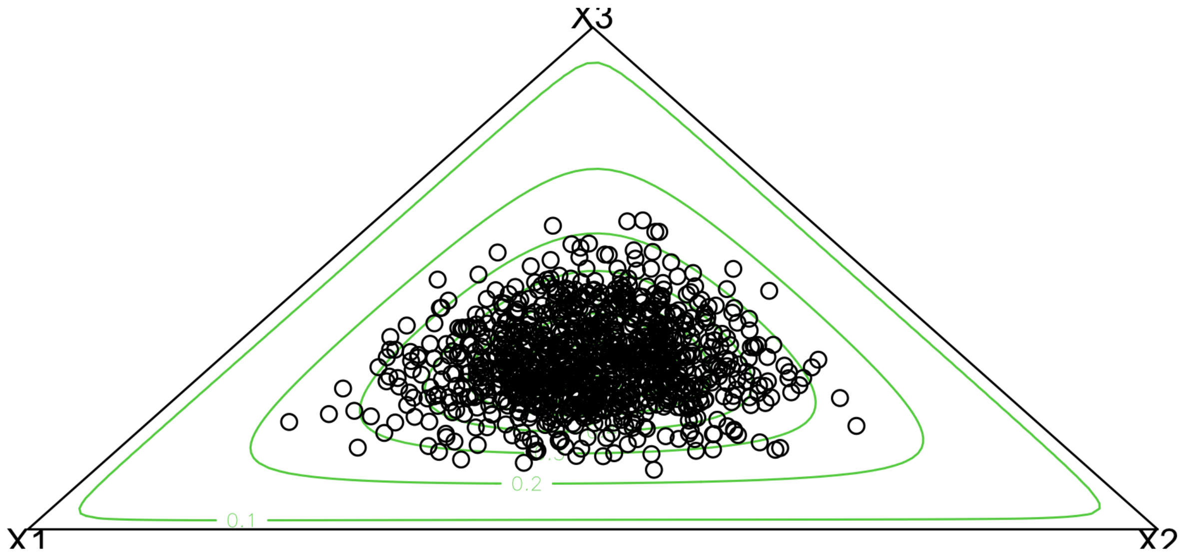
DIRICHLET EXAMPLES

$\text{Dirichlet}(1, 1, 1)$



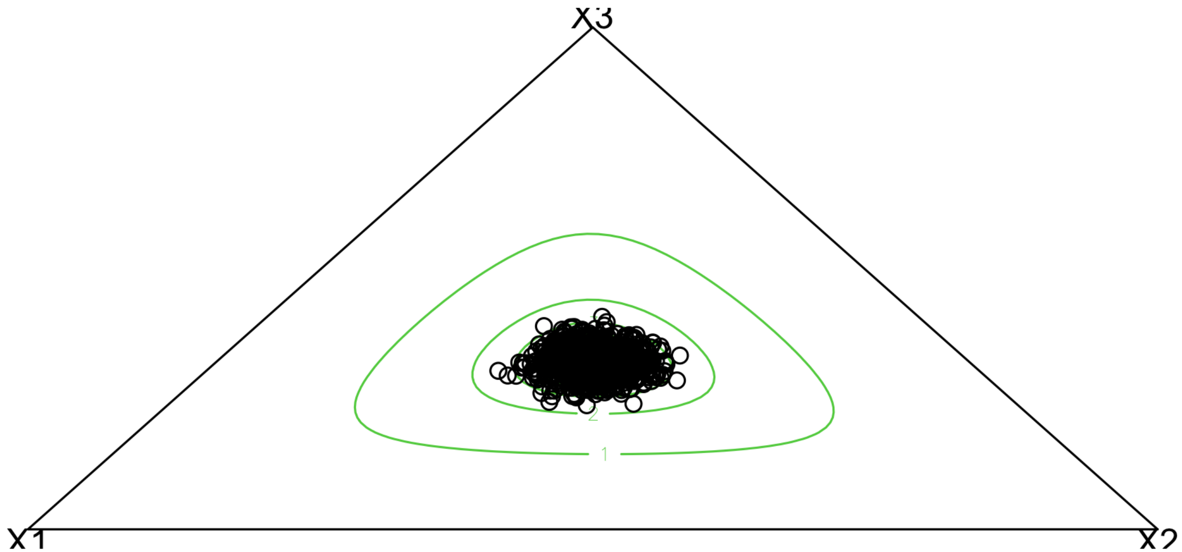
Dirichlet examples

$\text{Dirichlet}(10, 10, 10)$



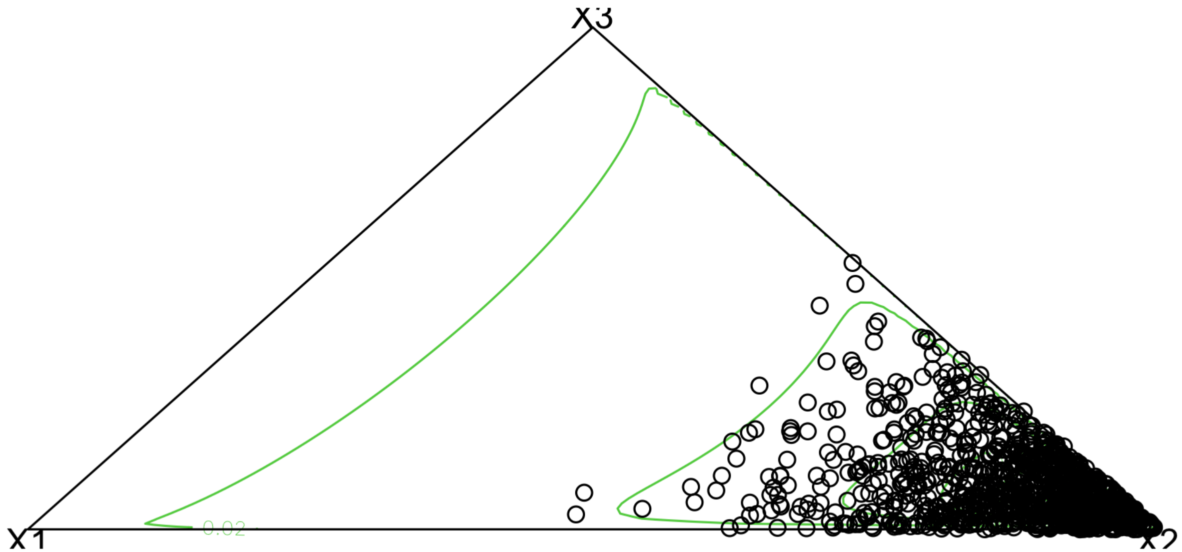
Dirichlet examples

$\text{Dirichlet}(100, 100, 100)$



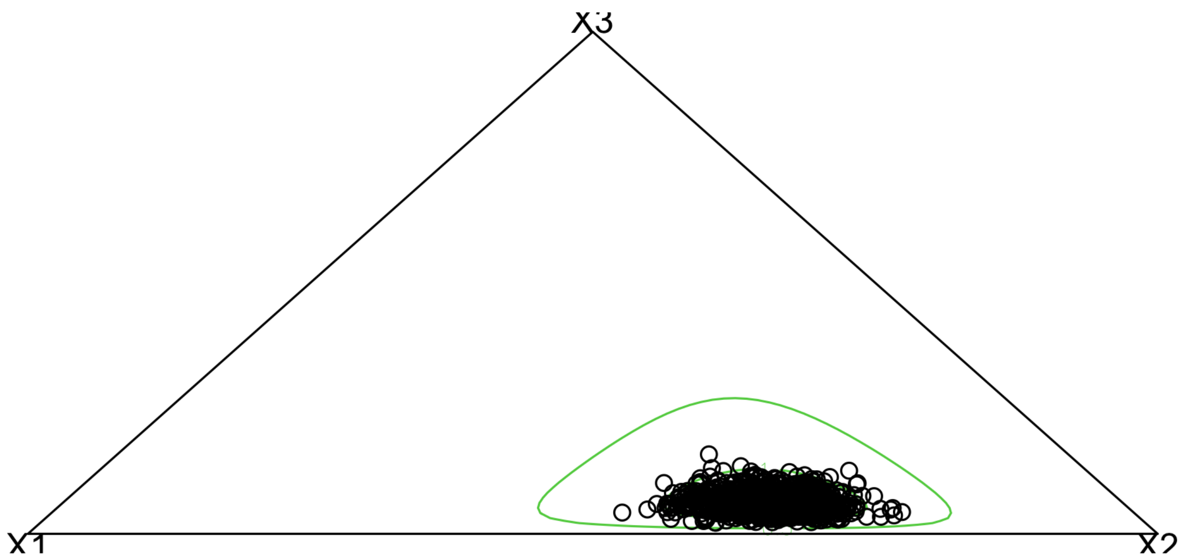
Dirichlet examples

$\text{Dirichlet}(1, 10, 1)$



Dirichlet examples

$\text{Dirichlet}(50, 100, 10)$



Likelihood

- Let $Y_i, \ldots, Y_n | \boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Recall

$$\Pr[y_i = d | \boldsymbol{\theta}] = \prod_{d=1}^{D} \theta_d^{1[y_i = d]}.$$

- Then,

$$p[Y | \boldsymbol{\theta}] = p[y_1, \ldots, y_n | \boldsymbol{\theta}] = \prod_{i=1}^{n} \prod_{d=1}^{D} \theta_d^{1[y_i = d]} = \prod_{d=1}^{D} \theta_d^{\sum_{i=1}^{n} 1[y_i = d]} = \prod_{d=1}^{D} \theta_d^{n_d}$$

where n_d is just the number of individuals in category d.

