# Robust Clustering With Subpopulation-Specific Deviations

Briana J. K. Stephenson, Amy H. Herring & Andrew Olshan

Taylor & Francis
Taylor & Francis Group

Check for updates

# Robust Clustering With Subpopulation-Specific Deviations

Briana J. K. Stephenson[a], Amy H. Herring[b], and Andrew Olshan[c,d]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC; [b]Department of Statistical Science, Duke University, Durham, NC; [c]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; [d]The National Birth Defects Prevention Study

### ABSTRACT

The National Birth Defects Prevention Study (NBDPS) is a case-control study of birth defects conducted across 10 U.S. states. Researchers are interested in characterizing the etiologic role of maternal diet, collected using a food frequency questionnaire. Because diet is multidimensional, dimension reduction methods such as cluster analysis are often used to summarize dietary patterns. In a large, heterogeneous population, traditional clustering methods, such as latent class analysis, used to estimate dietary patterns can produce a large number of clusters due to a variety of factors, including study size and regional diversity. These factors result in a loss of interpretability of patterns that may differ due to minor consumption changes. Based on adaptation of the local partition process, we propose a new method, robust profile clustering, to handle these data complexities. Here, participants may be clustered at two levels: (1) globally, where women are assigned to an overall population-level cluster via an overfitted finite mixture model, and (2) locally, where regional variations in diet are accommodated via a beta-Bernoulli process dependent on subpopulation differences. We use our method to analyze the NBDPS data, deriving prepregnancy dietary patterns for women in the NBDPS while accounting for regional variability. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

## 1. Introduction

### 1.1. Multivariate Categorical Dietary Data

Food frequency questionnaires (FFQ) are often used to measure an individual's dietary intake over a period of time. The standard FFQ queries consumption/intake levels for over 100 foods and beverages (Subar et al. 2001). Some researchers focus on individual foods or nutrients, but foods are not consumed in isolation, and many nutritionists argue that a more holistic approach is needed (Motulsky 1989). When data include a large number of exposures or, in this case, food items of an FFQ, data reduction techniques such as factor analysis, latent class analysis, or other clustering approaches are often used (Kant 2004; Venkaiah, Brahmam, and Vijayaraghavan 2011; Sotres-Alvarez, Herring, and Siega-Riz 2010; Keshteli et al. 2015).

Clustering methods generally assume participants within each cluster share dietary habits and aim to maximize differences across groups. At times, these methods may oversimplify dietary behaviors, and it can be difficult to determine when dimension reduction is generalizable across different populations. The generalizability issue poses a concern for large heterogeneous populations. Subjects from these populations may often share a combination of behaviors that could be general to an overall population, but also specific to a subject's subpopulation, defined as any group indicated by a categorical covariate (e.g., state residence, ethnicity, SES, etc.). For example, in the United States, if a subpopulation was defined by a subject's state of residence, the foods consumed to characterize "American" diets would look different, due to regional differences. A healthy diet may incorporate an increased consumption of regional foods indigenous to a specific state (e.g., more avocados in Texas). Reconciling these regional differences with a single overall clustering method presents a loss of granularity.

On the other hand, creating separate models for each subpopulation can greatly diminish statistical power, and can lead to misleading characterizations of diets when generalizing across the entire population sample. Individuals that are classified as having a "healthy" diet in North Carolina, or a subpopulation where poor eating behaviors are prevalent, may be classified as having an "unhealthy" diet in Massachusetts, where a more "health-conscious" subpopulation is prevalent. The differences found within these regional patterns are crucial for the improvement of national dietary recommendations that can accommodate heterogeneity of dietary behaviors.

### 1.2. Standard Clustering Methods

The latent class model, introduced by Lazarsfeld and Henry (1968), is the most common clustering method for handling multivariate categorical response data. Because the number of clusters is typically unknown, models with a varying number of clusters are fit. The best number of clusters is often chosen

via likelihood ratio tests, Bayesian information criteria, Akaike information criteria, or the Lo–Mendell–Rubin test (Nylund, Asparouhov, and Muthén 2007). In practice, these criteria tend to be "greedy" and select solutions with a large number of clusters. To avoid the challenges of a large number of clusters, in practice researchers will often add interpretability as a criterion for selecting the number of clusters and look for the best solution with a "manageable" number (Ford et al. 2010; Silverwood et al. 2011). With no separation across subpopulations, this restriction on the number of overall classes may mask any localized dietary behaviors that could be pronounced at a subpopulation level.

