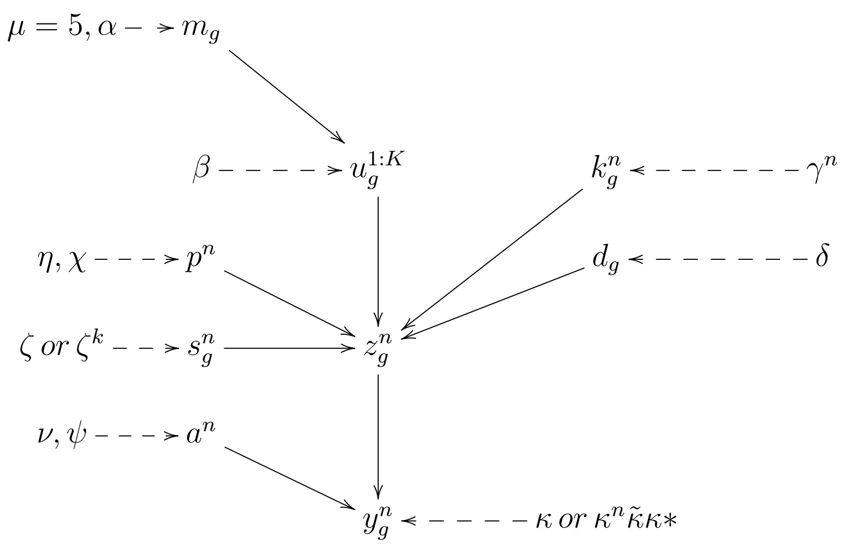
1 Project: A method to cluster single cell (SC) data.

1.1 Variables



Warning: this DAG is not 100% correct for all options of the model. For instance there should be an edge from k_g^n to s_g^n when ζ^k is introduced.

- N number of cells; G number of gene fragments; K is the number of cell clusters
- k_g^n represents the cluster of the cell n for gene g (with values in $1:K$)
- p^n: scaling before sampling specific to each cell n (cell size, capturing, amplification coefficient).
  - $p^n \sim \Gamma(mean = \eta, shape = \chi)$; $E(p^n) = \eta$; $Var(p^n) = \eta^2/\chi$
  - $\eta \sim \mathcal{E}(1)$;
  - $\chi \sim \mathcal{E}(1)$;
- s_g^n: a scaling factor for g and cell n
  - The algorithm can be run with $s_g^n = 1$ for $g = 1:G$ and $n = 1:N$ (no overdispersion of z_g^n around $p^n u_g^{k_g^n}$).
    OR
  - $s_g^n \sim \Gamma(mean = 1, shape = \zeta)$; $Var(s_g^n) = 1/\zeta$
  - $\zeta \sim \mathcal{E}(1)$;
    OR
  - $s_g^n \sim \Gamma(mean = 1, shape = \zeta^{k_g^n})$; $Var(s_g^n) = 1/\zeta^{k_g^n}$
  - $\zeta^{k_g^n} \sim \mathcal{E}(1)$;
- a^n: scaling of sampled molecules (fragment amplification and sampling)
  - $a^n \sim \Gamma(mean = \nu, shape = \psi)$; $E(p^n) = \eta$; $Var(p^n) = \nu^2/\psi$
  - $\nu \sim \mathcal{E}(1)$;
  - $\psi \sim \mathcal{E}(1)$;
- u_g^k expression of gene g in cluster k
  - $u_g^k \sim \Gamma(mean = m_g, shape = \beta)$; $scale = \frac{m_g}{\beta}$, $E(u_g^k) = m_g$; $var(u_g^k) = \frac{m_g^2}{\beta}$;
  - alternative writing: $u_g^k \sim m_g \cdot h_g^k$ where $h_g^k \sim \Gamma(mean = 1, shape = \beta)$
  - m_g mean expression for gene g: $m_g \sim \Gamma(mean = \mu, shape = \alpha)$

    * $\mu = 5$;
    * $\alpha \sim \mathcal{E}(1)$;
  - β shape of u_g^k distribution: $\beta \sim \mathcal{E}(1)$

