A Dirichlet Process Mixture Model for Clustering Longitudinal Gene Expression Data – Supplementary Materials

Jiehuan Sun[a], Jose D. Herazo-Maya[b], Naftali Kaminski[b], Hongyu Zhao[a], Joshua L. Warren[a,*,†]

[a]Department of Biostatistics, Yale University, New Haven, CT 06510, U.S.A.

[b]Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT 06519, U.S.A

[*]Correspondence to: Joshua L. Warren, Department of Biostatistics, Yale University, New Haven, CT 06510, U.S.A.

[†]Email: joshua.warren@yale.edu

1 Additional Simulations

We conduct additional simulations to study the robustness and performance of BClustLonG in different scenarios. Specifically, we simulate datasets with different number of genes, different number of clusters, and with different correlation structures among genes to study the performance of BClustLonG. Moreover, we simulate datasets where the intercepts and slopes are drawn from multivariate t distributions to study the robustness of BClustLonG to model misspecifications.

We adapt scenario RR in the paper to generate new simulation datasets as follows. The additional simulation scenarios have the same setup as in scenario RR for most parameters and the modified parameters are specified below.

- Scenario 1: We double the number of genes, that is $G = 80$. The intercepts and slopes for patients in each cluster are drawn from multivariate normal distributions. The means of intercepts and slopes are $\mathbf{1}_G$ for patients in cluster one and are $\mathbf{0}_G$ for patients in cluster two. The covariance matrices of the intercepts and slopes for patients in both clusters are taken to be the block diagonal matrix as $\left[\begin{smallmatrix} \boldsymbol{R} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{smallmatrix}\right]$.

- Scenario 2: We generate the intercepts and slopes from multivariate t distributions with degrees of freedom being 10. The means of intercepts and slopes are $\mathbf{1}_G$ for patients in cluster one and are $\mathbf{0}_G$ for patients in cluster two. The covariance matrices of the intercepts and slopes for patients in both clusters are taken to be $\boldsymbol{R}$.

- Scenario 3: The true number of clusters is four with each cluster having 25 patients. The intercepts and slopes for patients in each cluster are drawn from multivariate normal distributions. The means of intercepts (and slopes) are $-\mathbf{2}_G$, $\mathbf{0}_G$, $\mathbf{2}_G$, and $\mathbf{4}_G$ for patients in cluster one, two, three, and four, respectively. The covariance matrices of the intercepts and slopes for patients in all clusters are taken to be $\boldsymbol{R}$.

- Scenario 4: The setup is the same as that in scenario 3 except that the means and covariance matrices are changed. Specifically, the means of intercepts (and slopes) are $-\mathbf{1}_G$, $\mathbf{0}_G$, $\mathbf{1}_G$, and $\mathbf{2}_G$ for patients in cluster one, two, three, and four, respectively. The covariance matrices of the intercepts and slopes for patients in all clusters are taken to be the AR1 structure (autoregressive structure of order 1) with parameter 0.4.
- Scenario 5: The setup is the same as that in scenario 4 except that the covariance matrices are changed. Specifically, the covariance matrices of the intercepts and slopes for patients in all clusters are taken to be a randomly generated covariance matrix, which is done using the function *genPositiveDefMat* in R package "clusterGeneration".

Table S1: Comparisons of BClustLonG, BClustLonG0, MCLUST, EPGMM, and K-means in simulation settings. The numbers in each cell indicate the average adjusted Rand index (Avg.Rand) and the average number of clusters (Avg.Clust) estimated by each method under each scenario with standard deviations in parentheses.

Scenarios		BClustLonG	BClustLonG0	MCLUST	EPGMM	K-means
1	Avg.Rand	0.995 (0.013)	0.347 (0.048)	0.473 (0.085)	0.300 (0.470)	0.844 (0.088)
	Avg.Clust	2.1 (0.3)	8.5 (1.4)	4.2 (0.6)	1.3 (0.5)	2.0 (0.0)
2	Avg.Rand	0.924 (0.048)	0.251 (0.062)	0.293 (0.153)	0.000 (0.000)	0.429 (0.134)
	Avg.Clust	3.8 (1.4)	8.8 (1.8)	3.6 (0.9)	1.0 (0.0)	2.0 (0.0)
3	Avg.Rand	1.000 (0.000)	0.884 (0.097)	0.874 (0.207)	0.136 (0.216)	0.565 (0.184)
	Avg.Clust	4.0 (0.0)	5.5 (1.2)	4.8 (0.9)	2.1 (0.5)	2.3 (0.7)
4	Avg.Rand	0.937 (0.160)	0.864 (0.157)	0.562 (0.083)	0.000 (0.000)	0.488 (0.008)
	Avg.Clust	3.8 (0.6)	3.6 (0.6)	2.7 (1.0)	1.0 (0.0)	2.0 (0.0)
5	Avg.Rand	0.999 (0.006)	0.818 (0.198)	0.567 (0.095)	0.000 (0.000)	0.491 (0.004)
	Avg.Clust	4.0 (0.0)	3.5 (0.8)	2.9 (1.3)	1.0 (0.0)	2.0 (0.0)

