Full conditional distributions for the dependent Dirichlet process mixture model

I detail here how to arrive at the full conditional distributions for the dependent Dirichlet process mixture of normal distributions model.

Some facts that we will need are:

- Continuous version of Bayes' theorem

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}.$$

  The term in the denominator, normalising constant, does not depend on $\boldsymbol{\theta}$ and so we can write,

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta}),$$

  i.e.,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

- A conjugate prior leads to a posterior from the same parametric family as the prior. There are long lists of conjugate priors, see e.g.,

  `https://en.wikipedia.org/wiki/Conjugate_prior`

  Conjugate priors are used often for computational convenience because the posterior has a closed form. In fancier models, conjugate priors facilitate Gibbs sampling which is the easiest Bayesian computational algorithm.

Let us start with the simplest case where we have normally distributed data and where both the mean μ and variance σ^2 (or precision σ^{-2}) are unknown. For $\boldsymbol{\theta} = (\mu, \sigma^{-2})$, the likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} (\sigma^{-2})^{1/2} \exp\left\{ -\frac{1}{2} \sigma^{-2} (y_i - \mu)^2 \right\} \right] \\ &= (2\pi)^{-n/2} (\sigma^{-2})^{n/2} \exp\left\{ -\frac{1}{2} \sigma^{-2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \end{aligned}$$

We need to specify the prior $p(\mu, \sigma^{-2})$. Assuming a priori independence, we have $p(\mu, \sigma^{-2}) = p(\mu) p(\sigma^{-2})$. We further assume that $\mu \sim \mathrm{N}(m, S^2)$ and $\sigma^{-2} \sim \mathrm{Gamma}(a, b)$ (the parameterisation for which the mean is a/b). The joint posterior distribution is

$$\begin{aligned} p(\mu, \sigma^{-2} \mid \mathbf{y}) \propto{}& (2\pi)^{-n/2} (\sigma^{-2})^{n/2} \exp\left\{ -\frac{1}{2} \sigma^{-2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &\times \frac{1}{\sqrt{2\pi S^2}} \exp\left\{ -\frac{1}{2S^2} (\mu - m)^2 \right\} \\ &\times \frac{b^a}{\Gamma(a)} (\sigma^{-2})^{a-1} \exp\left\{ -b\sigma^{-2} \right\}. \end{aligned}$$

The joint posterior does not have a recognisable form but the full conditional distributions do. We have

$$\begin{aligned}
p(\mu \mid \sigma^{-2}, \mathbf{y}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\exp\left\{-\frac{1}{2S^2}(\mu-m)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\mu+\mu^2)\right\}\exp\left\{-\frac{1}{2S^2}(\mu^2-2\mu m+m^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(-2\mu n\bar{y}+n\mu^2)-\frac{1}{2S^2}(\mu^2-2\mu m)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{-2\mu n\bar{y}S^2+n\mu^2S^2+\mu^2\sigma^2-2\mu m\sigma^2}{\sigma^2S^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2(nS^2+\sigma^2)-2\mu(n\bar{y}S^2+m\sigma^2)}{\sigma^2S^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2-2\mu(n\bar{y}S^2+m\sigma^2)/(nS^2+\sigma^2)}{\sigma^2S^2/(nS^2+\sigma^2)}\right)\right\}.
\end{aligned}$$

Let

$$m^* = \frac{n\bar{y}S^2+m\sigma^2}{nS^2+\sigma^2}, \quad \text{and} \quad S^* = \frac{\sigma^2S^2}{nS^2+\sigma^2}.$$

