Supplementary Notes for Modeling Homophily in Dynamic Networks with Application to HIV Molecular Surveillance

1 Models for Homophily

1.1 Data Generating Process and Simulation

We model growth of clusters over time t ($1996 \leq t \leq 2018$). For each newly infected PWH$_i$ at t_i, there are k_i number of available clusters available at t_i, where $\mathbf{x}_j^i$ is a vector of cluster-level covariates for the j^{th} cluster at time t_i ($1 \leq j \leq k_i$). As noted earlier, for a cluster of size 1, $\mathbf{x}_j^i$ is the covariate for PWH_i, who defines the j^{th} cluster, while for a cluster of size larger than 1, $\mathbf{x}_j^i$ represents a function of the covariates for all of the PWH within cluster j. The newly infected PWH$_i$ either joins one of the clusters available at that time t_i or forms its own cluster. In the former case, the number of clusters at time t_i, is unchanged, i.e., $k_{i+1} = k_i$, but one of the clusters will have a new member, which we denote the i^{th} newly linked case (NLC^i). The covariate for NLC^i, denoted by $\mathbf{x}_{NLC}^i$, will be incorporated into the covariate for the cluster that was joined. In the latter case, the number of clusters at time t_{i+1} will increase by 1 to $k_{i+1} = k_i + 1$, with $\mathbf{x}_{NLC}^i$ forming the $(k_i + 1)$th cluster at t_{i+1}. We assume that $\mathbf{x}_j^i$ are independent ($1 \leq j \leq k_i$; $1 \leq i \leq n$).

We describe the data generating process for these two scenarios.

Scenario (1) $\mathbf{x}_{NLC}^i$ joins one of the jth clusters at time t_i, in which case the number of clusters at time t_{i+1} remains the same, $k_{i+1} = k_i$, but the jth cluster will have a new cluster-level covariate to reflect the addition of $\mathbf{x}_{NLC}^i$:

$$\left\{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_{k_i}^i\right\} = \left\{\mathbf{x}_1^i, ..., \mathbf{x}_j^{i+1} = \mathbf{h}\left(\mathbf{x}_j^i, \mathbf{x}_{NLC}^i\right), ..., \mathbf{x}_{k_i}^i\right\},$$

where $\mathbf{h}(\cdot, \cdot)$ is a vector-valued function that combines $\mathbf{x}_j^t$ and $\mathbf{x}_{NLC}^i$ to define $\mathbf{x}_j^{i+1} = \mathbf{h}\left(\mathbf{x}_j^i, \mathbf{x}_{NLC}^i\right)$.

Scenario (2) $\mathbf{x}_{NLC}^i$ forms its own cluster at time at time t_i, in which case the number of clusters at time t_{i+1} will grow by one to $k_{i+1} = k_i + 1$ and the k_{i+1} clusters at t_{i+1} are given by:

$$\left\{\mathbf{x}_1^{i+1}, \mathbf{x}_2^{i+1}, ..., \mathbf{x}_{k_i}^{i+1}, \mathbf{x}_{k_i+1}^{i+1}\right\} = \left\{\mathbf{x}_1^i, ..., \mathbf{x}_j^i ..., \mathbf{x}_{k_i}^i, \mathbf{x}_{NLC}^i\right\}.$$

We now discuss how the two scenarios are determined using Between-subject Multinomial Response models.