From Table S1, we can see that BClustLonG performs better than other methods in all of these scenarios, suggesting that BClustLonG is robust to varying number of genes and clusters, different covariance matrices, and model misspecifications.

2 Sensitivity Analysis

Next, we study the sensitivity of BClustLonG to the prior distributions for α_{a1} and α_{a2} (α_{b1} and α_{b2}), which are taken to be Gamma(2,1) for our analyses in simulations and real data application. Specifically, we adopt Gamma(5,2), Gamma(5,5), and Gamma(8,1) as the prior distributions for α_{a1} and α_{a2} and compare the results to that of using Gamma(2,1). For each of the prior distributions, we run one chain with the same configuration as that described in Section 3.2 of the main text.

As shown in Figure S1, the posterior similarity matrices resulted from using different prior distributions for α_{a1} and α_{a2} are similar. Some minor differences exist in the upper right corner, where the patients tend to be clustered together slightly more often in the setting of using Gamma(5,5) compared to other settings. However, the two major clusters using HCLUST are the same for all these settings with the same adjusted Rand index (0.94), suggesting that BClustLonG is not very sensitive to the prior distributions for α_{a1} and α_{a2}.

Furthermore, the prior distributions for α_{a1} and α_{a2} mainly determines how well the covariance matrices are approximated. We can see that BClustLonG performs consistently well for all scenarios, even where the covariance matrices differ substantially (shown in Section 1 above and the Simulation Section of the paper), again suggesting that the performance of BClustLonG is not very sensitive to the prior distributions for α_{a1} and α_{a2}.