Nonparametric Bayesian methods, for example, the Dirichlet process (DP) or overfitted finite mixture models, allow the number of clusters represented in a sample to grow as dimensions (sample size, number of variables) increase (Figueiredo and Jain 2002; Zhang, Ghahramani, and Yang 2004; Teh 2006; Miller and Harrison 2018; Rousseau and Mengersen 2011). In heterogeneous populations, a largely prevalent subpopulation may have its behaviors reflected in one of the overall clusters, while smaller subpopulation behaviors may still remain hidden in one of those same clusters. Further, while flexibly convenient, these models tend to overestimate the true number of clusters, permitting, at times, nonexistent clusters to appear (Miller and Harrison 2013). Outliers are often assigned to singleton clusters, which may measure lack of fit in the model more than a new pattern.

In multisite studies, often a hierarchical or nested approach is used to accommodate any potential differences among subpopulations. The hierarchical Dirichlet process assumes common clusters across groups (Teh et al. 2006). Nested approaches cluster subjects within a subpopulation and borrow information across subpopulations that share similar behaviors (Rodriguez, Dunson, and Gelfand 2008; Hu, Reiter, and Wang 2018). While useful in many applications, these techniques contain drawbacks. The number of nonempty clusters derived is highly sensitive to the selection of tuning parameters or hyperparameters. Often unrealistic, strong priors on these parameters are necessary to enforce sparsity and ensure subjects aggregate to a reasonable number of clusters, as interpretability once again becomes an issue.

What strains these clustering methods is the assumption of global clustering, where subjects belonging to the same cluster will exhibit the same expected set of responses for all variables included in the set. This is where subpopulation granularity is lost. Differences may exist for a subset of variables within that subpopulation. Local partition and hybrid Dirichlet process mixture models break this global clustering assumption by apply a two-tiered clustering scheme at a global and local level (Dunson 2009; Petrone, Guindani, and Gelfand 2009).

The hybrid Dirichlet process mixture model makes local cluster assignments to each individual variable and then clusters globally based on shared similarities of the local clusters with other subjects using a copula construction. The local clustering is considered a smaller subset of the overall population, but it is not specific to any identifiable subpopulation. Additionally, the hybrid Dirichlet prior is limited to continuous data. FFQ data are considered semiquantitative, as quantity of consumption is collected based on choices from several standardized portion

sizes. The frequency of consumption is collected in ordinal group levels (e.g., $X$ times per day, daily, weekly, monthly) (Subar et al. 2001). Given this data structure, the copula construction used in the hybrid DP is unable to generate unique solutions to discriminate global clusters when handling discrete data (Smith and Khaled 2012). A hierarchical hybrid Dirichlet could be proposed that would deal with discrete data and allow another level in the dependence structure. Such a model has not yet been proposed in the literature, to our knowledge.

Instead of clustering every variable individually, the local partition model allows an entire subset of variables to be partitioned to a local or global clustering system. Here, subjects cluster with other subjects who behave similarly for most or some of the variables analyzed. It is useful in characterizing the global cluster patterns because variables that do not provide much information to the overall population patterns can be assigned to a local cluster. As with the hybrid DP, the local level is not specific to any identifiable subpopulation. Those variables that are considered noise at the global level could be valuable at a subpopulation level. To identify which items are important in a general population setting and which items are important in a subpopulation setting, a statistically principled method is needed to identify and discriminate between the two levels of patterns, while still preserving a level of interpretability.

We organize this article as follows. Section 2 introduces the robust profile clustering (RPC) framework. Section 3 explores RPC functionality and performance against other methods using a simulation study. Section 4 presents a comprehensive analysis of the National Birth Defects Prevention Study (NBDPS) data and describes insights provided by the new methodology. We conclude with a short discussion in Section 5.

## 2. Robust Profile Clustering

In this section, we propose a novel class of RPC processes, which are designed to produce a robust set of "global" clusters summarizing the overall nutritional profile of an individual. Clustering does not follow the typical approach of restricting all of the measurements from individual $i$ to have the same cluster membership. Instead, local deviations from the global profile are allowed and explored within each identifiable subpopulation. We define this style as robust due to the model's ability to derive a global cluster that is robust to individual subpopulations that do not follow the global patterns everywhere. This robustness prevents RPC from introducing extraneous clusters to fit small deviations from global profiles within subpopulations.