- z_g^n number of cDNA full molecules for gene g and cell n
  - $z_g^n \mid u_g^{1:K}, k_g^n, p^n, s_g^n \sim \mathcal{P}(z_g^n; p^n \cdot u_g^{k_g^n} \cdot s_g^n)$
  - $E(z_g^n) = p^n \cdot u_g^{k_g^n}$; $Var(z_g^n \mid p^n \cdot u_g^{k_g^n}) = p^n \cdot u_g^{k_g^n} + (p^n \cdot u_g^{k_g^n})^2 \cdot \frac{1}{\zeta^{k_g^n}}$;

- y_g^n number of reads for g in cell n

$$y_g^n \mid z_g^n, a^n \quad \sim \quad \cdot\mathcal{NB}(y_g^n; mean = w_g \cdot z_g^n \cdot a^n, overdispersion = \kappa^n + \frac{\tilde{\kappa}_w}{w_g} + \frac{\kappa*}{z_g^n})$$

  where w_g is the unit size (length of gene).

- κ technical over-dispersion in read counts;
- d_g indicates if the expression of g (u_g^k) is the same among clusters: d_g is 0 if the g is expressed the same in all clusters and 1 if it expressed differently . If $d_g = 0$, the gene g will not be taken into account when building the clusters.

2 Gibbs algorithm

1. update u_g^k
   (a) $\pi(u_g^k \mid k_g^{1:N}, z_g^{1:N}, p^{1:N}, d_g, s_g^{1:N})$ - GDU (Gibbs direct update)
      - prior $\Gamma(shape = \beta, scale = \frac{m_g}{\beta})$
      - $\pi(u_g^k \mid k_g^{1:N}, z_g^{1:N}, p^{1:N}, s_g^{1:N}, d_g = 1) \propto \prod_n \mathcal{P}(z_g^n; u_g^{k^n} \cdot p^n \cdot s_g^n)^{\mathbb{1}_{k_g^n=k}} \cdot \Gamma(u_g^k; \beta, \frac{m_g}{\beta})$
        $\rightarrow$ GDU from posterior $\Gamma(shape = \beta + \sum_n \mathbb{1}_{k=k^n} z_g^n, scale = \frac{1}{\sum_n \mathbb{1}_{k=k^n} \cdot p^n \cdot s_g^n + \beta/m_g})$
      - $\pi(u_g \mid k^{1:N}, z_g^{1:N}, p^{1:N}, s_g^{1:N}, d_g = 0) \propto \prod_n \mathcal{P}(z_g^n; u_g \cdot p^n \cdot s_g^n) \cdot \Gamma(u_g\ \beta, \frac{m_g}{\beta})$
        $\rightarrow$ GDU from posterior $\Gamma(shape = \beta + \sum_n z_g^n, scale = \frac{1}{\sum_n p^n \cdot s_g^n + \beta/m_g})$. Set $u_g^{k^n} = u_g$ for $k^n = 1 : K$
   (b) $\pi(\beta \mid u_{1:G}^{1:K}, d_{1:G}) \propto \pi(u_{1:G}^{1:K} \mid \beta) \cdot \pi(\beta)$ - MH update
      posterior density $\propto \prod_{kg} (u_g^k)^{\beta-1} \cdot e^{-\sum_{kg} u_g^k \cdot \beta/m_g} \cdot \frac{(\beta/m_g)^{\sum_{kg} \beta}}{\Gamma(\beta)^{kg}} \cdot \pi(\beta)$
   (c) $\pi(m_g \mid u_g^{1:K}) \propto \pi(u_g^{1:K} \mid m_g) \cdot \pi(m_g)$
      posterior density $\propto e^{-\sum_k u_g^k \cdot \beta/m_g} \cdot m_g^{-\sum_k \beta} \cdot \Gamma(m_g; \alpha, \theta)$
   (d) $\mu = 5$
   (e) $\pi(\alpha \mid m_{1:G}, \mu) \propto \prod_g (m_g)^{\alpha-1} \cdot e^{-\sum_g m_g \cdot \alpha/\mu} \cdot \frac{(\alpha/\mu)^{G\alpha}}{\Gamma(\alpha)^G} \cdot e^{-\alpha}$ MH update
2. update p^n
   (a) $\pi(p^n \mid z_{1:G}^n, u_{1:G}^{k^n}) \propto \prod_g \mathcal{P}(z_g^n; u_g^{k^n} \cdot p^n \cdot s_g^n) \cdot \Gamma(p^n; mean = \eta, shape = \chi)$:
      - prior $\Gamma(p^n; mean = \eta, shape = \chi)$;
      - GDU, posterior $\Gamma(shape = \chi + \sum_g z_g^n, scale = \frac{1}{\sum_g u_g^{k^n} \cdot s_g^n + \chi/\eta})$
   (b) $\pi(\eta \mid p^{1:N}, \chi) \propto e^{-\sum_n p^n \cdot \chi/\eta} \cdot \eta^{-N \cdot \chi} \cdot e^{-\eta}$ MH update
   (c) $\pi(\chi \mid p^{1:N}, \eta) \propto \prod_n (p^n)^{\chi-1} \cdot e^{-\sum_n p^n \cdot \chi/\eta} \cdot \frac{(\chi/\eta)^{N\chi}}{\Gamma(\chi)^N} \cdot e^{-\chi}$ MH update
3. update s_g^n
   (a) $\pi(s_g^n \mid z_{1:G}^n, u_{1:G}^{k^n}, p^n) \propto \mathcal{P}(z_g^n; u_g^{k^n} \cdot p^n \cdot s_g^n) \cdot \Gamma(s_g^n; mean = 1, shape = \zeta)$:
      - prior $\Gamma(s_g^n; mean = 1, shape = \zeta)$;
      - GDU, posterior $\Gamma(shape = \zeta + z_g^n, scale = \frac{1}{u_g^{k^n} p^n + \zeta})$
      - $\pi(\zeta \mid s_{1:G}^{1:N}) \propto \prod_g \prod_n (s_g^n)^{\zeta-1} \cdot e^{-\sum_n \sum_g s_g^n \cdot \zeta} \cdot \frac{(\zeta)^{NG\zeta}}{\Gamma(\zeta)^{NG}} \cdot e^{-\zeta}$ MH update