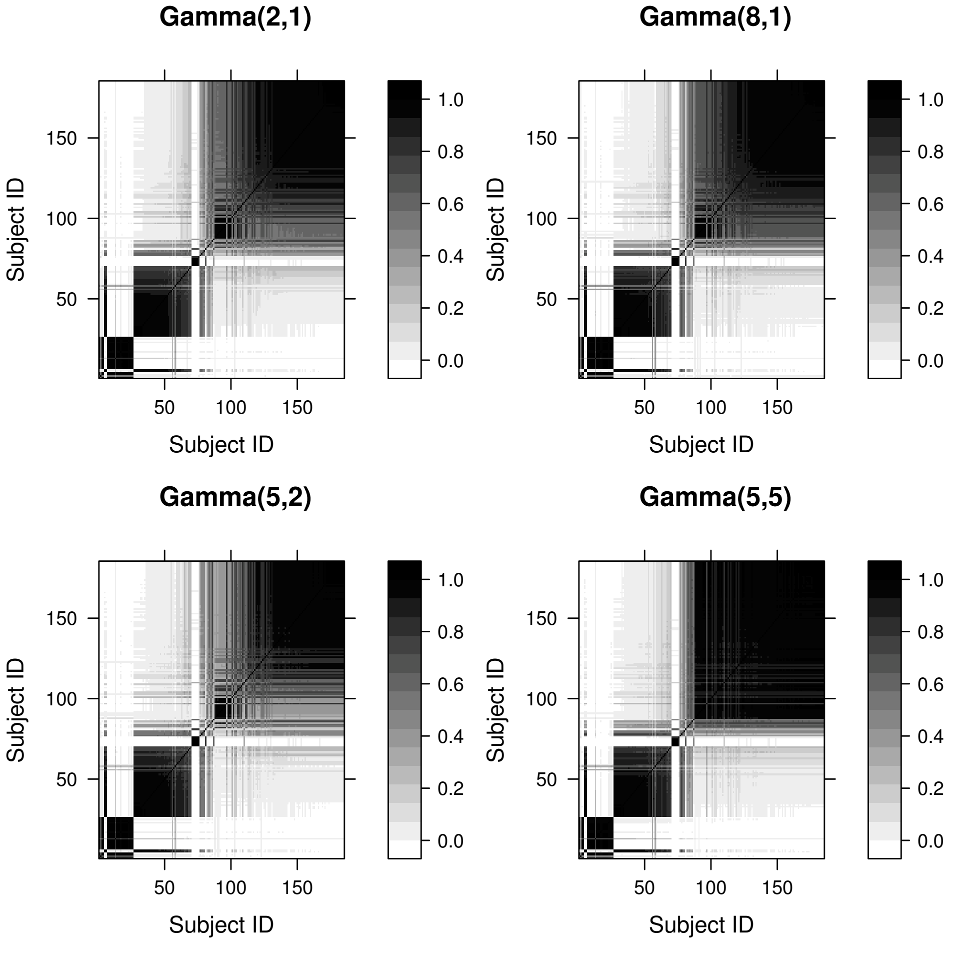

Figure S1: The posterior similarity matrices generated by using different hyperparameters specifications in the prior distributions for α_{a1} and α_{a2} in the injury data application.

3 MCMC algorithm

Finally, we provide the detailed MCMC sampling algorithm for our proposed method, BClustLonG, which clusters on both intercepts and slopes. The algorithms for BClustLonG that clusters only on intercepts or slopes can be easily adapted from the algorithm below. Notationally, let $[\boldsymbol{v}]_i$ denote the ith element of any vector $\boldsymbol{v}$ and $[\boldsymbol{V}]_{ij}$, $[\boldsymbol{V}]_{\cdot j}$, and $[\boldsymbol{V}]_{i\cdot}$ denote the (i,j)th entry, the jth column, the ith row of any matrix $\boldsymbol{V}$, respectively. Also, let K denotes the number of clusters at the current iteration (note that the number of clusters is changing from iterations to iterations) and n_k be the number of subjects belonging to kth cluster. And, let e_i be the cluster membership for subject i, that is $e_i = k$ if subject i belongs to the kth cluster.

The MCMC sampling proceeds in the following steps:

1. The random intercept coefficients for each subject are drawn as follows.

$$\boldsymbol{a}_i|... \sim \text{MVN}\left(\boldsymbol{\Omega}_a\left\{\boldsymbol{\Sigma}_{a0}^{-1}\boldsymbol{a}_{\mu i} + \sum_{t=1}^{T_i}\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{Y}_i(x_{it}) - \boldsymbol{b}_i x_{it}\right\}\right\}, \boldsymbol{\Omega}_a\right), \tag{S1}$$

   where $\boldsymbol{\Sigma}_{a0} = \boldsymbol{\Lambda}_a\boldsymbol{\Lambda}_a^T + \boldsymbol{\Sigma}_a$, $\boldsymbol{\Omega}_a^{-1} = \boldsymbol{\Sigma}_{a0}^{-1} + T_i\boldsymbol{\Sigma}^{-1}$, and $|...$ denotes the distribution is conditional on all other parameters.

2. The random slope coefficients for each subject are drawn as follows.

$$\boldsymbol{b}_i|... \sim \text{MVN}\left(\boldsymbol{\Omega}_b\left\{\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_{\mu i} + \boldsymbol{\Sigma}^{-1}\sum_{t=1}^{T_i} x_{it}\{\boldsymbol{Y}_i(x_{it}) - \boldsymbol{a}_i\}\right\}, \boldsymbol{\Omega}_b\right), \tag{S2}$$

   where $\boldsymbol{\Sigma}_{b0} = \boldsymbol{\Lambda}_b\boldsymbol{\Lambda}_b^T + \boldsymbol{\Sigma}_b$ and $\boldsymbol{\Omega}_b^{-1} = \boldsymbol{\Sigma}_{b0}^{-1} + \boldsymbol{\Sigma}^{-1}(\sum_{t=1}^{T_i} x_{it}^2)$.

3. The gene specific variances are drawn as follows.

$$[\boldsymbol{\Sigma}]_{gg}|... \sim \text{Inverse Gamma}\left(v_1 + \frac{1}{2}\sum_i T_i, v_2 + \frac{1}{2}\sum_i\sum_{t=1}^{T_i}\{Y_{ig}(x_{it}) - a_{ig} - b_{ig}x_{it}\}^2\right), \tag{S3}$$

   where v_1, v_2 are the hyperparameters in the prior distribution for $[\boldsymbol{\Sigma}]_{gg}$.