- Maximum likelihood estimate of θ_d is

$$\hat{\theta}_d = \frac{n_d}{n}, \quad d = 1, \ldots, D$$

STA 610L

Posterior

- Set $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_D)$. Then

$$\begin{aligned}\pi(\boldsymbol{\theta}|Y) &\propto p[Y|\boldsymbol{\theta}] \cdot \pi[\boldsymbol{\theta}] \\ &\propto \prod_{d=1}^{D} \theta_d^{n_d} \prod_{d=1}^{D} \theta_d^{\alpha_d - 1} \\ &\propto \prod_{d=1}^{D} \theta_d^{\alpha_d + n_d - 1} \\ &= \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_D + n_D)\end{aligned}$$

- Posterior expectation:

$$\mathbb{E}[\theta_d|Y] = \frac{\alpha_d + n_d}{\sum_{d^\star=1}^{D}(\alpha_{d^\star} + n_{d^\star})}.$$

- We can also extend the Dirichlet-multinomial model to more variables (contingency tables).

- First, what if our data actually comes from K different sub-populations of groups of people?

Finite mixture of multinomials

- For example, if our categorical data comes from men and women, and we don't expect marginal independence across the two groups, then we have a mixture of distributions.
- With our data coming from a "combination" or "mixture" of sub-populations, we no longer have independence across all observations, so that the likelihood $p[Y|\boldsymbol{\theta}] \neq \prod_{i=1}^{n} \prod_{d=1}^{D} \theta_j^{1[y_i=d]}$.
- However, we can still have "conditional independence" within each group.
- Unfortunately, we do not always know the indexes for those groups.
- That is, we know our data contains K different groups, but we actually do not know which observations belong to which groups.
- **Solution**: introduce a latent variable z_i representing the group/cluster indicator for each observation i, so that each $z_i \in \{1, \ldots, K\}$.