1.2 Data Generating Model

Given $\left\{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_{k_i}^i\right\}$ and $\mathbf{x}_{NLC}^i$, let $m_i = k_i + 1$ and consider a m_i-dimensional random vector $\mathbf{z}^i = \left(z_1^i, z_2^i, \ldots, z_{m_i}^i\right)^\top$, where z_l^i is a binary indicator and $\sum_{l=1}^{m_i} z_l^i = 1$. Let

$$\left\{d_1^i, d_2^i, ..., d_{(m_i-2)}^i, d_{(m_i-1)}^i\right\} = \left\{d\left(\mathbf{x}_1^i, \mathbf{x}_{NLC}^{i+1}\right), d\left(\mathbf{x}_2^i, \mathbf{x}_{NLC}^i\right), ..., d\left(\mathbf{x}_{k_i-1}^i, \mathbf{x}_{NLC}^i\right), d\left(\mathbf{x}_{k_i}^i, \mathbf{x}_{NLC}^i\right)\right\},$$

where $d(\cdot,\cdot)$ is a scalar similarity/dissimilarity function to determine if $\mathbf{x}_{NLC}^i$ joins $\mathbf{x}_j^i$ $(1 \le j \le k_i)$.

We assume that $\mathbf{z}^i$ conditional on $\mathbf{d}^i = \left\{d_1^i, d_2^i, ..., d_{(m_i-2)}^i, d_{(m_i-1)}^i\right\}$ follows a m_i-level Between-subject Multinomial Response model $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ defined as:

$$\begin{aligned}
&\mathbf{z}^i \mid \mathbf{d}^i \sim \text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right), \qquad (1)\\
&\boldsymbol{\eta}^i = \left(\eta_1^i, \eta_2^i, \ldots, \eta_{(m_i-2)}^i, \eta_{(m_i-1)}^i\right)^\top,\\
&\eta_j^i = \frac{\exp\left(\beta_0 + \beta_1 d_j^i\right)}{1 + \Sigma_i}, \quad 1 \le j \ \le m_i - 1, \quad \eta_{m_i}^i = \frac{1}{1 + \Sigma_i},\\
&\Sigma_i = \sum_{j=1}^{m_i-1} \exp\left(\beta_0 + \beta_1 d_j^i\right), \quad \sum_{j=1}^{m_i} z_j^i = 1.
\end{aligned}$$

The size of this multinomial is 1, i.e., $\sum_{j=1}^{m_i} z_j^i = 1$, and $\boldsymbol{\eta}^i$ denotes the vector of cell probabilities. Thus only one of the components of $\mathbf{z}^i$ is 1 and the rest is 0.

Note that we term the model in (1) a Between-subject Multinomial Response model because the cell probabilities $\boldsymbol{\eta}^i$ are determined by the covariates $\mathbf{d}^i$ that is a function of the NLC and clusters of subjects. This is different from the traditional within-subject Multinomial Response model, where the cell probabilities are determined by a single subject [Liu et al., 2021].

Scenario (1) This occurs when $z_j^i = 1$ for some j $(1 \le j \le m_i - 1)$. In this case, $\mathbf{x}_{NLC}^i$ joins the jth cluster at time t_i and the clusters at time t_{i+1} remain the same as at time t_i, except for the jth cluster that will have a new cluster-level covariate to reflect the addition of $\mathbf{x}_{NLC}^i$ to its members. Thus, $k_{i+1} = k_i$ and the cluster covariates for time t_{i+1} are given by:

$$\left\{\mathbf{x}_1^{i+1}, ..., \mathbf{x}_j^{i+1}, ..., \mathbf{x}_{k_{(i+1)}}^{i+1}\right\} = \left\{\mathbf{x}_1^i, ..., \mathbf{h}\left(\mathbf{x}_j^i, \mathbf{x}_{NLC}^i\right), ..., \mathbf{x}_{k_i}^i\right\}.$$

Scenario (2) This occurs when $z_{m_i}^i = 1$, in which case $\mathbf{x}_{NLC}^i$ forms a new cluster. The number of clusters at time t_{i+1} will grow by one to $k_i + 1$. Thus, $k_{i+1} = k_i + 1$ and the cluster

covariates for time t_{i+1} are given by:

$$\left\{\mathbf{x}_1^{i+1}, ..., \mathbf{x}_j^{i+1}, ..., \mathbf{x}_{k_{(i+1)}-1}^{i+1}, \mathbf{x}_{k_{(i+1)}}^{i+1}\right\} = \left\{\mathbf{x}_1^{i}, ..., \mathbf{x}_j^{i} ..., \mathbf{x}_{k_i}^{i}, \mathbf{x}_{NLC}^{i}\right\}.$$

We can readily fit the between-subject multinomial response models in (1) to the data generated. Alternatively, we can fit independent logistic regression, as we did in the analysis of the homophily data. We discuss the basis for this alternative approach next.