4. The cluster-specific mean of intercept coefficients ($\boldsymbol{a}_\mu^{(k)}$) are drawn as follows.

$$\boldsymbol{a}_\mu^{(k)}|... \sim \text{MVN}\left(\boldsymbol{\Omega}_a^{(k)}\left(\boldsymbol{\Sigma}_{a0}^{-1}\sum_{i:e_i=k}\boldsymbol{a}_i + \sigma_{a0}^{-2}\boldsymbol{I}_G\boldsymbol{a}_{\mu 0}\right), \boldsymbol{\Omega}_a^{(k)}\right), \tag{S4}$$

   where $\boldsymbol{\Omega}_a^{(k)} = \left\{\sigma_{a0}^{-2}\boldsymbol{I}_G + n_k\boldsymbol{\Sigma}_{a0}^{-1}\right\}^{-1}$, $\boldsymbol{\Sigma}_{a0} = \boldsymbol{\Lambda}_a\boldsymbol{\Lambda}_a^T + \boldsymbol{\Sigma}_a$, and $\{\boldsymbol{a}_{\mu 0}, \sigma_{a0}^2\}$ are the hyperparameters in the base distribution of the DP prior.

5. The overall mean of intercept coefficients ($\boldsymbol{a}_{\mu 0}$) are drawn as follows.

$$\boldsymbol{a}_{\mu 0}|... \sim \text{MVN}\left(\boldsymbol{\Omega}_{a0}\left(\sigma_{a0}^{-2}\sum_{k}\boldsymbol{a}_{\mu}^{(k)}\right), \boldsymbol{\Omega}_{a0}\right), \tag{S5}$$

where $\boldsymbol{\Omega}_{a0} = \left\{K\sigma_{a0}^{-2}\boldsymbol{I}_G + h^{-1}\boldsymbol{I}_G\right\}^{-1}$.

6. The elements of diagonal covariance matrix for intercepts are drawn as follows.

$$[\boldsymbol{\Sigma}_a]_{gg}|... \sim \text{Inverse Gamma}\left(v_1 + \frac{N}{2}, v_2 + \frac{1}{2}\sum_{i=1}^{N}\left(a_{ig} - [\boldsymbol{a}_{\mu i}]_g - [\boldsymbol{\Lambda}_a\boldsymbol{\eta}_{ai}]_g\right)^2\right), \tag{S6}$$

where v_1, v_2 are the hyperparameters in the prior distribution for $[\boldsymbol{\Sigma}_a]_{gg}$.

7. The hyperparameter σ_{a0}^2 in the base distribution for intercepts in the DP prior are drawn as follows.

$$\sigma_{a0}^2|... \sim \text{Inverse Gamma}\left(v_1 + \frac{N}{2}, v_2 + \frac{1}{2}\sum_{k=1}^{K}\sum_{g=1}^{G}\left([\boldsymbol{a}_{\mu}^{(k)}]_g - [\boldsymbol{a}_{\mu 0}]_g\right)^2\right), \tag{S7}$$

where v_1, v_2 are the hyperparameters in the prior distribution for σ_{a0}^2.

8. The subject specific factor scores for intercepts $\boldsymbol{\eta}_{ai}$ are drawn as follows.

$$\boldsymbol{\eta}_{ai}|... \sim \text{MVN}\left(\boldsymbol{\Omega}_{\eta a}\left\{\boldsymbol{\Lambda}_a^T\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{a}_i - \boldsymbol{a}_{\mu i})\right\}, \boldsymbol{\Omega}_{\eta a}\right), \tag{S8}$$

where $\boldsymbol{\Omega}_{\eta a}^{-1} = \boldsymbol{\Lambda}_a^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Lambda}_a + \boldsymbol{I}_{Ma}$.

9. The loading matrix $\boldsymbol{\Lambda}_a$ for the intercepts are drawn as follows.

$$[\boldsymbol{\Lambda}_a]_{g\cdot}|... \sim \text{MVN}\left(\boldsymbol{\Omega}_{\Lambda a}\{[\boldsymbol{\Sigma}_a]_{gg}^{-1}\boldsymbol{\eta}_a^T([\boldsymbol{A}]_{\cdot g} - [\boldsymbol{A}_{\mu}]_{\cdot g})\}, \boldsymbol{\Omega}_{\Lambda a}\right), \tag{S9}$$