      OR

      - prior $\Gamma(s_g^n; mean = 1, shape = \zeta^{k^n})$;
      - GDU, posterior $\Gamma(shape = \zeta^{k^n} + z_g^n, scale = \frac{1}{u_g^{k^n} p^n + \zeta^{k^n}})$
      - $\pi(\zeta^k \mid s_{1:G}^{1:N}) \propto \prod_g \prod_{n|k^n=k} (s_g^n)^{\zeta^k-1} \cdot e^{-\sum_n \sum_g s_g^n \cdot \zeta^k} \cdot \frac{(\zeta^k)^{N^k G\zeta}}{\Gamma(\zeta^k)^{N^k G}} \cdot e^{-\zeta^k}$ MH update
4. update a^n
   (a) $P(a^n \mid z_{1:G}^n, y_{1:G}^{1:N}) \propto \prod_g \mathcal{NB}(y_g^n; a^n \cdot b_g \cdot z_g^n \cdot w, overdis = \underline{\kappa^n}) \cdot \Gamma(a^n; mean = \nu, shape = \psi)$
      - prior $\Gamma(a^n; mean = \nu, shape = \psi)$;
      - MH update
   (b) $\pi(\nu \mid a^{1:N}, \psi) \propto e^{-\sum_n a^n \cdot \psi/\nu} \cdot \nu^{-N \cdot \psi} \cdot e^{-\nu}$ MH update
   (c) $\pi(\psi \mid a^{1:N}, \nu) \propto \prod_n (a^n)^{\psi-1} \cdot e^{-\sum_n a^n \cdot \psi/\nu} \cdot \frac{(\psi/\nu)^{N\psi}}{\Gamma(\psi)^N} \cdot e^{-\psi}$ MH update
5. update k_g^n
   (a) $P(k_g^n = k \mid u_g^{1:K}, y_g^n, z_g^{1:N}, d_g, \gamma)$
      - $P(k_g^n = k \mid u_g^{1:K}, z_g^n, d_g) \propto \mathcal{P}(z_g^n; u_g^{k_g} \cdot p^n \cdot s_g^n) \cdot P(k_g^n = k \mid \gamma^n)$
      - GDU from a categorical distribution $\big(k, \mathcal{P}(z_g^n; u_g^{k_g} \cdot p^n \cdot s_g^n) \cdot \gamma_k^n\big)_{1:K}$,
   (b) hyperpar. $\gamma^n \mid k_{1:G}^n \sim Dir(\sum_g 1_{k_g^n=k} + 1)$; prior $\gamma^n \sim Dir(1 \times K)$