1.3 Relationship between Independent Bernoulli and Multinomial Distribution

Consider k independent Bernoulli $z_j \sim Bern(\rho_j)$ $(1 \leq j \leq k)$ and let $\mathbf{z} = (z_1, z_2, \ldots, z_k)^\top$. Then, $\mathbf{z}$ given $\sum_{j=1}^n z_j = 1$ has a multinomial:

$$\mathbf{z}|\sum_{j=1}^{n} z_j = 1 \sim Mult(\mathbf{p},1), \quad \mathbf{p} = (p_1, p_2, \ldots, p_k)^\top,$$

$$p_j = \frac{\rho_j}{\sum_{l=1}^k \rho_l + \sum_{i\neq j}^k \frac{\rho_j(\rho_j-\rho_i)}{1-\rho_j}}, \quad 1 \leq j \leq k.$$

To show the above relationship between $\mathbf{p}$ and $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_k)^\top$, first consider p_1:

$$\begin{aligned}
p_1 &= \Pr\left(Z_1 = 1, Z_j = 0 \text{ for all } j \neq 1 \mid \sum_{j=1}^{k} Z_j = 1\right) \\
&= \frac{\rho_1 \prod_{j=2}^k (1-\rho_j)}{\sum_{j=1}^k \rho_j \prod_{l\neq j} (1-\rho_l)} \\
&= \frac{\rho_1 \prod_{j=2}^k (1-\rho_j)}{\sum_{j=1}^k \rho_j \prod_{l=2}^k (1-\rho_l) + \sum_{j=2}^k \rho_j \prod_{l\neq 1,j} (1-\rho_l)(\rho_j-\rho_1)} \\
&= \left(\frac{\sum_{j=1}^k \rho_j}{\rho_1} + \frac{\sum_{j=2}^k \rho_j \prod_{l\neq 1,j} (1-\rho_l)(\rho_j-\rho_1)}{\rho_1 \prod_{j=2}^k (1-\rho_j)}\right)^{-1} \\
&= \left(\frac{\sum_{j=1}^k \rho_j}{\rho_1} + \sum_{j=2}^{k} \frac{\rho_j(\rho_j-\rho_1)}{\rho_1(1-\rho_j)}\right)^{-1} \\
&= \frac{\rho_1}{\sum_{j=1}^k \rho_j + \sum_{j=2}^k \frac{\rho_j(\rho_j-\rho_1)}{(1-\rho_j)}}.
\end{aligned}$$

In general, the multinomial event probability p_j is given by:

$$\Pr\left(Z_j = 1, Z_l = 0 \text{ for all } l \neq j \mid \sum_{l=1}^{k} Z_l = 1\right) = \frac{\rho_j}{\sum_{l=1}^n \rho_l + \sum_{l\neq j} \frac{\rho_l(\rho_l-\rho_j)}{1-\rho_l}}.$$

Thus by viewing the data generating process as the conditional distribution of independent Bernonlli's $z_1, z_2, \ldots, z_k^\top$ under the constraint $\sum_{j=1}^{n} z_j = 1$, we can fit the data generated using independent logistic regression.

1.4 Model Fitting

1.4.1 Model Fitting based on Between-subject Multinomial Response Model

Given n newly infected cases, we have observed data:

$$\mathbf{x}_{NLC}^i, \quad \left\{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_{k_i}^i\right\}, \quad \mathbf{z}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i+1}^i\right)^\top, \quad 1 \leq i \leq n. \tag{2}$$

If $\mathbf{x}_{NLC}^i$ joins a cluster at time t_i, say, j, then $z_j^i = 1$ for some j $(1 \leq j \leq k_i)$. If $\mathbf{x}_{NLC}^i$ forms its own cluster, then $z_{k_i+1}^i = 1$.