where $\boldsymbol{\Omega}_{\Lambda a}^{-1} = [\boldsymbol{\Sigma}_a]_{gg}^{-1}\boldsymbol{\eta}_a^T\boldsymbol{\eta}_a + \boldsymbol{D}_g$, $\boldsymbol{D}_g$ is a diagonal matrix with $[\boldsymbol{D}_g]_{mm} = [\boldsymbol{\phi}_a]_{gm}[\boldsymbol{\tau}_a]_m$,$\boldsymbol{\eta}_a^T = [\boldsymbol{\eta}_{a1}, ..., \boldsymbol{\eta}_{an}]$, $[\boldsymbol{A}]_{\cdot g}$ is the g_{th} column of $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]^T$, and $[\boldsymbol{A}_{\mu}]_{\cdot g}$ is the g_{th} column of $\boldsymbol{A}_{\mu} = [\boldsymbol{a}_{\mu 1}, ..., \boldsymbol{a}_{\mu n}]^T$.

This updating rule is for a fixed number of factors M_a. A straightforward adaptive updating rule can be derived to tune the number of factors as the sampler progresses. Specifically, let $p(s) = \exp\{\kappa_1 + \kappa_2 s\}$ (κ_1 and κ_2 are some pre-specified negative numbers) and $u(s)$ be a Uniform(0,1) random number generated at the sth iteration. At the sth iteration, if $u(s) \leq p(s)$, we discard the columns in the loading matrix having all elements within some pre-specified small neighborhood of zero. If no such columns exists, we add a column to the loading matrix. We refer to [1] for more detailed discussions on this adaptive updating rule.

10. The variance parameters $\boldsymbol{\phi}_a$ for the loading matrix of intercepts are drawn as follows.

$$[\boldsymbol{\phi}_a]_{gm}|... \sim \text{Gamma}\left(v + \frac{1}{2}, v + \frac{1}{2}[\boldsymbol{\Lambda}_a]^2_{gm}[\boldsymbol{\tau}_a]_m\right), \tag{S10}$$

    where v are the hyperparameters in the prior distribution for $\boldsymbol{\phi}_a$.

11. The parameters $\boldsymbol{\gamma}_a$, which determine the variance parameters $\boldsymbol{\tau}_a$ for the loading matrix of intercepts, are drawn as follows.

$$[\boldsymbol{\gamma}_a]_m|... \sim \text{Gamma}\left(\alpha_{a2} + \frac{1}{2}G(M_a - m + 1), 1 + \frac{1}{2}\sum_{l=m}^{M_a}[\boldsymbol{\tau}_a]_l^{(m)}\left(\sum_{g=1}^{G}[\boldsymbol{\phi}_a]_{gl}[\boldsymbol{\Lambda}_a]^2_{gl}\right)\right), \quad \forall m \geq 2, \tag{S11}$$

    where α_2 are the hyperparameters in the prior distribution for $\boldsymbol{\gamma}_a$ and $[\boldsymbol{\tau}_a]_l^{(m)} = \frac{[\boldsymbol{\tau}_a]_l}{[\boldsymbol{\gamma}_a]_m}$. For $m = 1$, the conditional distribution is in the same form except that the α_{a2} is replaced by α_{a1}.

12. The hyperparameters α_{a2} and α_{a1} are drawn using Metropolis Hasting. The conditional density of α_{a2} and α_{a1} can be written as

$$f(\alpha_{a1}|...) \propto \Psi(\alpha_{a1}; 2, 1)\Psi([\boldsymbol{\gamma}]_1; \alpha_{a1}, 1), \tag{S12}$$

$$f(\alpha_{a2}|...) \propto \Psi(\alpha_{a2}; 2, 1)\prod_{m=2}^{M_a}\Psi([\boldsymbol{\gamma}_a]_m; \alpha_{a1}, 1), \tag{S13}$$

    where $\Psi(\cdot; a, b)$ is the density function of a gamma distribution with shape and scale being a and b, respectively. Then, a Metropolis Hasting updating algorithm can be used to draw α_{a1} using $\text{N}(0, s^2)$ as proposing distribution on the transformed variable $\log(\alpha_{a1})$, where s^2 is the variance parameter that can be used to tune the rejection rate. The parameter α_{a2} can be drawn using a similar strategy.

13. The corresponding parameters for the slopes $\{\boldsymbol{b}_\mu^{(k)}, \boldsymbol{b}_{\mu 0}, \boldsymbol{\Sigma}_b, \sigma^2_{b0}, \boldsymbol{\eta}_{bi}, \boldsymbol{\Lambda}_b, \boldsymbol{\phi}_b, \boldsymbol{\gamma}_b, \alpha_{b1}, \alpha_{b2}\}$ can be drawn similarly as shown in Equations (S4) - (S13).

