Reduction to the dirichlet Distribution

The Dirichlet distribution is defined for $k = 1, \ldots, K$ taxa, where $\sum_{i=1}^{K} x_j = 1$

On the other hand, the Generalized Dirichlet distribution is defined for $k = 1, \ldots, K-1$ taxa, and $x_K = 1 - \sum_{i=1}^{K-1} X_j$.

Claim: The Generalized Dirichlet distribution reduces to the generalized distribution when $\beta_j = \alpha_{j+1} + \beta_{j+1}$

The probability density function of the generalized dirichlet distribution for $x_1, \ldots, x_{k-1}$ is

$$\left[\prod_{i=1}^{k-1} B(\alpha_i, \beta_i)\right]^{-1} x_k^{\beta_{k-1}-1} \prod_{i=1}^{k-1} \left[x_i^{\alpha_i - 1} \left(\sum_{j=1}^{k} x_j \right)^{\beta_{i-1} - (\alpha_i + \beta_i)} \right]$$

where $x_k = 1 - \sum_{i=1}^{k-1} x_i$

(Where in this implies ordering matters?)

When $\beta_j = \alpha_{j+1} + \beta_{j+1}$,

$$\begin{aligned}
\left[\prod_{i=1}^{k-1} B(\alpha_i, \beta_i)\right]^{-1} x_k^{\beta_{k-1}-1} \prod_{i=1}^{k-1} \left[x_i^{\alpha_i - 1} \left(\sum_{j=1}^{k} x_j \right)^{\beta_{i-1} - (\alpha_i + \beta_i)} \right] &= \left[\prod_{i=1}^{k-1} B(\alpha_i, \beta_i)\right]^{-1} x_k^{\beta_k + \alpha_k - 1} \prod_{i=1}^{k-1} \left[x_i^{\alpha_i - 1} \left(\sum_{j=1}^{k} x_j \right)^{\beta_{i-1} - \beta_{i-1}} \right] \\
&= \left[\prod_{i=1}^{k-1} B(\alpha_i, \beta_i)\right]^{-1} x_k^{\beta_k + \alpha_k - 1} \prod_{i=1}^{k-1} \left[x_i^{\alpha_i - 1} \right] \\
&= \left[\prod_{i=1}^{k-1} B(\alpha_i, \beta_i)\right]^{-1} x_k^{\beta_k} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \\
&= \left[\prod_{i=1}^{k-1} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\right] x_k^{\beta_k} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \\
&= \left[\frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_k + \beta_k) \prod_{i=1}^{k-1} \Gamma(\alpha_i)}\right] x_k^{\beta_k} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \\
&= \left[\frac{\Gamma(\alpha_1 + \beta_k + \sum_{i=2}^{k} \alpha_i)}{\Gamma(\alpha_k + \beta_k) \prod_{i=1}^{k-1} \Gamma(\alpha_i)}\right] x_k^{\beta_k} \prod_{i=1}^{k} x_i^{\alpha_i - 1}
\end{aligned}$$

Does this mean we set $\beta_k = 0$?

Is this order invariant?

Dirichlet distribution density for $(x_1, \ldots, x_k)$

$$\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

Generalized Dirichlet Distribution in terms of Beta RVs

$Z_j \sim Beta(\alpha_j, \alpha_{j+1} + \beta_{j+1})$ for $j = 1, \dots, k-1$

Equivalently, $Z_j \sim \text{Beta}(\alpha_j, \sum_{i=j+1}^{k} \alpha_i + \beta_k)$

$$X_1 = Z_1, \qquad X_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i) \qquad X_k = 1 - \sum_{i=1}^{k-1} X_i$$

Then, $\mathbf{X} = (X_1, \dots, X_k)$ is GDM distributed

Compare to Dirichlet:

Dirichlet distribution marginals: $X_j \sim \beta(\alpha_j, \alpha_0 - \alpha_j)$

If we can use the standard Dirichlet distribution, but use the 'stick-breaking' approach that builds the GD, and add zero inflation there, should have fewer parameters needed to estimate.

Zero-Inflated Dirichlet Distribution

Use the same idea of adding zero-inflation to the Z_j Beta distributions that create the Generalized distribution, but use parameters for the Z_j so the GD reduces to the Dirichlet.

Let $\Delta_i \sim Bern(\pi_i)$ be the indication that taxa i is absent. ($\Delta_i = 1$ implies taxa i is a structural zero, $\Delta_i = 0$ implies taxa i follows a Beta distribution)

Let $Z_1, \dots, Z_K$ be independent, $Z_i = 0$ if $\Delta_i = 1$, $Z_i \sim Beta(\alpha_i, \beta_k + \sum_{j=i+1}^{k} \alpha_i)$ if $\Delta_i = 0$

Let

$$X_1 = Z_1, \quad X_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i)$$

Then,

$$
\begin{aligned}
E(X_1) &= E(Z_1) \\
&= (1-\pi_1)\frac{\alpha_1}{\alpha_1 + \sum_{i=2}^k \alpha_i} \\
&= (1-\pi_1)\frac{\alpha_1}{\alpha_0} \\
E(X_2) &= E(Z_1(1-Z_2)) \\
&= E(Z_2)E(1-Z_1) \quad \text{Since independent} \\
&= (1-\pi_2)\frac{\alpha_2}{\sum_{i=2}^k \alpha_2}\left[\pi_1 + (1-\pi_1)\frac{\sum_{i=2}^k \alpha_i}{\alpha_0}\right] \qquad \text{Since } 1-Z_1 \sim \text{Beta}(\sum_{i=2}^k \alpha_i, \alpha_1) \\
E(X_j) &= E(Z_j)E(1-Z_{j-1})\cdots E(1-Z_1) \\
&= (1-\pi_j)\frac{\alpha_j}{\sum_{i=j}^k \alpha_i}\prod_{i=1}^{j-1}\left[\pi_i + (1-\pi_i)\frac{\sum_{i=2}^k \alpha_i}{\alpha_0}\right]
\end{aligned}
$$

Does ordering still matter here?

Without zero-inflation:

assuming $i < j$:

$$
\begin{aligned}
Cov(X_i, X_j) &= E(X_iX_j) - E(X_i)E(X_j) \\
E(X_iX_j) &= E\left[Z_iZ_j\prod_{s=1}^{i-1}(1-Z_s)\prod_{t=1}^{j-1}(1-Z_t)\right] \\
&= E\left[Z_iZ_j\prod_{s=1}^{i-1}(1-Z_s)^2\prod_{t=i}^{j-1}(1-Z_t)\right] \\
&= E\left[\left(\prod_{s=1}^{i-1}(1-Z_s)^2\right)Z_i(1-Z_i)\left(\prod_{t=i+1}^{j-1}(1-Z_t)\right)Z_j\right] \\
&= \left(\prod_{s=1}^{i-1}E\left[(1-Z_s)^2\right]\right)E(Z_i(1-Z_i))\left(\prod_{t=i+1}^{j-1}E(1-Z_t)\right)E(Z_j)
\end{aligned}
$$

We have that for $Z_j \sim Beta(\alpha_j, \beta_j)$,

$$
\begin{aligned}
E(Z_j(1-Z_j)) &= \frac{\Gamma(\alpha_j+\beta_j)\Gamma(\alpha_j+1)\Gamma(\beta_j+1)}{\Gamma(\alpha_j)\Gamma(\beta_j)\Gamma(\alpha_j+\beta_j+2)} \\
&= \frac{\alpha_j\beta_j}{(\alpha_j+\beta_j)(\alpha_j+\beta_j+1)}
\end{aligned}
$$