First, we compute the similarity/dissimilarity variables (that serve as homophily covariates) for the clusters:

$$\mathbf{d}^i = \left\{d_1^i, d_2^i, ..., d_{k_i}^i\right\} = \left\{d\left(\mathbf{x}_1^i, \mathbf{x}_{NLC}^i\right), d\left(\mathbf{x}_2^i, \mathbf{x}_{NLC}^i\right), ..., d\left(\mathbf{x}_{k_i}^i, \mathbf{x}_{NLC}^i\right)\right\}, \quad 1 \leq i \leq n. \tag{3}$$

Then, we fit a (k_i+1)-level between-subject multinomial response model $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ using maximum likelihood:

$$\begin{aligned}
&\mathbf{z}^i \mid \mathbf{d}^i \sim \text{Multi}_b\left(\boldsymbol{\eta}^i\right), \\
&\boldsymbol{\eta}^i = \left(\eta_1^i, \eta_2^i, \ldots, \eta_{k_i}^i, \eta_{k_i+1}^i\right) \\
&\eta_j^i = \frac{\exp\left(\beta_0 + \beta_1 d_j^i\right)}{1 + \Sigma_i}, \quad 1 \leq j \ \leq k_i, \quad \eta_{k_i+1}^i = \frac{1}{1+\Sigma_i}, \\
&\Sigma_i = \sum_{j=1}^{k_i} \exp\left(\beta_0 + \beta_1 d_j^i\right), \quad \sum_{j=1}^{k_i+1} z_j^i = 1, \quad 1 \leq i \leq n.
\end{aligned} \tag{4}$$

The log-likelihood function is given by:

$$l = \sum_{i=1}^{n} l_i = \sum_{t=1}^{n} \sum_{j=1}^{k_i} z_j^i \log(\eta_j^i).$$

1.4.2 Model Fitting based on Independent Logistic Regression

When fitting the Between-subject Multinomial response model to the data generated, we need to have full data that includes both Scenario 1 and Scenario 2. Within the context of the paper, we only have data for Scenario 1, i.e., we only have data when a newly infected case $\mathbf{x}_{NLC}^i$ joins a

cluster at time t_i. Thus, we only have a subset of the data in (2) under the constraint $z_j^i = 1$ for some j $(1 \leq j \leq k_i)$, or $\sum_{j=1}^{k_i} z_j^i = 1$ $(1 \leq i \leq n)$.

Based on the relationship between independent Bernoulli and multinomial distribution, we can model the subset of the observed data $\mathbf{z}^i$ with $\sum_{j=1}^{k_i} z_j^i = 1$ using $\sum_{t=1}^{n} k_i I\left(\sum_{j=1}^{k_i} z_j^i = 1\right)$ independent logistic regression models, where $I(A)$ denotes an indicator with $I(A) = 1$ if A is true and $I(A) = 1$ otherwise. Thus, given

$$\mathbf{x}_{NLC}^i, \quad \left\{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_{k_i}^i\right\}, \quad \mathbf{z}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i}^i\right)^\top, \quad \sum_{j=1}^{k_i} z_j^i = 1, \quad 1 \leq i \leq n, \tag{5}$$

we can fit k_i independent logistic regression models as described in the main paper.

2 Simulation Study

2.1 Study One

We use simulated data to examine (1) if data generated from the between-subject multinomial response model in Section 1.2 can be equivalently modeled by independent logistic regression as described in Section 1.3, and (2) performance of the independent logistic regression when fit to data generated from the between-subject multinomial response model.

