**Global data-generation parameters** We define a separate factor for each component's data-generating parameters $q(\Lambda_k)$, to approximate the posterior $p(\Lambda_k|\mathbf{x}, \mathbf{z}, \ldots)$. Each factor is Wishart with parameters $\hat{\nu}_k, \hat{W}_k$, updated as follows

$$q(\Lambda_k) = \text{Wishart}(\Lambda_k|\hat{\nu}_k, \hat{W}_k) \tag{17}$$

$$\hat{\nu}_k = \nu + \hat{N}_k, \qquad \hat{W}_k^{-1} = W^{-1} + s_k(\mathbf{x}) \tag{18}$$

Given $\hat{\nu}_k, \hat{W}_k^{-1}$, we compute the expected log probability under component $k$ for each data item $x_n$

$$\mathbb{E}_q\Big[\log p(x_n|\phi_k)\Big] = -\frac{D}{2}\log[2\pi] + \frac{1}{2}\mathbb{E}_q\Big[\log|\Lambda_k|\Big] - \frac{1}{2}\text{tr}(\mathbb{E}_q[\Lambda_k]x_n x_n^T) \tag{19}$$

Here, we use basic expectations under the Wishart distribution:

$$\mathbb{E}_q[\Lambda_k] = \hat{\nu}_k\hat{W}_k, \qquad \mathbb{E}_q[\log|\Lambda_k|] = \psi_D\Big(\frac{\hat{\nu}_k}{2}\Big) + D\log 2 + \log|\hat{W}_k| \tag{20}$$

where $\psi_D(a) = \sum_{d=1}^{D}\psi(a + \frac{1-d}{2})$ is the multivariate digamma function of dimension $D$.

## 2 Birth Moves for Mixture Models

**Overview.** As input, our birth procedure takes an existing variational model $q$ with $K$ components, together with global sufficient statistics $S^0 = [S_1^0\ S_2^0\ \ldots\ S_K^0]$ for the full dataset $\mathbf{x}$. The algorithm consists of 3 steps: *collection* of a subsample dataset $\mathbf{x}'$, *creation* of brand-new components by a fresh DP mixture model variational analysis of $\mathbf{x}'$, and *adoption* of these fresh new components by the full dataset $\mathbf{x}$. The output will be an expanded model $q^*$ with $K + J'$ components.

### 2.1 Collection of the target dataset $\mathbf{x}'$

We find it simplest to focus on a birth move which *targets* a specific component $k'$. After selecting the component $k'$, the birth move proceeds to subsample data $\mathbf{x}'$ associated with $k'$, using the existing local assignment factors $q(z_n)$ to identify which data items to subsample. Certainly other ways of subsampling exist, but this has an intuitive interpretation as targeting a single sub-optimal component which may be too coarse (explaining multiple ideal subclusters) and refining it.

**Selecting the target component $k'$.** The procedure for selecting which component $k'$ to target is not complicated. For understanding the mechanics of birth moves, it is fine to simply select the component $k'$ uniformly at random. If we have $K$ active components in original model $q$, then

$$k' \sim \text{Unif}(\{1, 2, \ldots K\}) \tag{21}$$

Many other schemes for choosing $k'$ can be considered. But the above is perfectly sufficient, albeit potentially slow at trying a diverse set of possible moves in a short timespan.

In practice, we recommend sampling $k'$ at random, but in a way that *biases* towards choosing components that (1) have more mass and (2) have not been targeted in the last few moves. Let $\hat{N}_k^0$ give the current expected count on the full dataset, and $L_k$ denote the number of passes through the data since component $k$ was last chosen for a birth move.

$$p(k' = k) \propto (\hat{N}_k^0) * (L_k)^2 \tag{22}$$

Squaring the $L_k$ term forces the algorithm to not wait very long between trying all possible components, ensuring good coverage of the space of all possible moves. We found that this revised selection procedure improved the speed with which our algorithm recovered all missing components, but uniform selection should eventually reach the same high-quality configurations.

**Sampling a dataset targeted on component $k'$.** After selecting $k'$, next we collect a *targeted* dataset $\mathbf{x}'$ with size at most $N'$. We recommend choosing $N'$ large enough that necessary "undiscovered" components (not in the existing set $\{1, 2, \ldots K\}$ can be learned, but still small enough that running many batch VB iterations does not take more than a few seconds. We found $N' = 10000$