Replacing,

$$\begin{aligned}
p(\mu \mid \sigma^{-2}, \mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2-2\mu m^*}{S^*}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2-2\mu m^*+(m^*)^2-(m^*)^2}{S^*}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{(\mu-m^*)^2-(m^*)^2}{S^*}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{(\mu-m^*)^2}{S^*}\right)\right\},
\end{aligned}$$

which we recognise as the kernel of a normal distribution with mean m^* and variance S^*. That is

$$\mu \mid \sigma^2, \mathbf{y} \sim \mathrm{N}\left(\frac{n\bar{y}S^2+m\sigma^2}{nS^2+\sigma^2}, \frac{\sigma^2S^2}{nS^2+\sigma^2}\right),$$

or, equivalently, dividing everything by σ^2S^2,

$$\mu \mid \sigma^2, \mathbf{y} \sim \mathrm{N}\left(\frac{\frac{m}{S^2}+\frac{n\bar{y}}{\sigma^2}}{\frac{1}{S^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{S^2}+\frac{n}{\sigma^2}}\right).$$

In a similar way,

$$\begin{aligned}
p(\sigma^{-2} \mid \mu, \mathbf{y}) &\propto (\sigma^{-2})^{n/2}\exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}(\sigma^{-2})^{a-1}\exp\left\{-b\sigma^{-2}\right\} \\
&= (\sigma^{-2})^{a+n/2-1}\exp\left\{-\sigma^{-2}\left(b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)\right\},
\end{aligned}$$

which we recognise as a kernel of a gamma distribution with shape parameter $a + n/2$ and rate parameter $b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$, i.e.,

$$\sigma^{-2} \mid \mu, \mathbf{y} \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

Let us now move to the case of Dirichlet process mixtures (DPM) of normals, which is a fancy name for infinite mixture models. To motivate DPMs, we start with a formulation based on finite mixtures of normal distributions, which are known to approximate any continuous smooth distribution (Lo ,1984). A finite normal mixture model posits

$$f(y) = \sum_{k=1}^{K} p_k \phi(y \mid \mu_k, \sigma_k^{-2}), \tag{1}$$

Under this framework, each observation arises from one of K mixture components, which each have a specific mean and variance/precision. The model in (1) can be equivalently written as

$$f(y) = \int \phi(y \mid \mu, \sigma^{-2}) \mathrm{d}G(\mu, \sigma^{-2}),$$

where G is a discrete (mixing) distribution given by

$$G(\cdot) = \sum_{k=1}^{K} p_k \delta_{(\mu_k, \sigma_k^{-2})}(\cdot), \tag{2}$$

and where δ_a denotes a point mass at a. To proceed with Bayesian inference, prior distributions are required for the weights, means, and variances. The vector of weights $(p_1, \ldots, p_K)$ is often assigned a Dirichlet prior distribution and, leveraging conjugacy properties, the prior for (μ_k, σ_k^{-2}), say $G_0(\mu, \sigma^{-2})$, often is a normal-gamma distribution. Placing a prior distribution on $\{(p_k, \mu_k, \sigma_k^{-2}) \mid k = 1, \ldots, K\}$ is equivalent to placing a prior distribution on the mixing distribution G. Finite mixture modelling provides a very flexible framework for density estimation. However, choosing the number of mixture components K is not a trivial task in general. It is common to use the value that optimises a selection criterion across candidate models with different numbers of components. Another approach is to place a prior on K, but this can be computationally difficult to implement efficiently in practice (e.g., involving reversible jump MCMC).

A powerful alternative involves a Bayesian nonparametric Dirichlet process (Ferguson, 1973) prior for G, which induces a DPM of normal distributions on f. In addition to the theoretical advantage of the Dirichlet process (DP) mixture of normal model having full support on the space of absolutely continuous distributions (Lo, 1984), the DP prior has the practical advantage of automatically determining the number of components that best fits a given data set. We write $G \sim \text{DP}(\alpha, G_0)$ to indicate that G follows a DP prior. The prior mean $E(G) = G_0$ is a parametric base/centering distribution. The positive precision parameter α determines, among other important characteristics, the variation of G around the prior mean G_0, with smaller (larger) values of α implying higher (lower) uncertainty. According to Sethuraman (1994), the DP prior can be represented as