For (1), we started with 5 clusters and set 10 as the total number of clusters. The cluster-level covariates for the beginning 5 clusters, $\{x_1, x_2, ..., x_5\}$, were simulated from 5 different normal distributions:

$$\left\{x_1^1, x_2^1, ..., x_5^1\right\}, \quad x_j^1 \sim N(2j - 1, 0.1), \quad 1 \leq j \leq 5.$$

To simulate a sample of size n from the between-subject Multinomial response model $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$, for $1 \leq i \leq n$ each newly infected case $\mathbf{x}_{NLC}^i$ was generated from a normal distribution:

$$\mathbf{x}_{NLC}^i \sim N(2r - 1, 1), \quad r \sim U_d\{1, 2, ..., k_i\}, \quad 2 \leq i \leq n,$$

where $U_d\{1, 2, ..., k_i\}$ denotes a discrete uniform distribution with values $1, 2, ..., k_i$. We use $d_j^i = d\left(x_j^i, x_{NLC}^i\right) = \left|x_j^i - x_{NLC}^i\right|^{-1}$ (truncated at 0.001 and 100) as the similarity/dissimilarity function and $h\left(x_j^i, x_{NLC}^i\right) = \frac{1}{2}\left(x_j^i + x_{NLC}^i\right)$ as the function to integrate each x_{NLC}^i with the cluster-level covariate x_j^t of the cluster that x_{NLC}^i joins. We set $n = 1,000$ for our simulation. The event probability $\boldsymbol{\eta}^i$ for the between-subject Multinomial response model for generating the multinomial

response $\mathbf{z}^i$ in (1) is given by:

$$\eta_j^i = \frac{\exp\left(\beta_0 + \beta_1 d_j^i\right)}{1+\Sigma_i} = \frac{\exp\left(-0.8 + 0.1 d_j^i\right)}{1+\Sigma_i}, \quad m_i = k_i + 1 \tag{6}$$

$$\eta_{m_i}^i = \frac{1}{1+\Sigma_i}, \quad \Sigma_i = \sum_{j=1}^{m_i-1} \exp\left(-0.8 + 0.1 d_j^{t+1}\right).$$

To reduce sampling variability as well as to compare estimates from fitting the between-subject multinomial response and independent logistic regression model, we controlled the total number of clusters to 10, i.e., $k_n = 10$, with the fixed sample size n. Thus, a newly infected case x_{NLC}^i will not form its own cluster for every t_i when $\sum_{j=1}^{k_i} z_j^i = 0$ until a certain number of samples are generated from $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ for a given number of clusters. For example, to grow the initial 5 clusters to 10 clusters, we select 6 subsample sizes, $1 < n_5 < n_6 \ldots < n_9 < n_{10}$, such that for $n_{l-1}+1 \le t \le n_l$ we generate $\mathbf{z}^i$ from $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ with k_i clusters, where $n_4 = 0$ and $5 \le l \le 10$. For our simulation study, we set $n_5 = 300, n_6 = 500, n_7 = 700, n_8 = 900, n_9 = 1100$ and $n_{10} = 1500$.

We fit both the between-subject Multinomial response model and independent logistic regression as described in Section in (1.3). For the between-subject Multinomial response model, we obtained: $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)^\top = (-0.7995, 0.0999)^\top$, which were quite close to the respective true values $\beta_0 = -0.8$ and $\beta_1 = 0.1$. For the independent logistic regression, we obtained: $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \widehat{\gamma}_1)^\top = (-2.0401, \widehat{\gamma}_1 = 0.0963)^\top$. Although $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ are not directly comparable, we can compare the fitted $\widehat{\boldsymbol{\eta}}^{t+1}$ and $\widehat{\boldsymbol{\xi}}^{t+1}$ based on substituting $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ in place of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Since the independent logistic regression is fit to the subject vector $\mathbf{z}_{sub}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i}^i\right)^\top$ of $\mathbf{z}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i}^i, z_{k_i+1}^i\right)^\top$, we compare $\widehat{\boldsymbol{\xi}}^i = \left(\widehat{\xi}_1^i, \widehat{\xi}_2^i, \ldots, \widehat{\xi}_{k_i}^i\right)^\top$ with the normalized subvector $\widehat{\boldsymbol{\eta}}_{sub}^i = \frac{1}{\widehat{s}_i}\left(\widehat{\eta}_1^i, \widehat{\eta}_2^i, \ldots, \widehat{\eta}_{k_i}^i\right)^\top$, where $\widehat{s}_i = \sum_{j=1}^{k_i} \widehat{\eta}_j^i$.