Our data are nested within subpopulations. This allows local deviations to have a subpopulation-specific form. Introducing some notation, we let $i = 1, \ldots, n$ index individuals in a study, $s_i \in \{1, \ldots, S\}$ index the known subpopulation (essentially a categorical covariate) of individual $i$, and $C_i$ index the (unknown) global profile membership of subject $i$. Additionally, each subject has a multivariate data vector, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. Individual $i$ may not follow her global cluster allocation for all elements of this multivariate vector but may deviate for some as needed. We let $G_{ij} = 1$ if item $j$ is attributed to global cluster $C_i$ for individual $i$ and $G_{ij} = 0$ otherwise. We let $L_{ij}$ denote the local cluster allocation conditionally on $G_{ij} = 0$ and $s_i = s$.

An RPC process is then induced through probability models containing three components: (i) the global clustering, $C_i$, (ii) the variable deviation indicator, $G_{ij}$, and (iii) the local clustering membership, $L_{ij}$. There are a wide variety of choices that can be used for (i)–(iii), and to put in general form we let

$$\Pr(C_i = h) = \pi_h,$$
$$\Pr(G_{ij} = 1|s_i = s) = \nu_j^{(s)}, \quad (1)$$
$$\Pr(L_{ij} = l|s_i = s) = \lambda_l^{(s)}.$$

The data $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ are assumed to be drawn independently from a set of global density parameters $\Theta_{0 \cdot C_i, \cdot} = \{\theta_{0jC_i, \cdot}\}_{j=1}^p$ or local density parameters $\Theta_{1 \cdot L_{ij}, \cdot}^{(s)} = \{\theta_{1jL_{ij}, \cdot}^{(s)}\}_{j=1}^p$. For example, in a multivariate categorical case, we can use an equal number of categories $d$ for simplicity in exposition, where $\theta_{0jC_i, \cdot} = (\theta_{0jC_i, 1}, \ldots, \theta_{0jC_i, d})'$ and $\theta_{1jL_{ij}, \cdot} = (\theta_{1jL_{ij}, 1}^{(s)}, \ldots, \theta_{1jL_{ij}, d}^{(s)})'$. An extension to a variable number of categories is easily accommodated. We then let

$$y_{ij} \sim \begin{cases} \text{Mult}(\{1, \ldots, d\}, \theta_{0jC_i, \cdot}) & \text{if } G_{ij} = 1 \\ \text{Mult}(\{1, \ldots, d\}, \theta_{1jL_{ij}, \cdot}^{(s)}) & \text{if } G_{ij} = 0, s_i = s, \end{cases} \quad (2)$$

where the cluster- and food item-specific probability vectors $\{\theta_{0jC_i, \cdot}, \theta_{1jL_{ij}, \cdot}^{(s)}\} \overset{\text{iid}}{\sim} \text{Dirichlet}(1, \ldots, 1)$ a priori for $j = 1, \ldots, p$, $C_i = 1, 2, \ldots, L_{ij} = 1, 2, \ldots, s = 1, 2, \ldots, S$. Although $y_{ij}$ and $y_{ij'}$ are conditionally independent given $\boldsymbol{\theta}$, dependence is induced in marginalizing out the global cluster index $C_i$, as shown in expression (6).

We assume binary deviation vectors, $G_{i \cdot} = (G_{i1}, \ldots, G_{ip})$, for all $i \in (1, \ldots, n)$, are independent and identically distributed given $s_i = s$ with deviation probability $\nu_j^{(s)}$. Given the binary structure of a variable's deviation to one of two levels, we model each subpopulation with a beta-Bernoulli process to exploit its convenient conjugacy properties.

$$G_{ij} \sim \text{Bern}(\nu_j^{(s)}), \quad \nu_j^{(s)} \sim \text{Be}(1, \beta^{(s)}), \quad \beta^{(s)} \sim \text{Ga}(a, b). \quad (3)$$

The hyperparameters $(a, b)$ control the overall weight given to each local component (deviated food item) of its respective subpopulation. We let $a = b = 1$ as a default to place equal probability a priori on the global and local components, while allowing substantial uncertainty.