$$G(\cdot) = \sum_{k=1}^{\infty} p_k \delta_{(\mu_k, \sigma_k^{-2})}(\cdot), \tag{3}$$

where $(\mu_k, \sigma_k^{-2}) \overset{\text{iid}}{\sim} G_0$ and $p_1 = v_1$, $p_k = v_k \prod_{m=1}^{k-1}(1 - v_m)$ for $k \geq 2$, with $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, independently of $\{(\mu_k, \sigma_k^{-2}) \mid k \geq 1\}$. Notice that under Sethuraman's representation, uncertainty about each weight parameter

p_k is induced from uncertainty about the v_k's.
The proposed model for the density dunction is thus given by a DPM of normals, namely

$$f(y) = \int \phi(y \mid \mu, \sigma^{-2}) \mathrm{d}G(\mu, \sigma^{-2}), \quad G \sim \mathrm{DP}(\alpha, G_0). \tag{4}$$

Due to conjugacy properties, we take $G_0(\mu, \sigma^{-2}) = \mathrm{N}(\mu \mid m, S^2)\mathrm{Gamma}(\sigma^{-2} \mid a, b)$. For ease of posterior simulation and because it provides a highly accurate approximation, we used a truncated version of the DPM model (Ishwaran and James, 2001). Specifically, the mixing distribution G in (3) is replaced with

$$G^L(\cdot) = \sum_{k=1}^{L} p_k \delta_{(\mu_k, \sigma_k^{-2})}(\cdot),$$

with L being pre-specified and where the p_k's result from a truncated version of the stick-breaking construction: $p_1 = v_1$, $p_k = v_k \prod_{m=1}^{k-1}(1 - v_m)$ for $k = 2, \ldots, L$, $v_1, \ldots, v_{L-1} \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$, and $v_L = 1$. An appropriate value of L can be determined by considering the properties of the high order weight values in the infinite sum representation (3). For instance, $E\left(\sum_{k=L+1}^{\infty} p_k \mid \alpha\right) = \alpha^L(1 + \alpha)^{-L}$. The truncated DPM model can be expressed as

$$f(y) = \sum_{k=1}^{L} p_k \phi(y \mid \mu_k, \sigma_k^{-2}),$$

where $\mu_k \overset{\text{iid}}{\sim} \mathrm{N}(m, S^2)$ and $\sigma_k^{-2} \overset{\text{iid}}{\sim} \mathrm{Gamma}(a, b)$ for $k = 1, \ldots, L$, and the weights p_k arise from the truncated stick-breaking representation described above. Note that L is not the exact number of components we expect to observe but instead an upper bound on the number of components.
The likelihood of the data is then

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{L} p_k \phi(y_i \mid \mu_k, \sigma_k^{-2}) \right\}, \quad \boldsymbol{\theta} = (v_1, \ldots, v_{L-1}, \mu_1, \ldots, \mu_L, \sigma_1^{-2}, \ldots, \sigma_L^{-2}).$$

which is not analytically tractable. However, this issue can be easily overcomed by data augmentation. To this end, consider the latent variable $z_i \in \{1, \ldots, L\}$ with $z_i = k$ indicating that observation y_i comes from component k, i.e., from the normal component with mean μ_k and precision σ_k^{-2}. We can then re-write the likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) &= \prod_{i=1}^{n} p_{z_i} \phi(y_i \mid \mu_{z_i}, \sigma_{z_i}^{-2}) \\ &= \prod_{k=1}^{L} \prod_{i: z_i = k} p_k \phi(y_i \mid \mu_k, \sigma_k^{-2}) \\ &= \prod_{k=1}^{L} p_k^{n_k} \prod_{i: z_i = k} \phi(y_i \mid \mu_k, \sigma_k^{-2}), \end{aligned}$$