Shown below are the averaged $\widehat{\boldsymbol{\eta}}_{sub}^i$ and $\widehat{\boldsymbol{\xi}}^i$ over the samples $1 \le i \le n_5$ with 5 clusters, $\overline{\widehat{\boldsymbol{\eta}}}_{sub}^5$ and $\overline{\widehat{\boldsymbol{\xi}}}_{sub}^5$, and over the samples $n_9 \le i \le n_{10}$, $\overline{\widehat{\boldsymbol{\eta}}}_{sub}^{10}$ and $\overline{\widehat{\boldsymbol{\xi}}}_{sub}^{10}$:

$$\overline{\widehat{\boldsymbol{\eta}}}_{sub}^5 = (0.1973, 0.1967, 0.2006, 0.2060, 0.1994)^\top,$$

$$\overline{\widehat{\boldsymbol{\xi}}}_{sub}^5 = (0.2001, 0.1977, 0.2008, 0.2037, 0.1977)^\top,$$

$$\overline{\widehat{\boldsymbol{\eta}}}_{sub}^{10} = (0.1019, 0.1045, 0.1002, 0.0951, 0.1024, 0.0970, 0.0978, 0.1019, 0.0934, 0.1060)^\top,$$

$$\overline{\widehat{\boldsymbol{\xi}}}_{sub}^{10} = (0.1017, 0.1091, 0.1022, 0.0875, 0.1004, 0.1008, 0.1028, 0.0997, 0.0938, 0.1021)^\top.$$

In both cases, the averaged event probabilities from the fitted between-subject Multinomial response and independent logistic regression model were quite close to each other.

To see if estimates of $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \widehat{\gamma}_1)^\top$ in the independent logistic regression converge, we simulated a sample size of 450 and 900 homophily cases, i.e., all simulated 450 (900) newly infected cases that join existing clusters. The two sets of estimates are $(\widehat{\gamma}_0, \widehat{\gamma}_1) = (-2.104, 0.113)$ for 450 and $(\widehat{\gamma}_0, \widehat{\gamma}_1) = (-2.051, 0.105)$ for 900. They are close to each other.

2.2 Study Two

In this simulation study, we simulate data from one continuous and one binary predictor based on the distributions of the observed birthy year and ethnicity. The mean birth year is 1973 and the proportion of Hispanic Ethnicity (HE) is 35%. We again set the total number of clusters to 10. We simulate the continuous cluster-level covariate from 10 normal distributions with different means and the binary cluster-level covariate from 10 Beroulli distributions with different means. To make the 10 distributions of the cluster-level covariates similar to those of the study data, the 10 normal means average to 1973 and the 10 Bernoulli means average to 35%.

We again started with 5 clusters with the continuous cluster-level birth year covariate, $\{x_{11}, x_{12}, ..., x_{15}\}$, following the 5 different normal distributions:

$$\left\{x_{11}^1, x_{12}^1, ..., x_{15}^1\right\},$$

$$x_{11}^1 \sim N\left(1968, 1\right), x_{21}^1 \sim N\left(1971, 1\right), x_{31}^1 \sim N\left(1975, 1\right), x_{41}^1 \sim N\left(1978, 1\right), x_{51}^1 \sim N\left(1981, 1\right).$$