STA 610L

Finite mixture of multinomials

- Given the cluster indicator z_i for observation i, write
  - $\Pr(y_i = d|z_i) = \psi_{z_i,d} \equiv \prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i=d|z_i]}$, and
  - $\Pr(z_i = k) = \lambda_k \equiv \prod_{k=1}^{K} \lambda_k^{1[z_i=k]}$.
- Then, the marginal probabilities we care about will be

$$\begin{aligned}\theta_d &= \Pr(y_i = d) \\ &= \sum_{k=1}^{K} \Pr(y_i = d|z_i = k) \cdot \Pr(z_i = k) \\ &= \sum_{k=1}^{K} \lambda_k \cdot \psi_{k,d},\end{aligned}$$

which is a finite mixture of multinomials, with the weights given by λ_k.

Posterior inference

- Write
  - $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, and
  - $\boldsymbol{\psi} = \{\psi_{z_i,d}\}$ to be a $K \times D$ matrix of probabilities, where each kth row is the vector of probabilities for cluster k.
- The observed data likelihood is

$$
\begin{aligned}
p\left[Y = (y_1, \ldots, y_n) | Z = (z_1, \ldots, z_n), \boldsymbol{\psi}, \boldsymbol{\lambda}\right] &= \prod_{i=1}^{n} \prod_{d=1}^{D} \Pr\left(y_i = d | z_i, \psi_{z_i,d}\right) \\
&= \prod_{i=1}^{n} \prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i = d | z_i]},
\end{aligned}
$$

which includes products (and not the sums in the mixture pdf), and as you will see, makes sampling a bit easier.

- Next we need priors.

STA 610L

Posterior inference

- First, for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, the vector of cluster probabilities, we can use a Dirichlet prior. That is,

$$\pi[\boldsymbol{\lambda}] = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \propto \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}.$$

- For $\boldsymbol{\psi}$, we can assume independent Dirichlet priors for each cluster vector $\boldsymbol{\psi}_k = (\psi_{k,1}, \ldots, \psi_{k,D})$. That is, for each $k = 1, \ldots, K$,

$$\pi[\boldsymbol{\psi}_k] = \text{Dirichlet}(a_1, \ldots, a_d) \propto \prod_{d=1}^{D} \psi_{k,d}^{a_d - 1}.$$

- Finally, from our distribution on the z_i's, we have

$$p\left[Z = (z_1, \ldots, z_n) | \boldsymbol{\lambda}\right] = \prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}.$$

STA 610L

Posterior inference

- Note that the unobserved variables and parameters are $Z = (z_1, \ldots, z_n)$, $\boldsymbol{\psi}$, and $\boldsymbol{\lambda}$.
- So, the joint posterior is

$$
\begin{aligned}
\pi\left(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y\right) &\propto p\left[Y | Z, \boldsymbol{\psi}, \boldsymbol{\lambda}\right] \cdot p(Z | \boldsymbol{\psi}, \boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) \\
&\propto \left[\prod_{i=1}^{n} \prod_{d=1}^{D} p\left(y_i = d | z_i, \psi_{z_i,d}\right)\right] \cdot p(Z | \boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\psi}) \cdot \pi(\boldsymbol{\lambda}) \\
&\propto \left(\prod_{i=1}^{n} \prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i = d | z_i]}\right) \\
&\quad \times \left(\prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{1[z_i = k]}\right) \\
&\quad \times \left(\prod_{k=1}^{K} \prod_{d=1}^{D} \psi_{k,d}^{a_d - 1}\right) \\
&\quad \times \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}\right).
\end{aligned}
$$

