A major component of the Bayesian hierarchical clustering algorithm is the marginal likelihood of the data from the combination of two previous clusters assuming that the combined data comes from the same probabilistic model $p(\mathbf{x}|\theta)$. Per Heller & Ghahramani, let $\mathcal{D}_k = \mathcal{D}_i \cup \mathcal{D}_j$ be the the data under the union of the two clusters i and j. From equation 1 of Heller, the probability of the data $\mathcal{D}_k$ under the hypothesis $\mathcal{H}_1^k$ that the data comes from the same distribution is:

$$\begin{aligned} p(\mathcal{D}_k|\mathcal{H}_1^k) &= \int p(\mathcal{D}_k|\theta)p(\theta|\beta)d\theta \\ &= \int \left[\prod_{\mathbf{x}^{(i)}\in\mathcal{D}_k} p(\mathbf{x}^{(i)}|\theta)\right] p(\theta|\beta)d\theta \end{aligned}$$

As per Heller, models with conjugate priors, such as Normal-Inverse Wishart priors for Normal continuous data pr Dirichlet priors for Multinomial discrete data, leads to tractable integrals where the results are simple functions of sufficient statistics of $\mathcal{D}_k$. However, Heller does not provide these results, so they are derived below.

1 Multinomial Data

For multinomial data, the parameter θ is a q-dimensional vector $\mathbf{p}$ where q is the total number of categories in the data. The conjugate prior is a Dirichlet distribution with concentration parameter β. Therefore, the pdfs for these distributions are:

$$\begin{aligned} p(\mathcal{D}_k|\mathbf{p}) &= \frac{n!}{x_1!\dots x_q!}p_1^{x_1}\dots p_q^{x_q} \\ p(\mathbf{p}|\beta) &= \frac{1}{B(\beta)}p_1^{\beta_1-1}\dots p_q^{\beta_q-1} \end{aligned}$$

where $n = \sum x_i$ and B is the multinomial beta function, $B(\beta) = \dfrac{\prod_{i=1}^k \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^k \beta_i)}$.

It follows that

$$p(\mathcal{D}_k, \mathbf{p}|\beta) = \frac{n!}{B(\beta)x_1!\dots x_q!}p_1^{x_1+\beta_1-1}\dots p_q^{x_q+\beta_q-1}$$

This has the kernel of another Dirichlet distribution, with parameters $x_1+\beta_1, x_2+\beta_2, \dots$. Therefore, multiply and divide by the normalizing constant for a Dirichlet distribution with these parameters. We can then integrate out the Dirichlet distribution over $\mathbf{p}$, leaving us with:

$$\begin{aligned} p(\mathcal{D}_k|\beta) &= \frac{n!}{x_1!\dots x_q!}\frac{B(\beta^*)}{B(\beta)} \\ &= \frac{n!}{x_1!\dots x_q!}\frac{\prod_{i=1}^k \Gamma(x_i+\beta_i)}{\prod_{i=1}^k \Gamma(\beta_i)}\frac{\Gamma(\sum_{i=1}^k \beta_i)}{\Gamma(\sum_{i=1}^k x_i+\beta_i)} \\ &= \frac{n!}{x_1!\dots x_q!}\frac{\prod_{i=1}^k \Gamma(x_i+\beta_i)}{\prod_{i=1}^k \Gamma(\beta_i)}\frac{\Gamma(\sum_{i=1}^k \beta_i)}{\Gamma(n+\sum_{i=1}^k \beta_i)} \end{aligned}$$

Assuming that β_i are constant hyperparameters, their sum can be precomputed to accelerate processing.

2 Multivariate Normal Data