For the global clustering process, we assume an overfitted finite mixture model (van Havre et al. 2015), which greatly simplifies computation relative to the LPP. Let $K$ be a conservative upper bound on the number of clusters (say, $K = 50$). Then we have

$$\Pr(C_i = h) = \pi_h$$
$$\pi_{\cdot} = (\pi_1, \ldots, \pi_K)' \sim \text{Dir}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right). \quad (4)$$

For the local clustering process, we use a parallel formulation, letting

$$\Pr(L_{ij} = l \mid s_i = s) = \lambda_l^{(s)}$$
$$\lambda_{\cdot}^{(s)} = (\lambda_1^{(s)}, \ldots, \lambda_K^{(s)}) \sim \text{Dir}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right). \quad (5)$$

The induced subject-specific likelihood $\mathbf{y}_i$ conditionally on $G_{i \cdot} = \{G_{ij}\}_{j=1}^p$, $\Theta_{0 \cdot C_i, \cdot} = \{\theta_{0jC_i, \cdot}\}_{j=1}^p$, $\Theta_{1 \cdot L_{ij}, \cdot}^{(s)} = \{\theta_{1jL_{ij}, \cdot}^{(s)}\}_{j=1}^p$, but marginalizing out $C_i$ and $L_{i \cdot} = \{L_{ij}\}_{j=1}^p$ is given by

$$f(\mathbf{y}_i|-) = \left[\sum_{h=1}^K \pi_h \prod_{j:G_{ij}=1}^p \prod_{r=1}^d \theta_{0jh,r}^{\mathbf{1}(y_{ij}=r)}\right] \prod_{j:G_{ij}=0}^p$$
$$\left[\sum_{l=1}^K \lambda_l^{(s)} \prod_{r=1}^d (\theta_{1jl,r}^{(s)})^{\mathbf{1}(y_{ij}=r)}\right]. \quad (6)$$

## 3. Simulation Study

We use this section to explore the performance of existing methods with the newly proposed RPC via a simulation study. We observed model performances under seven scenarios: (1) strictly global, where data contain only true underlying global clusters, and no local clusters, (2) strictly local, where data contain only true underlying local clusters, but no global clusters, (3) hybrid case, where each subpopulation contained an increasing proportion of deviated variables to a local profile, (4) null case where the data contained no true underlying clusters at the global or local level, (5) a combination of scenarios (1)–(3) to mimic a large, heterogeneous population as expected in the NBDPS dietary dataset, (6) various settings violating our Beta-Bernoulli assumption, and (7) an exploration of the model under a continuous data setting. A total of 500 datasets were generated for each scenario across a preidentified finite set of subpopulations $S$. Each dataset described a total of $p = 50$ variables that each contain $d = 4$ categorical response variables. Continuous variables were randomly drawn from a normal distribution, according to its predefined cluster assignment. Variables were probabilistically allocated to global or local levels by prespecified fixed values of $\nu_j^{(s)}$. Exact specifications for each simulation case are detailed in Appendix A.

### 3.1. MCMC Computation

Discrete data were analyzed under five different models: (1) traditional LCA model with four classes (tLCA) (Lazarsfeld and Henry 1968), (2) Dirichlet process mixture model (DPM) (Dunson and Xing 2009), (3) over-fitted finite mixture model of $K = 50$ classes (oFMM) (Rousseau and Mengersen 2011), (4) local partition process model (LPP₂) (Dunson 2009), and (5) the newly proposed RPC model. Continuous data were analyzed under a sparse finite Gaussian mixture (GMM) of $K = 50$ classes and RPC models. While models (1)–(4) ignore any subpopulation effect, we include them to demonstrate how the RPC properties compare to the methods currently available.

Estimation was performed using a Gibbs sampler of 20,000 runs after a 5000 burn-in. Posterior medians of all model parameters were calculated from the MCMC output. To encourage mixing, label switching moves were imposed to allow better exploration of the parameter space. The random permutation sampler was applied for finite mixture model cases (tLCA, oFMM, GMM, RPC) (Frühwirth-Schnatter 2001). The DPM and LPP₂, which involve a Dirichlet process mixture model, had two label switching moves imposed to favor swapping of both

equal and unequal size clusters (Papaspiliopoulos and Roberts 2008).