6. update $k_g \mid \gamma^{1:N}, k_g$ reshuffle the cluster labels
   For each $g = 1 : G$:
   (a) sample clusters k_1 and k_2 (uniform)
   (b) compute the weight of each cluster in cells that are assigned for one of the two clusters: $n_v^l = \prod_n (\gamma_l^n)^{1_{k_g^n=v}}$ for $v, l \in \{k_1, k_2\}$
   (c) re-sample labels $(k_1, k_2) \sim \left(n_{k_2}^{k_1} \cdot n_{k_1}^{k_2}, n_{k_2}^{k_2} \cdot n_{k_1}^{k_1}\right)$

   Recompute γ.

   Or more systematic shuffling of cluster names.
   For gene $g = 1 \ldots G$, select randomly k_1 and draw k_1' such as $k_1 \leftarrow k_1'$ and $k_1' \leftarrow k_1$ from the GDU pdf:

$$\mathbb{P}(k_1' = k) \quad \propto \quad \prod_n (\gamma_k^n / \gamma_{k_1}^n)^{\mathbb{I}\{k_g^n = k_1\}} (\gamma_{k_1}^n / \gamma_k^n)^{\mathbb{I}\{k_g^n = k\}}$$

   if the model involves ζ^k this equation becomes

$$\mathbb{P}(k_1' = k) \quad \propto \quad \prod_n \left(\frac{\gamma_k^n \Gamma(s_g^n; mean = 1, shape = \zeta^k)}{\gamma_{k_1}^n \Gamma(s_g^n; mean = 1, shape = \zeta^{k_1})}\right)^{\mathbb{I}\{k_g^n = k_1\}} \left(\frac{\gamma_{k_1}^n \Gamma(s_g^n; mean = 1, shape = \zeta^{k_1})}{\gamma_k^n \Gamma(s_g^n; mean = 1, shape = \zeta^k)}\right)^{\mathbb{I}\{k_g^n = k\}}$$

   Warning: with the proposed extension for versatile sharing of u_g^k between clusters the symmetry that allowed this simple move is broken. A slightly more complicated move and acceptance probability is used (see below).

7. update z_g^n
   - $P(z_g^n \mid y_g^n, u_g^{k^n}, p^n, s_g^n) \propto \mathcal{NB}(y_g^n; z_g^n \cdot a^n \cdot w_g, \kappa^n + \frac{\tilde{\kappa}_w}{w_g} + \frac{\kappa *}{z_g^n}) \cdot \mathcal{P}(z_g^n; p^n \cdot s_g^n \cdot u_g^{k^n})$ - MH

8. update κ^n
   - $\pi(\kappa \mid z_{1:G}^n, y_g^n) \propto \prod_g \mathcal{NB}(y_g^n; a^n \cdot z_g^n \cdot w, overdis = \kappa^n + \frac{\tilde{\kappa}_w}{w_g} + \frac{\kappa *}{z_g^n}) \cdot \Gamma(\kappa^n, shape = pa, scale = pb)$ - MH update
   - same for $\tilde{\kappa}_w, \kappa *$