where $n_k = \sum_{i=1}^{n} I(z_i = k)$. The joint posterior is

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y}) \propto & \prod_{k=1}^{L} \left(v_k \prod_{m<k} (1 - v_m) \right)^{n_k} \prod_{i:z_i=k} \phi(y_i \mid \mu_k, \sigma_k^{-2}) \\ & \times \prod_{k=1}^{L} \left[\frac{1}{\sqrt{2\pi\ S^2}} \exp\left\{ -\frac{1}{2S^2} (\mu_k - m)^2 \right\} \right] \\ & \times \prod_{k=1}^{L} \left[\frac{b^a}{\Gamma(a)} (\sigma_k^{-2})^{a-1} \exp\{-b\sigma_k^{-2}\} \right] \\ & \times \prod_{k=1}^{L-1} \left[v_k^{1-1} (1 - v_k)^{\alpha-1} \right]. \end{aligned}$$

From similar arguments to the ones seen above, we have that

$$\begin{aligned} \mu_k \mid \text{else} \sim & \text{N} \left(\frac{\frac{m}{S^2} + \frac{n\bar{y}_k}{\sigma_k^2}}{\frac{1}{S^2} + \frac{n_k}{\sigma_k^2}}, \frac{1}{\frac{1}{S^2} + \frac{n_k}{\sigma_k^2}} \right), \qquad k = 1, \ldots, L, \\ \sigma_k^{-2} \mid \text{else} \sim & \text{Gamma} \left(a + \frac{n_k}{2}, b + \frac{1}{2} \sum_{i:z_i=k} (y_i - \mu_k)^2 \right), \qquad k = 1, \ldots, L, \end{aligned}$$

where $\bar{y}_k$ is the mean of observations in component k, i.e., $\bar{y}_k = \sum_{i:z_i=k} y_i / n_k$ Now, let us look at the v_ks,

$$\begin{aligned} p(v_1 \mid \text{else}) & \propto v_1^{n_1} \times (1 - v_1)^{n_2} \times \ldots \times (1 - v_1)^{n_L} \times (1 - v_1)^{\alpha-1} \\ & = v_1^{n_1} (1 - v_1)^{\alpha + \sum_{k=2}^{L} n_k - 1}, \end{aligned}$$

which we recognise as the kernel of a beta distribution with parameters $n_1 + 1$ and $\alpha + \sum_{k=2}^{L} n_k$, i.e.,

$$v_1 \mid \text{else} \sim \text{Beta} \left(n_1 + 1, \alpha + \sum_{k=2}^{L} n_k \right).$$

We can generalise and state that

$$v_k \mid \text{else} \sim \text{Beta} \left(n_k + 1, \alpha + \sum_{l=k+1}^{L} n_l \right), \quad k = 1, \ldots, L - 1.$$

Finally, note that

$$\Pr(z_i = k) \propto \ p_k \phi(y_i \mid \mu_k, \sigma_k^{-2}),$$

which implies that

$$z_i \mid \text{else} \sim \text{Multinomial}(p_{i1}, \ldots, p_{iL}), \quad p_{ik} = \Pr(z_i = k \mid \text{else}) = \frac{p_k \phi(y_i \mid \mu_k, \sigma_k^{-2})}{\sum_{l=1}^{L} p_l \phi(y_i \mid \mu_l, \sigma_l^{-2})}.$$

Now let us consider our case, the linear dependent Dirichlet process mixture of normals (de Iorio, 2009), which is given by

$$f(y \mid \mathbf{x}) = \sum_{k=1}^{L} p_k \phi(y \mid \mathbf{x}^T \boldsymbol{\beta}_k, \sigma_k^{-2}),$$

where the weights p_ks arise from a truncated stick breaking representation and $(\boldsymbol{\beta}_k, \sigma_k^{-2}) \overset{\text{iid}}{\sim} G_0 \equiv N_p(\boldsymbol{\beta}_k \mid \mathbf{m}_{\boldsymbol{\beta}}, \mathbf{S}_{\boldsymbol{\beta}})\text{Gamma}(\sigma_k^{-2} \mid a, b)$. In order to allow for some more flexibility, we let $\mathbf{m}_{\boldsymbol{\beta}} \sim \text{N}_p(\mathbf{m}_0, \mathbf{S}_0)$ and $\mathbf{S}_{\boldsymbol{\beta}}^{-1} \sim \text{Wishart}(\nu, (\nu\boldsymbol{\psi})^{-1})$. Using again the latent variables trick, we have