Mixing efficiency and convergence were evaluated using trace plots of model parameters and randomly selected variables. Parameters were relabeled using a similarity matrix from the MCMC output that contained the pairwise posterior probabilities of two subjects being clustered together in a given iteration. Hierarchical clustering was then performed on the similarity matrix, using the complete linkage approach (Krebs 1989; Medvedovic and Sivaganesan 2002). A nonempty cluster was identified as any cluster with a posterior probability weight greater than 0.01.

Dirichlet hyperpriors of the finite mixture component weights found in the tLCA, oFMM, GMM, RPC were preset to $1/K_{\max}$, where $K_{\max}$ is the preset maximum number of clusters allowed in the model. The concentration hyperparameter of models containing a Dirichlet process was preset to 1. Parameters estimated from all existing methods were sampled in accordance with their algorithms found in their respective literature (Nylund, Asparouhov, and Muthén 2007; Dunson and Xing 2009; Rousseau and Mengersen 2011; Dunson 2009).

In the continuous case, normal gamma priors were generated based on the range of input data, such that the prior mean for each cluster was drawn from a normal distribution with mean 0 and variance $\tau = 10$. Cluster specific variance at both the global and local level were drawn from a Gamma distribution with shape $a = 0.1$ and scale $b = 10$.

Profile patterns from each respective model were derived using the posterior median of cluster density parameter estimates. Each cluster density contained a vector of posterior probabilities of a subject responding at a given level within that cluster. The response level containing the maximum probability of each variable was designated as the modal cluster pattern response. The modal cluster responses were used to identify any potential redundant clusters present. Sensitivity was measured as the proportion of subjects belonging to the same preassigned cluster were assigned to the same cluster by the model. Specificity was defined as the proportion of subjects who were not preassigned together remained not assigned together. Density concordance was measured by comparing how well each model-derived cluster density compared to the true cluster density. This was calculated by calculating the mean-squared error ($\text{MSE}_\theta$) of the true density parameters, $\{\theta_{0jC_i}\}_{(j=1)^p}$, to corresponding model cluster density parameters, $\{\hat{\theta}_{0jC_i}\}_{(j=1)^p}$. To evaluate RPCs discriminating feature of determining if a variable is global ($\hat{v}_j^{(s)} \to 1$) or local ($\hat{v}_j^{(s)} \to 0$), total MSE was calculated for each dataset, such that $\text{MSE}_v = \frac{\sum_s \sum_j (\hat{v}_j^{(s)} - v_j^{(s)})}{S \cdot p}$, where $\hat{v}_j^{(s)}$ is the posterior median estimate of $v_j^{(s)}$. Variables predicted as global will have higher probabilities approaching 1 and variables predicted as local will have lower probabilities approaching 0. A small $\text{MSE}_v$ implies a strong discrimination of the RPC model across all $p$ variables.

## 3.2. Results

All models were run using MATLAB (version 2017a). Examination of trace plots indicated good mixing and convergence of model parameters. The median number of nonempty clusters was collected across all 500 simulated sets for each case. The median summaries across all 500 simulated sets is summarized in Table 1. The number of nonempty global clusters derived in each model is found where $K_0 = \sum_k \mathbf{1}(\hat{\pi}_k > 0.01)$, where $\hat{\pi}_k$ is the posterior median component weight of cluster $k$. An issue with redundant clustering was evident in the RPC and on occasion the LPP$_2$ and GMM model. This is a common consequence when using sparse mixture models (Malsiner-Walli, Frühwirth-Schnatter, and Grün 2016). These models required an additional post-processing step to remove redundant clusters. The final set of unique nonempty clusters can be found by subtracting the number of redundant clusters from $K_0$. Global/local discrimination results were illustrated in the last column, $\text{MSE}_v$.