9. update d_g
   (a) $P(d_g \mid p^{1:N}, z_g^{1:N})$; integrate out μ_g
      - $\pi(d_g = 0 \mid z_g^{1:N}, p^{1:N}, s_g^{1:N}) = \Gamma(\beta + \sum_n z_g^n) \cdot \frac{1}{\sum_n p^n \cdot s_g^n + \beta / m_g}^{\beta + \sum_n z_g^n} \frac{1}{\Gamma(\beta) \cdot (m_g / \beta)^\beta} \cdot \delta \cdot ct_a$
      - $\pi(d_g = 1 \mid z_g^{1:N}, p^{1:N}, s_g^{1:N}) = \prod_{k=1:K} \Gamma(\beta + \sum_n \mathbb{1}_{k = k_g^n} z_g^n) \cdot \frac{1}{\sum_n \mathbb{1}_{k = k_g^n} p^n \cdot s_g^n + \beta / m_g}^{\beta + \sum_n \mathbb{1}_{k = k_g^n} z_g^n} \frac{1}{\Gamma(\beta) \cdot (m_g / \beta)^\beta} \cdot (1 - \delta) \cdot ct_a$
   (b) $\pi(\delta \mid d_{1:G}) \propto \pi(d_{1:G} \mid \delta) \cdot \pi(\delta)$
      GDU from posterior density $\delta \mid d_{1:G} \sim \beta(1 + \sum_g \mathbb{1}_{d_g = 0}, 1 + \sum_g \mathbb{1}_{d_g = 1})$; Beta prior $\beta(1, 1)$

10. compute likelihood $\pi(y \mid k, u, z, a, d) = \prod_g \prod_n \pi(y_g^n \mid z_g^n, a^n, \kappa)$

2.1 Extension for versatile sharing of u_g^k's between clusters

The model described above distinguishes only two cases for a given gene:

- if $d_g = 0$ each $(u_g^1, \ldots, u_g^K)$ are iid $u_g^k \sim \Gamma(mean = m_g, shape = \beta)$,
- otherwise $(d_g = 1)$ $u_g^1 = \ldots = u_g^K = u$ where $u \sim \Gamma(mean = m_g, shape = \beta)$.

Our goal is to propose an extension allowing more versatile sharing of u_g^k's values between clusters.
We introduce the random variables

- $i_g^k \in (1, \ldots, K)$ such as $i_g^k \sim Mult(1; \theta_1^k, \ldots, \theta_K^k)$
- and $v_g^k \in \mathbb{R}^+$ such as $v_g^k \sim \Gamma(mean = m_g, shape = \beta)$.

and we set $u_g^k = v_g^{i_g^k}$, i.e. $u_g^k = v_g^q$ with probability θ_q^k.

This modeling makes it possible to account for complex relationships between the clusters. In particular $\mathbb{P}(u_g^k = u_g^{k'}) = \sum_q \theta_q^k \theta_q^{k'}$. Denoting $S_{k,k'} = \mathbb{P}(u_g^k = u_g^{k'})$, $S = (S_{k,k'})_{(k,k')\in(1,\ldots,K)^2}$, and $\Theta = (\theta_{k,q})_{(k,q)\in(1,\ldots,K)^2}$ we have $S = \Theta\Theta^T$. We can also easily write the probability for all u_g^k being equal $\mathbb{P}(u_g^1 = \ldots = u_g^K) = \sum_q \prod_k \theta_q^k$.
MCMC moves:

- Update of $(\theta_1^k, \ldots, \theta_K^k)$ for $k = 1 \ldots K$. We assume a Dirichlet prior on the θ^k's.

$$(\theta_1^k, \ldots, \theta_K^k) \mid \ldots \quad \sim \quad Dir(1 + \sum_g \mathbb{I}\{i_g^k = 1\}, \ldots, 1 + \sum_g \mathbb{I}\{i_g^k = K\})\,.$$

- Update of v_g^k for $k = 1 \ldots K$ and $g = 1 \ldots G$. The new v_g^k are propagated to u_g^k.

$$v_g^k \mid \ldots \quad \sim \quad \Gamma(mean = \frac{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = k\} z_g^n}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = k\} p_n s_g^n}, shape = \beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = k\} z_g^n)$$

scale $\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k^n g} = k\} p_n s_g^n}$

- Update of i_g^k

$$\mathbb{P}(i_g^1, \ldots, i_g^K \mid \ldots) \quad \propto \quad \prod_k \theta_{i_g^k}^k \times \prod_q \Gamma(\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} z_g^n) \left[\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} p_n s_g^n}\right]^{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} z_g^n}$$