The 5 binary cluster-level HE covariates, $\{x_{21}, x_{22}, ..., x_{25}\}$, following the 5 different Bernoulli distributions:

$$\left\{x_{21}^1, x_{22}^1, ..., x_{25}^1\right\}, x_{21}^1 \sim Ber(0.25), x_{22}^1 \sim Ber(0.3), x_{23}^1 \sim Ber(0.35), x_{24}^1 \sim Ber(0.4), x_{25}^1 \sim Ber(0.45).$$

To simulate a sample of size n from the between-subject Multinomial response model $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$, for $1 \leq i \leq n$ each newly infected case $\mathbf{x}_{NLC}^i = \left(x_{NLC1}^i, x_{NLC2}^i\right)^\top$ was generated from a normal and a Beroulli distribution as follows:

$$\mathbf{x}_{NLC1}^i \sim N\left(V_{1r}, 1\right), \quad \mathbf{x}_{NLC2}^i \sim Ber\left(V_{2r}\right), \quad r \sim U_d\{1, 2, ..., k_i\}, \quad 2 \leq i \leq n,$$

where $U_d\{1, 2, ..., k_i\}$ denotes a discrete uniform distribution with values $1, 2, ..., k_i$, V_{1r} and V_{2r} denotes the r-th row of vector V_1 and V_2, and $V_1 = (1968, 1971, 1975, 1978, 1981, 1955, 1961, 1965, 1985, 1991)$, $V_2 = (0.25, 0.3, 0.35, 0.4, 0.45, 0.1, 0.15, 0.2, 0.5, 0.55)$.

We use $d_{1j} = d\left(x_{1j}^i, x_{NLC1}^i\right) = \left|x_{1j}^i - x_{NLC1}^i\right|$ and $d_{2j} = d\left(x_{2j}^i, x_{NLC2}^i\right) = p_j^{x_{NLC2}^i}\left(1 - p_j\right)^{1 - x_{NLC2}^i}$ as the similarity/dissimilarity function for the continuous and binary cluster-level covariate. We use $\mathbf{h}\left(\mathbf{x}_j^i, \mathbf{x}_{NLC}^i\right) = \frac{1}{2}\left(\mathbf{x}_j^i + \mathbf{x}_{NLC}^i\right)$ as the function to integrate each $\mathbf{x}_{NLC}^i$ with the cluster-level covariate $\mathbf{x}_j^i$ of the cluster that $\mathbf{x}_{NLC}^i$ joins. We set $n = 1,000$ for our simulation. The event

probability $\boldsymbol{\eta}^i$ for the between-subject Multinomial response model for generating the multinomial response $\mathbf{z}^i$ in (1) is given by:

$$\eta_j^i = \frac{\exp\left(\beta_0 + \beta_1 d_{1j}^i + \beta_2 d_{2j}^i\right)}{1 + \Sigma_{t+1}}, \quad \eta_{m_i}^i = \frac{1}{1+\Sigma_i}, \quad m_i = k_i + 1$$

$$\Sigma_i = \sum_{j=1}^{m_i - 1} \exp\left(\beta_0 + \beta_1 d_{1j}^i + \beta_2 d_{2j}^i\right).$$

We set $\beta_0 = 1$, $\beta_1 = -0.15$ and $\beta_2 = 0.9$. We use a negative β_1 and positive β_2 so that smaller d_{1j} and larger d_{2j} will increase the probability of joining the

To reduce sampling variability as well as to compare estimates from fitting the between-subject multinomial response and independent logistic regression model, we controlled the total number of clusters to 10, i.e., $k_n = 10$, with the fixed sample size n. Thus, a newly infected case x_{NLC}^i will not form its own cluster for every i when $\sum_{j=1}^{k_i} z_j^i = 0$ until a certain number of samples are generated from $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ for a given number of clusters. For example, to grow the initial 5 clusters to 10 clusters, we select 6 subsample sizes, $1 < n_5 < n_6 \ldots < n_9 < n_{10}$, such that for $n_{l-1} + 1 \leq t \leq n_l$ we generate $\mathbf{z}^i$ from $\text{Multi}_b\left(\boldsymbol{\eta}^i, 1\right)$ with k_i clusters, where $n_4 = 0$ and $5 \leq l \leq 10$. For our simulation study, we set $n_5 = 300, n_6 = 500, n_7 = 700, n_8 = 900, n_9 = 1100$ and $n_{10} = 1500$.