DPM, OFM, LPP$_2$, and RPC models performed well under the completely global case (1) and completely local case (2). The true cluster modes were identified with no additional clustering. With the exception of the RPC, all of the expected cluster patterns in the model were reflected at the global level in both cases. In the completely local (case 2) setting, patterns were masked in the LCA model due to the forced model restriction to $K = 4$ classes. In both of these cases, the RPC performed well in being able to recognize that all of the variables in the completely global case (1) should be assigned at the global level with an overall mean $v_j^{(s)} = 0.75 \pm 0.001$ across all variables $j \in (1, \ldots, p)$ and subpopulations $s \in (1, \ldots, S)$. As a result, all of the patterns in case 1 were reflected at the global level. Similarly, in the completely local setting (case 2), the RPC was able to recognize that all of the variables should be local with an overall mean $v_j^{(s)} = 0.03 \pm .0002$ across all variables and subpopulations. Consequently, the patterns in the completely local case were reflected at the local level. With all of the variables in the global setting (case 1) and local setting (case 2) favoring one level in the RPC model, few if any clusters were identifiable at the non-favored level. For example, two global profiles were derived in the local setting (case 2). However, the posterior probability corresponding to the modal response level of these profiles was 0.25, indicating no strong modal response for any variables or profiles at the global level.

With no true pattern evident in the null case (4), all of the models collapsed participants into a single cluster, with the exception of LPP$_2$. The LPP$_2$ identified an overwhelming number of small clusters (median = 47). The RPC model indicated a strong tendency to the single global cluster with $v_j^{(s)} > 0.99$ across all variables and subpopulations. As a result, at the local level no identifiable clusters were formed within each subpopulation. These results are consistent in reflecting no true pattern in the population set.

All of the models were able to identify the expected modes of each global profile pattern in case (3). However, LCA, DPM, OFM, LPP$_2$ models had a decreased sensitivity (57%) as additional clusters were derived to accommodate the hybridization of global and local variables. This resulted in subjects who originally belonged to the same cluster being dispersed across several similar, but not identical clusters. The RPC was the only model that had a high sensitivity (99%) and specificity (99%). It derived the true global patterns without additional clustering. At the local level, the model's imposed sparsity resulted in a

**Table 1.** Summary of simulated cases (1)–(7) for 500, respectively, simulated datasets.

| Case | Model | $K_0$ (IQR) | Redundant Clusters | Concordance $MSE_\theta$ | $MSE_\nu$ |
|---|---|---|---|---|---|
| | LCA | 3(0) | 0 | 0.09 | NA |
| | DPM | 3(0) | 0 | < 0.01 | NA |
| Global | OFM | 3(0) | 0 | 0.08 | NA |
| | LPP | 3(0) | 0 | < 0.01 | NA |
| | RPC | 5(0) | 2 | 0.01 | 0.07 |
| | LCA | 4(0) | 0 | 0.07 | NA |
| | DPM | 8(0) | 0 | < 0.01 | NA |
| Local | OFM | 8(0) | 0 | 0.08 | NA |
| | LPP | 7(1) | 0 | < 0.01 | NA |
| | RPC | 2(0)* | 0 | 0.01 | <0.01 |
| | LCA | 4(0) | 0 | 0.06 | NA |
| | DPM | 17(2) | 0 | < 0.01 | NA |
| Hybrid | OFM | 15(1) | 0 | 0.08 | NA |
| | LPP | 8(1) | 0 | 0.01 | NA |
| | RPC | 6(1) | 3 | 0.01 | 0.03 |
| | LCA | 1(0) | 0 | NA | NA |
| | DPM | 1(1) | 0 | NA | NA |
| Null | OFM | 1(1) | 0 | NA | NA |
| | LPP | 47(3) | 0 | NA | NA |
| | RPC | 1(0) | 0 | NA | NA |
| | LCA | 3(1) | 0 | 0.57 | NA |
| | DPM | 18(2) | 2 | < 0.01 | NA |
| Mock NBDPS | OFM | 17(1) | 0 | 0.08 | NA |
| | LPP | 12(1) | 1 | < 0.01 | NA |
| | RPC | 8(1) | 4 | 0.01 | 0.03 |
| Beta-Bernoulli | RPC | 5(0) | 2 | < 0.01 | 0.03 |
| | Separate: GMM | 3(0) | 1 | < 0.01** | NA |
| | Separate: RPC | 3(0) | 1 | < 0.01** | 0.06 |
| Continuous | Partial: GMM | 6(0) | 4 | < 0.01** | NA |
| | Partial: RPC | 6(2) | 4 | < 0.01** | 0.03 |
| | Overlap: GMM | 0(0) | 0 | NA | NA |
| | Overlap: RPC | 34(4) | 10 | 0.06** | 0.07 |

NOTES: $K_0$: median number of nonempty global clusters with a weight larger than 0.01. Concordance: median mean square error of true cluster density parameters and corresponding model cluster density parameters.
*Local profile patterns were not reflected at the global level.
**For the continuous case, concordance was only compared for variables that were preallocated global across both subpopulations.

single local cluster per subpopulation. These singleton clusters derived reflected a modal split between the two expected local profile modes. For example, if local variable $j$, had a response level of 4 for half the subpopulation and a response level of 3 for the remaining half of the subpopulation, the single local cluster would reflect an approximate probability of 0.5 at response level 3 and also at response level 4.