$$\begin{aligned}
p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y}) \propto & \prod_{k=1}^{L} \left(v_k \prod_{m<k} (1 - v_k) \right)^{n_k} \prod_{i:z_i=k} (\sigma_k^{-2})^{1/2} \exp\left\{ -\frac{1}{2} \sigma_k^{-2} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2 \right\} \\
& \times \prod_{k=1}^{L} |\mathbf{S}_{\boldsymbol{\beta}}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) \right\} \\
& \times \prod_{k=1}^{L} \left[(\sigma_k^{-2})^{a-1} \exp\{-b\sigma_k^{-2}\} \right] \\
& \times \prod_{k=1}^{L-1} \left[v_k^{1-1} (1 - v_k)^{\alpha-1} \right] \\
& \times \exp\left\{ -\frac{1}{2} (\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0) \right\} \\
& \times |\mathbf{S}_{\boldsymbol{\beta}}^{-1}|^{(\nu-p-1)/2} \exp\left\{ -\frac{1}{2} \text{tr}(\nu\boldsymbol{\psi}\mathbf{S}_{\boldsymbol{\beta}}^{-1}) \right\},
\end{aligned}$$

where $\boldsymbol{\theta} = (v_1, \ldots, v_{L-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L, \sigma_1^{-2}, \ldots, \sigma_L^{-2}, \mathbf{m}_{\boldsymbol{\beta}}, \mathbf{S}_{\boldsymbol{\beta}})$.

The most 'challenging' to obtain full conditional distribution is $p(\boldsymbol{\beta}_k \mid \text{else})$. Let $y_{(k)} = \{y_i : i = k\}$ and let $\mathbf{x}_{(k)}$ be a $p \times k$ matrix whose columns are the $\mathbf{x}_i$ whose $i = k$.

$$\begin{aligned}
p(\boldsymbol{\beta}_k \mid \text{else}) &\propto \exp\left\{ -\frac{1}{2} \sigma_k^{-2} \sum_{i:z_i=k} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2 \right\} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\sigma_k^{-2} (y_{(k)} - \mathbf{x}_{(k)}^T \boldsymbol{\beta}_k)^T (y_{(k)} - \mathbf{x}_{(k)}^T \boldsymbol{\beta}_k) + (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) \right) \right\}
\end{aligned}$$

Now, working with the term in the exponential function

$$\begin{aligned}
& \sigma_k^{-2} (y_{(k)} - \mathbf{x}_{(k)}^T \boldsymbol{\beta})^T (y_{(k)} - \mathbf{x}_{(k)}^T \boldsymbol{\beta}) + (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) \\
&= \sigma_k^{-2} y_{(k)}^T y_{(k)} - \sigma_k^{-2} y_{(k)}^T \mathbf{x}_{(k)}^T \boldsymbol{\beta}_k - \sigma_k^{-2} \boldsymbol{\beta}_k^T \mathbf{x}_{(k)} y_{(k)} + \sigma_k^{-2} \boldsymbol{\beta}_k^T \mathbf{x}_{(k)} \mathbf{x}_{(k)}^T \boldsymbol{\beta}_k \\
&\quad + \boldsymbol{\beta}_k^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_{\boldsymbol{\beta}}^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}_k + \mathbf{m}_{\boldsymbol{\beta}}^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \mathbf{m}_{\boldsymbol{\beta}} \\
&= \boldsymbol{\beta}_k^T \left(\sigma_k^{-2} \mathbf{x}_{(k)} \mathbf{x}_{(k)}^T + \mathbf{S}_{\boldsymbol{\beta}}^{-1} \right) \boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T \left(\sigma_k^{-2} \mathbf{x}_{(k)} y_{(k)} + \mathbf{S}_{\boldsymbol{\beta}}^{-1} \mathbf{m}_{\boldsymbol{\beta}} \right) - \left(\sigma_k^{-2} y_{(k)}^T \mathbf{x}_{(k)}^T + \mathbf{m}_{\boldsymbol{\beta}}^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \right) \boldsymbol{\beta}_k \\
&\quad + \sigma_k^{-2} y_{(k)}^T y_{(k)} + \mathbf{m}_{\boldsymbol{\beta}}^T \mathbf{S}_{\boldsymbol{\beta}}^{-1} \mathbf{m}_{\boldsymbol{\beta}}
\end{aligned}$$

