Supplementary Material for Bayesian Hierarchical Clustering with Exponential Family: Small-Variance Asymptotics and Reducibility

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

1 Derivations in Section 3.2

For $\boldsymbol{X}_c = \{\boldsymbol{x}_i\}_{i\in c} \overset{\text{iid}}{\sim} p(\boldsymbol{x}|\beta,\boldsymbol{\theta})$, we have

$$p(\bar{t}_c|\beta,\boldsymbol{\theta}) = \exp\left\{\beta|c|\langle \bar{t}_c, \boldsymbol{\theta}\rangle - \beta|c|\psi(\boldsymbol{\theta}) - \sum_{i\in c} h_\beta(\boldsymbol{x}_i)\right\}. \tag{1}$$

For notational simplicity, we let $\bar{t}_c = \boldsymbol{y}$ from now. By the normalization property,

$$\beta|c|\psi(\boldsymbol{\theta}) = \log \int \exp\left\{\beta|c|\langle \boldsymbol{y}, \boldsymbol{\theta}\rangle - \sum_{i\in c} h_\beta(\boldsymbol{x}_i)\right\} d\boldsymbol{y}. \tag{2}$$

Differentiating both sides by $\boldsymbol{\theta}$ yields

$$\beta|c|\frac{d\psi(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \int \beta|c|\boldsymbol{y}\cdot p(\boldsymbol{y}|\beta,\boldsymbol{\theta})d\boldsymbol{y}, \quad \frac{d\psi(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{y}]. \tag{3}$$

Also, we have

$$\begin{aligned}\beta|c|\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} &= \int \beta|c|y_j p(\boldsymbol{y}|\beta,\boldsymbol{\theta})\left(\beta|c|y_j - \beta|c|\frac{\partial\psi(\boldsymbol{\theta})}{\partial\theta_k}\right)dy_j \\ &= \beta^2|c|^2\mathbb{E}[y_j y_k] - \beta^2|c|^2\mathbb{E}[\boldsymbol{y}]_j\mathbb{E}[\boldsymbol{y}]_k = \beta^2|c|^2\mathrm{cov}(y_j,y_k).\end{aligned} \tag{4}$$

Hence,

$$\frac{1}{\beta|c|}\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} = \mathrm{cov}(y_j,y_k) = \int (y_j - \mathbb{E}[\boldsymbol{y}]_j)(y_k - \mathbb{E}[\boldsymbol{y}]_k)p(\boldsymbol{y}|\beta,\boldsymbol{\theta})dy_j dy_k. \tag{5}$$

Differentiating this again yields

$$\begin{aligned}\frac{1}{\beta|c|}\frac{\partial^3\psi(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l} &= \int (y_j - \mathbb{E}[\boldsymbol{y}]_j)(y_k - \mathbb{E}[\boldsymbol{y}]_k)(y_l - \mathbb{E}[\boldsymbol{y}]_l)p(\boldsymbol{y}|\beta,\boldsymbol{\theta})dy_j dy_k dy_l \\ &= \mathbb{E}[(y_j - \mathbb{E}[\boldsymbol{y}]_j)(y_k - \mathbb{E}[\boldsymbol{y}]_k)(y_l - \mathbb{E}[\boldsymbol{y}]_l)].\end{aligned} \tag{6}$$

Unfortunately, this relationship does not continue after the third order; the fourth derivative of $\psi(\boldsymbol{\theta})$ is not exactly match to the fourth order central moment of $\boldsymbol{y}$. However, one can easily maintain the mth order central moment by manipulating the mth order derivative of $\psi(\boldsymbol{\theta})$, and mth order central moment always have the constant term $(\beta|c|)^{-m}$.

Equation (40) of the paper is a simple consequence of the equation (6). To prove the equation (41) of the paper, we use the following relationship:

$$\mathbb{E}\left[\frac{\epsilon_\phi(\bar{\boldsymbol{x}}_c)}{\Delta_\phi(\bar{\boldsymbol{x}}_c)}\right] \approx \frac{\mathbb{E}[\epsilon_\phi(\bar{\boldsymbol{x}}_c)]}{\mathbb{E}[\Delta_\phi(\bar{\boldsymbol{x}}_c)]} - \frac{\mathrm{cov}(\epsilon_\phi(\bar{\boldsymbol{x}}_c),\Delta_\phi(\bar{\boldsymbol{x}}_c))}{\mathbb{E}[\Delta_\phi(\bar{\boldsymbol{x}}_c)]^2} + \frac{\mathbb{E}[\epsilon_\phi(\bar{\boldsymbol{x}}_c)]\mathrm{var}[\Delta_\phi(\bar{\boldsymbol{x}}_c)]}{\mathbb{E}[\Delta_\phi(\bar{\boldsymbol{x}}_c)]^3}. \tag{7}$$

Now it is easy to show that this equation converges to zero when $\beta \to 0$; all the expectations and variances can be obtained by differentiating $\psi(\boldsymbol{\theta})$ for as many times as needed.