These results were consistent in the expanded mock NBDPS case (5). Due to subpopulations containing an increasing proportion of locally allocated variables, all of the models with the exception of RPC derived several extraneous clusters. The RPC had only one extraneous cluster, which was due to a subpopulation that reflected a completely local profile. The discrimination feature of the RPC model performed well in both the hybrid case (3) and mock NBDPS case (5) easily identifying which variables should be global or local ($\overline{MSE}_\nu = 0.03$, Figure B.1).

The $\{\beta^{(s)}\}_{s=1}^S$ parameter in the RPC model illustrates the weight of each subpopulation to the model's global or localized components. Lower values indicate a strong representation of global components. For example, in global case (1), the $\beta^{(s)}$ for each of the subpopulations were all less than 1. Higher values indicate a stronger representation of local components, as illustrated in local case (2), where $\beta^{(s)} = 19$ for each of the subpopulations.

Consistent with prior simulation cases, RPC was still able to preserve a strong level of specificity (99%) and sensitivity (99%) to the global and local profile patterns. The $\overline{MSE}_\nu$ for each subpopulation were consistent with RPC results of prior simulation cases. This consistency implies the model's robustness to violations of the Beta-Bernoulli assumption.

In the continuous case (7), redundancy in cluster profiles increased as the standard deviation increased for both RPC and GMM. The RPC performed well in highlighting the global/local discrimination feature with an $\overline{MSE}_\nu \leq 0.07$ in all three subcases. This feature led to a split in pattern identification. The expected global profile information was consistent with the expected global mean only when it was assigned to the global level. Likewise, expected local information was consistent with the expected local mean only when it was assigned to the local level. For example, all variables assigned to global profile 1 had a mean near 2.0, as expected. If variable 6 within subpopulation 1 was expected to deviate to the local level, the local profile densities at variable 6 reflected an expected mean of 5 for local profile 1 and −5 for local profile 2, as expected. When concordance was compared with only variables that were global across all subpopulations, both GLM and RPC were able to identify the parameter values.

## 4. Analysis of National Birth Defects Prevention Study Data

### 4.1. Multivariate Categorical Dietary Data

The NBDPS is an ongoing multistate population-based, case-control study of birth defects in the United States (Yoon et al. 2001). Infants were identified via birth defect surveillance in Arkansas, California, Iowa, Massachusetts, New Jersey, New York, Texas, Georgia, North Carolina, and Utah. We focus this analysis on dietary habits of control participants. Participants for this study included mothers with expected due dates from 1997 to 2009, totaling 9747 controls. Controls were defined as any live-born infant without any birth defects and were randomly selected from birth certificates or hospital records. Following Sotres-Alvarez et al. (2013), subjects were excluded who had multiple births, a prior history of birth defects, preexisting diabetes, or folate antagonist medication use from three months before pregnancy to the end of pregnancy, and unexplained. Mothers with daily energy intake below the 2.5th and above the 97.5th percentiles were also excluded to prevent inclusion of unlikely intake data. After exclusion criteria were applied, a total of 9010 controls were included for this analysis.

Food consumption was measured in grams per day and calculated by multiplying frequency of consumption by the standard portion size for each food item as outlined in Sotres-Alvarez et al. (2013). Because respondents are prone to over-estimate or under-estimate total intake (Jønneland et al. 1991; Haraldsdottir 1993), percentiles were computed by dividing an individual food consumed over the total foods consumed, using grams per day as the consumption metric. The distribution of food items showed a spike at zero, which is well known in the literature (Kipnis et al. 2009; Zhang et al. 2011). Keeping with common dietary analysis practices, these percentiles were aggregated into four relative consumption levels: no consumption, low consumption (0–33% consumed), medium consumption (33–66% consumed), and high consumption (66–100% consumed). A total of $p = 63$ food items were included in the study with four consumption levels ($d = 4$) fit using a categorical distribution.

