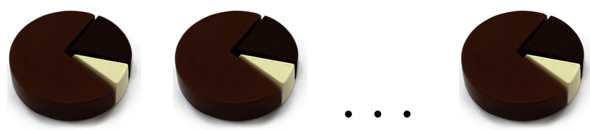
MG-RAST Nested Data

scat

February 16, 2011



Have some cheesecake.

1 DP Data Mining – Nice tool to have

For $i = 1, \cdots, n$,

$$\begin{aligned} Y_i &\equiv \{X_{i1}, X_{i2}, \cdots, X_{iK}\} \\ Y_i &\sim \text{Multinomial}(N_i, p_i) \\ X_{ik} &\in \mathbb{N}^+ \cup \{0\} \\ N_i &= \sum_{k=1}^{K} X_{ik} \end{aligned}$$

Model

$$\begin{aligned} p_i &\sim \text{Dirichlet}(\theta_i) \\ \theta_i &\sim G \\ G &\sim \text{DP}(\alpha G_0) \\ \alpha &\sim \text{Gamma}(a, b) \\ G_0 &\equiv \text{Normal}_+(\mu, \Sigma) \end{aligned}$$

Or if that's not quite appropriate for modeling the p_i, model, for example,

$$\begin{aligned} p_i &\sim \text{Dirichlet}(\theta + O_i) \\ O_i &\sim G \\ G &\sim \text{DP}(\alpha G_0) \\ \alpha &\sim \text{Gamma}(a, b) \\ G_0 &\equiv \pi\delta_0 + (1 - \pi)\text{Normal}(\mu, \Sigma) \end{aligned}$$

Choices for G_0 are important. There's lots of data to do this in an empirical fashion. Inference on θ locates interesting pies Y_i (clusters, perhaps). This could be applied across any single hierarchical data level. And then I don't have to look through this data for interesting things myself.

2 Three level nested data – DP hierarchical modeling

For $i = 1, \cdots, n$,

$$\begin{aligned}
Y_i^{(1)} &\equiv \{X_{i1}^{(1)}, X_{i2}^{(1)}, \cdots, X_{iQ^{(1)}}^{(1)}\} \\
Y_i^{(1)} &\sim \text{Multinomial}(N_i, p_i^{(1)}) \\
X_{ij}^{(1)} &\in \mathbb{N}^+ \cup \{0\} \\
N_i &= \sum_{j=1}^{Q^{(1)}} X_{ij}^{(1)}
\end{aligned}$$

$$\begin{aligned}
Y_{ij}^{(2)} &\equiv \{X_{ij1}^{(2)}, X_{ij2}^{(2)}, \cdots, X_{ijQ^{(2)}}^{(2)}\} \\
Y_{ij}^{(2)} &\sim \text{Multinomial}(X_{ij}^{(1)}, p_{ij}^{(2)}) \\
X_{ijk}^{(2)} &\in \mathbb{N}^+ \cup \{0\} \\
X_{ij}^{(1)} &= \sum_{k=1}^{Q^{(2)}} X_{ijk}^{(2)}
\end{aligned}$$

$$\begin{aligned}
Y_{ijk}^{(3)} &\equiv \{X_{ijk1}^{(3)}, X_{ijk2}^{(3)}, \cdots, X_{ijkQ^{(3)}}^{(3)}\} \\
Y_{ijk}^{(3)} &\sim \text{Multinomial}(X_{ijk}^{(2)}, p_{ijk}^{(3)}) \\
X_{ijkl}^{(3)} &\in \mathbb{N}^+ \cup \{0\} \\
X_{ijk}^{(2)} &= \sum_{l=1}^{Q^{(3)}} X_{ijkl}^{(3)}
\end{aligned}$$

This is a three level nested data structure:

- Individual i has a 1^{st} level multinomial random variable, $Y_i^{(1)}$.
- The j^{th} category in $Y_i^{(1)}$ (with $X_{ij}^{(1)}$ counts) may itself be subdivided into a multinomial random variable, $Y_{ij}^{(2)}$.

- The k^{th} category in $Y_{ij}^{(2)}$ (with $X_{ijk}^{(2)}$ counts) may itself be subdivided into a multinomial random variable, $Y_{ijk}^{(3)}$.
- The l^{th} category in $Y_{ijk}^{(3)}$ will have $X_{ijkl}^{(2)}$ counts.

Consider just the top two levels (ignore the bottom third level), and model

$$
\begin{aligned}
p_i^{(1)} &\sim \text{Dirichlet}(\theta^{(1)} + A_i^{(1)}) \\
\theta^{(1)} &\sim \text{Normal}_+(\mu_\theta^{(1)}, \Sigma_\theta^{(1)}) \\
A_i^{(1)} &\equiv \{A_{i1}^{(1)}, A_{i2}^{(1)}, \cdots, A_{iQ^{(1)}}^{(1)}\} \\
A_{ij}^{(1)} &\sim G_j^{(1)} \\
G_j^{(1)} &\sim \text{DP}_j(\alpha_j^{(1)} G_0^{(1)}) \\
\alpha_j^{(1)} &\sim \text{Gamma}_j(a^{(1)}, b^{(1)}) \\
G_0 &\equiv \pi^{(1)}\delta_0 + (1 - \pi^{(1)})\text{Normal}(\mu_A^{(1)}, \sigma_A^{(1)})
\end{aligned}
$$

$$
\begin{aligned}
p_{ij}^{(2)} &\sim \text{Dirichlet}(\theta_{ij}^{(2)} + A_{ij}^{(2)}) \\
\theta_{ij}^{(2)} | A_{ij}^{(1)} = a_{ij} &\sim \text{Normal}_+(\mu_{\theta a_{ij}}^{(2)}, \Sigma_{\theta a_{ij}}^{(2)}) \\
A_{ij}^{(2)} &\equiv \{A_{ij1}^{(2)}, A_{ij2}^{(2)}, \cdots, A_{ijQ^{(2)}}^{(2)}\} \\
A_{ijk}^{(2)} | A_{ij}^{(1)} = a_{ij} &\sim G_{jka_{ij}}^{(2)} \\
G_{jka_{ij}}^{(2)} &\sim \text{DP}_{jka_{ij}}(\alpha_{jka_{ij}}^{(2)} G_0^{(2)}) \\
\alpha_{jka_{ij}}^{(2)} &\sim \text{Gamma}_{jka_{ij}}(a^{(2)}, b^{(2)}) \\
G_0 &\equiv \text{Normal}(\mu_A^{(2)}, \sigma_A^{(2)})
\end{aligned}
$$

- The DPs cluster the i's element wise on the A adjustments.
- Each of the elements in the A adjustment become a new multinomial random variable. The clusters get the same $\theta^{(2)}$.
- The extension from the second to the third level would be analogous to that of the extension from the first to the second.
- I view this as some kind of bubbling process. When there's a cluster in a top level category, those in the cluster grow a separate parameter in the next level.
- Writing this out is all well and good... fitting this thing is going to be something else.

- I also am not sure if a difference between a category at one level means that there should be a difference in the next level breakdown of that category... but that's what we're doing here.