Let

$$\mathbf{V}_1 = \left(\sigma_k^{-2}\mathbf{x}_{(k)}\mathbf{x}_{(k)}^T + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_1 = \mathbf{V}_1\left(\sigma_k^{-2}\mathbf{x}_{(k)}y_{(k)} + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}\right)$$

Then, the last expression from above is equal to

$$\boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\beta} + \sigma_k^{-2}y_{(k)}^T y_{(k)} + \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}$$

Plugging in this into the full conditional distribution

$$\begin{aligned}
p(\boldsymbol{\beta}_k \mid \text{else}) &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\beta} + \sigma_k^{-2}y_{(k)}^T y_{(k)} + \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\beta}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\beta} + \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left((\boldsymbol{\beta}_k - \boldsymbol{\mu}_1)^T\mathbf{V}_1^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\mu}_1) - \boldsymbol{\mu}_1^T\mathbf{V}_1^{-1}\boldsymbol{\mu}_1\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left((\boldsymbol{\beta}_k - \boldsymbol{\mu}_1)^T\mathbf{V}_1^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\mu}_1)\right)\right\},
\end{aligned}$$

which we recognise as the kernel of a multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\mathbf{V}_1$. Then,

$$\boldsymbol{\beta}_k \mid \text{elseN}_p(\boldsymbol{\mu}_1, \mathbf{V}_1), \qquad k = 1, \ldots, L,$$

with $\mathbf{V}_1 = \left(\sigma_k^{-2}\mathbf{x}_{(k)}\mathbf{x}_{(k)}^T + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\right)^{-1} = \left(\sigma_k^{-2}\sum_{i:z_i=k}\mathbf{x}_i\mathbf{x}_i^T + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$, and

$$\boldsymbol{\mu}_1 = \mathbf{V}_1\left(\sigma_k^{-2}\mathbf{x}_{(k)}y_{(k)} + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}\right) = \mathbf{V}_1\left(\sigma_k^{-2}\sum_{i:z_i=k}\mathbf{x}_i y_i + \mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}\right)$$

We also have as before

$$\sigma_k^{-2} \mid \text{else} \sim \text{Gamma}\left(a + \frac{n_k}{2}, b + \frac{1}{2}\sum_{i:z_i=k}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2\right), \qquad k = 1, \ldots, L.$$

and

$$v_k \mid \text{else} \sim \text{Beta}\left(n_k + 1, \alpha + \sum_{l=k+1}^{L} n_l\right), \qquad k = 1, \ldots, L-1,$$

as well as

$$z_i \mid \text{else} \sim \text{Multinomial}(p_{i1}, \ldots, p_{iL}), \quad p_{ik} = \Pr(z_i = k \mid \text{else}) = \frac{p_k\phi(y_i \mid \mu_k, \sigma_k^{-2})}{\sum_{l=1}^{L} p_l\phi(y_i \mid \mu_l, \sigma_l^{-2})}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, L.$$

$$p(\mathbf{m}_{\boldsymbol{\beta}} \mid \text{else}) \propto \exp\left\{-\frac{1}{2}\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})\right\}\exp\left\{-\frac{1}{2}(\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)\right\}$$
$$= \exp\left\{-\frac{1}{2}\left(\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) + (\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)\right)\right\}.$$