STA 610L

Posterior inference

- First, we need to sample the z_i's, one at a time, from their 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

$$\begin{aligned}
\Pr[z_i = k | \ldots] &= \Pr[z_i = k | y_i, \boldsymbol{\psi}_k, \lambda_k] \\
&= \frac{\Pr[y_i, z_i = k | \boldsymbol{\psi}_k, \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i, z_i = l | \boldsymbol{\psi}_l, \lambda_l]} \\
&= \frac{\Pr[y_i | z_i = k, \boldsymbol{\psi}_k] \cdot \Pr[z_i = k, \lambda_k]}{\sum_{l=1}^{K} \Pr[y_i | z_i = l, \boldsymbol{\psi}_l] \cdot \Pr[z_i = l, \lambda_l]} \\
&= \frac{\psi_{k,d} \cdot \lambda_k}{\sum_{l=1}^{K} \psi_{l,d} \cdot \lambda_l}.
\end{aligned}$$

STA 610L

Posterior inference

- Next, sample each cluster vector $\boldsymbol{\psi}_k = (\psi_{k,1}, \ldots, \psi_{k,D})$ from

$$\begin{aligned}
\pi[\boldsymbol{\psi}_k | \ldots] &\propto \pi\left(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y\right) \\
&\propto \left(\prod_{i=1}^{n} \prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i=d|z_i]}\right) \cdot \left(\prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{1[z_i=k]}\right) \cdot \left(\prod_{k=1}^{K} \prod_{d=1}^{D} \psi_{k,d}^{a_d-1}\right) \cdot \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k-1}\right) \\
&\propto \left(\prod_{d=1}^{D} \psi_{k,d}^{n_{k,d}}\right) \cdot \left(\prod_{d=1}^{D} \psi_{k,d}^{a_d-1}\right) \\
&= \left(\prod_{d=1}^{D} \psi_{k,d}^{a_d+n_{k,d}-1}\right) \\
&\equiv \text{Dirichlet}\left(a_1 + n_{k,1}, \ldots, a_D + n_{k,D}\right).
\end{aligned}$$

where $n_{k,d} = \sum_{i:z_i=k} 1[y_i = d]$, the number of individuals in cluster k that are assigned to category d of the levels of y.

STA 610L

Posterior inference

- Finally, sample $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, the vector of cluster probabilities from

$$
\begin{aligned}
\pi[\boldsymbol{\lambda}| \ldots] &\propto \pi\left(Z, \boldsymbol{\psi}, \boldsymbol{\lambda} | Y\right) \\
&\propto \left(\prod_{i=1}^{n}\prod_{d=1}^{D} \psi_{z_i,d}^{1[y_i=d|z_i]}\right) \cdot \left(\prod_{i=1}^{n}\prod_{k=1}^{K} \lambda_k^{1[z_i=k]}\right) \cdot \left(\prod_{k=1}^{K}\prod_{d=1}^{D} \psi_{k,d}^{a_d-1}\right) \cdot \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k-1}\right) \\
&\propto \left(\prod_{i=1}^{n}\prod_{k=1}^{K} \lambda_k^{1[z_i=k]}\right) \cdot \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k-1}\right) \\
&\propto \left(\prod_{k=1}^{K} \lambda_k^{n_k}\right) \cdot \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k-1}\right) \\
&\propto \left(\prod_{k=1}^{K} \lambda_k^{\alpha_k+n_k-1}\right) \\
&\equiv \text{Dirichlet}\left(\alpha_1 + n_1, \ldots, \alpha_K + n_K\right),
\end{aligned}
$$

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

STA 610L

Categorical data: bivariate case

- How can we extend the same ideeas to multiple categorical variables?
- Well let's start small. 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.

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.

STA 610L

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

STA 610L

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

STA 610L

Posterior inference

- Option 3: Log-linear model
  - $\theta_{d_1 d_2} = \dfrac{e^{\alpha_{d_1} + \beta_{d_2} + \gamma_{d_1 d_2}}}{\displaystyle\sum_{d_2=1}^{D_2} \sum_{d_1=1}^{D_1} e^{\alpha_{d_1} + \beta_{d_2} + \gamma_{d_1 d_2}}}$;
  - Specify priors (perhaps normal) on the parameters.

STA 610L

Posterior inference

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

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

  - This is once again, a finite mixture of multinomial distributions.

STA 610L

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!