We fit both the between-subject Multinomial response model and independent logistic regression as described in Section in (1.3). For the between-subject Multinomial response model, we obtained: $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\right)^\top = (1.0047, -0.1509, 0.8931)^\top$, which were quite close to the respective true values $\beta_0 = 1$, $\beta_1 = -0.15$ and $\beta_2 = 0.9$. For the independent logistic regression, we obtained: $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2)^\top = (-0.9914, -0.1279, -0.0352)^\top$. Although $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ are not directly comparable, we can compare the fitted $\widehat{\boldsymbol{\eta}}^i$ and $\widehat{\boldsymbol{\xi}}^i$ based on substituting $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ in place of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Since the independent logistic regression is fit to the subject vector $\mathbf{z}_{sub}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i}^i\right)^\top$ of $\mathbf{z}^i = \left(z_1^i, z_2^i, \ldots, z_{k_i}^i, z_{k_i+1}^i\right)^\top$, we compare $\widehat{\boldsymbol{\xi}}^i = \left(\widehat{\xi}_1^i, \widehat{\xi}_2^i, \ldots, \widehat{\xi}_{k_i}^i\right)^\top$ with the normalized subvector $\widehat{\boldsymbol{\eta}}_{sub}^i = \frac{1}{\widehat{s}_i}\left(\widehat{\eta}_1^i, \widehat{\eta}_1^i, \ldots, \widehat{\eta}_{k_i}^i\right)^\top$, where $\widehat{s}_i = \sum_{j=1}^{k_i} \widehat{\eta}_j^i$.

Shown below are the averaged $\widehat{\boldsymbol{\eta}}_{sub}^i$ and $\widehat{\boldsymbol{\xi}}^i$ over the samples $1 \leq i \leq n_5$ with 5 clusters, $\overline{\widehat{\boldsymbol{\eta}}}_{sub}^5$

and $\overline{\widehat{\boldsymbol{\xi}}}^{5}_{sub}$, and over the samples $n_9 \leq i \leq n_{10}$, $\overline{\widehat{\boldsymbol{\eta}}}^{10}_{sub}$ and $\overline{\widehat{\boldsymbol{\xi}}}^{10}_{sub}$:

$$\overline{\widehat{\boldsymbol{\eta}}}^{5}_{sub} = (0.1979, 0.2001, 0.2017, 0.2019, 0.1984)^{\top},$$

$$\overline{\widehat{\boldsymbol{\xi}}}^{5}_{sub} = (0.1954, 0.1953, 0.2023, 0.2031, 0.2040)^{\top},$$

$$\overline{\widehat{\boldsymbol{\eta}}}^{10}_{sub} = (0.0943, 0.1098, 0.1012, 0.1041, 0.1073, 0.0820, 0.0994, 0.1077, 0.0921, 0.1021)^{\top},$$

$$\overline{\widehat{\boldsymbol{\xi}}}^{10}_{sub} = (0.0949, 0.1007, 0.0978, 0.1047, 0.1043, 0.0939, 0.0964, 0.0983, 0.1043, 0.1045)^{\top}.$$

In both cases, the averaged event probabilities from the fitted between-subject Multinomial response and independent logistic regression model were quite close to each other.

References

J Liu, X Zhang, T Chen, T Wu, T Lin, L Jiang, S Lang, L Liu, L Natarajan, JX Tu, et al. A semi-parametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics*, 2021.