Noting that

$$\begin{aligned}
&\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}}) + (\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_0)\\
&= \sum_{k=1}^{L}\boldsymbol{\beta}_k^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_k - \sum_{k=1}^{L}\boldsymbol{\beta}_k^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\sum_{k=1}^{L}\boldsymbol{\beta}_k + L\mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}\\
&\quad + \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_0^{-1}\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_{\boldsymbol{\beta}} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\\
&= \mathbf{m}_{\boldsymbol{\beta}}^T(L\mathbf{S}_{\boldsymbol{\beta}}^{-1} + \mathbf{S}_0^{-1})\mathbf{m}_{\boldsymbol{\beta}} - \mathbf{m}_{\boldsymbol{\beta}}^T\left(\mathbf{S}_{\boldsymbol{\beta}}^{-1}\sum_{k=1}^{L}\boldsymbol{\beta}_k + \mathbf{S}_0^{-1}\mathbf{m}_0\right) - \left(\sum_{k=1}^{L}\boldsymbol{\beta}_k^T\mathbf{S}_{\boldsymbol{\beta}}^{-1} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\right)\mathbf{m}_{\boldsymbol{\beta}} + \sum_{k=1}^{L}\boldsymbol{\beta}_k^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_k + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0
\end{aligned}$$

By letting $\mathbf{V} = (L\mathbf{S}_{\boldsymbol{\beta}}^{-1} + \mathbf{S}_0^{-1})^{-1}$ and $\boldsymbol{\mu} = \mathbf{V}\left(\mathbf{S}_{\boldsymbol{\beta}}^{-1}\sum_{k=1}^{L}\boldsymbol{\beta}_k + \mathbf{S}_0^{-1}\mathbf{m}_0\right)$, we have

$$p(\mathbf{m}_{\boldsymbol{\beta}} \mid \text{else}) \propto \exp\left\{-\frac{1}{2}(\mathbf{m}_{\boldsymbol{\beta}} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{m}_{\boldsymbol{\beta}} - \boldsymbol{\mu})\right\},$$

i.e.,

$$\mathbf{m}_{\boldsymbol{\beta}} \mid \text{else} \sim \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{V})$$

$$\begin{aligned}
p(\mathbf{S}_{\boldsymbol{\beta}}^{-1} \mid \text{else}) &\propto |\mathbf{S}_{\boldsymbol{\beta}}^{-1}|^{L/2}\exp\left\{-\frac{1}{2}\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})\right\}|\mathbf{S}_{\boldsymbol{\beta}}^{-1}|^{(\nu-p-1)/2}\exp\left\{-\frac{1}{2}tr(\nu\boldsymbol{\psi}\mathbf{S}_{\boldsymbol{\beta}}^{-1})\right\}\\
&= |\mathbf{S}_{\boldsymbol{\beta}}^{-1}|^{(L+\nu-p-1)/2}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left(\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T\mathbf{S}_{\boldsymbol{\beta}}^{-1}\right) + \mathrm{tr}(\nu\boldsymbol{\psi}\mathbf{S}_{\boldsymbol{\beta}}^{-1})\right)\right\}\\
&= |\mathbf{S}_{\boldsymbol{\beta}}^{-1}|^{(L+\nu-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T + \nu\boldsymbol{\psi}\right]\mathbf{S}_{\boldsymbol{\beta}}^{-1}\right)\right\},
\end{aligned}$$

which can be recognised as the kernel of a Wishart distribution with parameters $L + \nu$ and $\left(\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T + \nu\boldsymbol{\psi}\right)^{-1}$, that is,

$$\mathbf{S}_{\boldsymbol{\beta}}^{-1} \mid \text{else} \sim \text{Wishart}_p\left(\nu + L, \left(\sum_{k=1}^{L}(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \mathbf{m}_{\boldsymbol{\beta}})^T + \nu\boldsymbol{\psi}\right)^{-1}\right).$$