We use the notations $tt1 = \sum_n \mathbb{I}\{k_g^n = k\} z_g^n$ and $tt2 = \sum_n \mathbb{I}\{k_g^n = k\} p_n s_g^n$. The first type of move would consist in trying a new value $\tilde{i_g^k}$ for i_g^k. MH move with acceptation probability

$$\alpha \quad = \quad \frac{\theta_{\tilde{i_g^k}}^k}{\theta_{i_g^k}^k} \times \frac{\Gamma(\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = i_g^k\} z_g^n - tt1) \left[\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} p_n s_g^n - tt2}\right]^{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} z_g^n - tt1}}{\Gamma(\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = i_g^k\} z_g^n) \left[\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} p_n s_g^n}\right]^{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = q\} z_g^n}}$$
$$\times \frac{\Gamma(\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} z_g^n + tt1) \left[\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} p_n s_g^n + tt2}\right]^{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} z_g^n + tt1}}{\Gamma(\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} z_g^n) \left[\frac{1}{\beta/m_g + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} p_n s_g^n}\right]^{\beta + \sum_n \mathbb{I}\{i_g^{k_g^n} = \tilde{i_g^k}\} z_g^n}}$$

following the same strategy we could do a GDU of the value of i_g^k. The $v_g^{i_g^k}$ and $v_g^{\tilde{i_g^k}}$ need to redrawn after an accepted move of this type.

Another interesting update could consist in randomly picking q_1 and q_2 and then to set $\tilde{i_g^k} = q_2$ when $i_g^k = q_1$ and $\tilde{i_g^k} = q_1$ when $i_g^k = q_2$. After accepting a move of this type Tthe v_g^q's need to swapped accordingly: $v_g^{\tilde{q_1}} = v_g^{q_2}$ and $v_g^{\tilde{q_2}} = v_g^{q_1}$. This corresponding MH move has acceptation probability

$$\alpha \quad = \quad \frac{\prod_k (\theta_{q_2}^k)^{\mathbb{I}\{i_g^k = q_1\}} (\theta_{q_1}^k)^{\mathbb{I}\{i_g^k = q_2\}}}{\prod_k (\theta_{q_1}^k)^{\mathbb{I}\{i_g^k = q_1\}} (\theta_{q_2}^k)^{\mathbb{I}\{i_g^k = q_2\}}}\,.$$

The update move aiming at shuffling the labels of the clusters (k_g^n for $n = 1 : N$) associated with a given gene g could be implemented as follows. Randomly select a k_1 and draw k_1' such as the labels are swapped $k_1 \leftrightarrow k_1'$, meaning that cells in $k_g^n = k_1 \leftarrow k_g^n = k_1'$ and $k_g^n = k_1' \leftarrow k_g^n = k_1$. Simultaneously we swap values $u_g^{k_1} \leftrightarrow u_g^{k_1'}$ and $i_g^{k_1} \leftrightarrow i_g^{k_1'}$ but the underlying v_g^q's for $q = 1 \ldots K$ remain unchanged.

$$\mathbb{P}(k_1' = k) \quad \propto \quad \left[\prod_n \left(\frac{\gamma_k^n}{\gamma_{k_1}^n}\right)^{\mathbb{I}\{k_g^n = k_1\}} \left(\frac{\gamma_{k_1}^n}{\gamma_k^n}\right)^{\mathbb{I}\{k_g^n = k\}}\right] \times \frac{\theta_{i_g^k}^{k_1} \theta_{i_g^{k_1}}^k}{\theta_{i_g^{k_1}}^{k_1} \theta_{i_g^k}^k}\,.$$

When the model involves ζ^k, the equation becomes

$$\mathbb{P}(k_1' = k) \quad \propto \quad \left[\prod_n \Big(\frac{\gamma_k^n \Gamma(s_g^n; mean = 1, shape = \zeta^k)}{\gamma_{k_1}^n \Gamma(s_g^n; mean = 1, shape = \zeta^{k_1})} \Big)^{\mathbb{I}\{k_g^n = k_1\}} \Big(\frac{\gamma_{k_1}^n \Gamma(s_g^n; mean = 1, shape = \zeta^{k_1})}{\gamma_k^n \Gamma(s_g^n; mean = 1, shape = \zeta^k)} \Big)^{\mathbb{I}\{k_g^n = k\}} \right]$$
$$\times \frac{\theta^{k_1}_{i_g^k} \theta^{k}_{i_g^{k_1}}}{\theta^{k_1}_{i_g^{k_1}} \theta^{k}_{i_g^k}} .$$