14. The concentration parameter c in the DP prior is drawn using Metropolis Hasting. It is well known that the distribution of c depends only on K and the number of subjects N [2]. More specifically,

$$f(c|N, K) \propto c^K\Gamma(c)\pi(c)/\Gamma(c + N), \tag{S14}$$

    where $\pi(c)$ is the prior distribution on c, which is Uniform$(0, 10)$ in our case. Then, a Metropolis Hasting updating algorithm can be used to draw c using $\text{N}(0, s^2)$ as proposing distribution on the transformed variable $\log(\frac{10-c}{c})$, where s^2 is the variance parameter that can be used to tune the rejection rate.

15. The cluster membership indicators for each subject are drawn as follows. For cluster k, which is occupied by some subjects excluding subject i, we have

$$P(e_i = k|e_{-i}, ...) = ln_k^{(-i)}\Phi\left(\boldsymbol{a}_i; \boldsymbol{a}_\mu^{(k)}, \boldsymbol{\Sigma}_a^{(k)}\right)\Phi\left(\boldsymbol{b}_i; \boldsymbol{b}_\mu^{(k)}, \boldsymbol{\Sigma}_b^{(k)}\right) \tag{S15}$$

where l is some normalizing constant shared across all clusters, $n_k^{(-i)}$ is the number of subjects in the kth cluster excluding subject i, $\Phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function for the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\{\boldsymbol{a}_\mu^{(k)}, \boldsymbol{\Sigma}_a^{(k)}, \boldsymbol{b}_\mu^{(k)}, \boldsymbol{\Sigma}_b^{(k)}\}$ are the cluster specific means and variances for intercepts and slopes, respectively, which can be calculated as

$$\begin{aligned}
\boldsymbol{a}_\mu^{(k)} &= \boldsymbol{\Sigma}_a^{(k)}\left(\boldsymbol{\Sigma}_{a0}^{-1}\sum_{j:e_j=k,j\neq i}\boldsymbol{a}_j + \sigma_{a0}^{-2}\boldsymbol{I}_G\boldsymbol{a}_{\mu 0}\right), &\text{(S16)}\\
\boldsymbol{\Sigma}_a^{(k)} &= \left\{\sigma_{a0}^{-2}\boldsymbol{I}_G + \boldsymbol{\Sigma}_{a0}^{-1}n_k^{(-i)}\right\}^{-1}, &\text{(S17)}\\
\boldsymbol{b}_\mu^{(k)} &= \boldsymbol{\Sigma}_b^{(k)}\left(\boldsymbol{\Sigma}_{b0}^{-1}\sum_{j:e_j=k,j\neq i}\boldsymbol{b}_j + \sigma_{b0}^{-2}\boldsymbol{I}_G\boldsymbol{b}_{\mu 0}\right), &\text{(S18)}\\
\boldsymbol{\Sigma}_b^{(k)} &= \left\{\sigma_{b0}^{-2}\boldsymbol{I}_G + \boldsymbol{\Sigma}_{b0}^{-1}n_k^{(-i)}\right\}^{-1}, &\text{(S19)}
\end{aligned}$$

where $\boldsymbol{\Sigma}_{a0}$ and $\boldsymbol{\Sigma}_{b0}$ are the same as in Equations (S1) and (S2).

For a new cluster $\tilde{k}$, we have

$$P(e_i = \tilde{k}|e_{-i}, ...) = lc\Phi\left(\boldsymbol{a}_i; \boldsymbol{a}_{\mu 0}, \sigma_{a0}^2\boldsymbol{I}_G\right)\Phi\left(\boldsymbol{b}_i; \boldsymbol{b}_{\mu 0}, \sigma_{b0}^2\boldsymbol{I}_G\right), \tag{S20}$$

where c is the concentration parameter in the DP prior and $\{\boldsymbol{a}_{\mu 0}, \sigma_{a0}^2, \boldsymbol{b}_{\mu 0}, \sigma_{b0}^2\}$ are the means and variances for intercepts and slopes, respectively, in the base distribution.

Then, the indicators are drawn from a multinomial distribution for all possible clusters with corresponding probabilities given by Equations (S15) and (S20).

References

[1] A. Bhattacharya and D. B. Dunson, "Sparse bayesian infinite factor models," *Biometrika*, vol. 98, no. 2, pp. 291–306, 2011.

[2] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, no. 430, pp. 577–588, 1995.