For our RPC model, we define $\theta_{0jk,r}$ as the probability of a subject having a consumption level of $r$ from food item $j$, given allocation to global dietary profile $C_i = k$. Similarly, we define $\theta_{1jl,r}^{(s)}$ as the probability of a subject from subpopulation $s$ having a consumption level $r$ from food item $j$ given allocation to local dietary cluster $L_{ij} = l$.

The cluster-specific parameters were each drawn from a flat, symmetric Dirichlet distribution, $\theta_{0jh,\cdot} = \{\theta_{0jh,1}, \ldots, \theta_{0jh,d}\} \sim \text{Dir}_d(\eta)$, $\theta_{1jl,\cdot}^{(s)} = \{\theta_{1jl,1}^{(s)}, \ldots, \theta^{(s)1jl,d}\} \sim \text{Dir}_d(\eta)$, where $\eta = 1$. The hyperparameters of the Beta-Bernoulli process component of the RPC were drawn from a gamma prior, $\beta^{(s)} \sim \Gamma(1, 1)$. To encourage a less informative Beta-Bernoulli process, $\beta^{(s)}$ was set to 1 for simplicity.

### 4.2. MCMC Performance

Following the simulation, inference is based on an MCMC run of 20,000 iterations, after a 5000 burn-in. Given the tendency, acknowledged in the simulation study, of parameters gravitat-ing to a preferred node and remaining there for subsequent iterations in large samples, the random permutation sampler was applied to encourage mixing (Frühwirth-Schnatter 2001). Furthermore, the overfitted model is also prone to generating extraneous and redundant clustering (van Havre et al. 2015). Redundancies were removed by creating a posterior pairwise comparison matrix based on the full MCMC output. Hierar-chical clustering was then performed using the complete link-age approach, restricting to the median number of nonempty clusters larger than 5% in size (Krebs 1989; Medvedovic and Sivaganesan 2002). This threshold was determined from the simulation study to focus on identifying the clusters of global interest. The trace plots of $\beta^{(s)}$ and $\pi_{1:K}$ showed a good mixing and rapid convergence. All model parameters were estimated by calculating the posterior median and 95% credible intervals.

### 4.3. NBDPS Results

#### 4.3.1. RPC Results

The RPC model identified a total of seven nonempty global cluster patterns. A heat map illustrating the patterns of the global profiles is provided in Figure 1. Global profile behaviors were described by foods most likely to have a given consumption level within each cluster. Figure 2 illustrates the top five foods with the highest probability of having a given consumption level for each global profile. Foods are listed from top to bottom in order of frequency of consumption, with the largest level of consumption illustrated at the bottom. Global profile 1 had a high consump-tion of meats and fatty foods (candy, chocolate, beef, chicken starches). Global profile 2 had a high consumption of fast foods (soda, white bread, French fries, ground beef, potato chips). Global profile 3 had a high consumption of chicken, cheese, and beef. Foods commonly found in a "Tex-Mex" or Latino style diet were likely to be consumed at a high level in global profile 4. Global profile 5 had a high consumption of "snack-style" foods (soda, French fries, tea, potato chips) as well as a medium consumption of pork and frankfurters. Global profile 6 had a high consumption of caffeine products (soda, coffee, tea). Global profile 7 was the most prudent profile with highly consumed foods including wheat bread, fruit cocktail, and low-fat milk.

The distribution of these profiles by each subpopulation is illustrated in Figure 3. Global profile 3 (gray) was the largest and most prominent profile in all states except for Arkansas and California. In Arkansas, global profile 5 was most prominent. Texas had a strong representation of global profile 4, which had a high consumption of foods commonly found in a "Tex-Mex" style diet (tortillas, refried beans, salsa).

While the food patterns found at the global level were shared among subjects from different subpopulations, unique behav-iors were more pronounced within each state at the local level. A subset of foods were found to deviate at the subpopulation level for all 10 states. A food with a tendency to deviate in favor of a subpopulation pattern was identified as $v_j^{(s)} < 0.50$. Eight foods were found to share a tendency to deviate in favor of nonconsumption in all 10 states (butter, squash, beef liver, beef tongue, coffee, tea, diet soda, folate cereal). Many of these foods were found to have consumption levels at the global level.
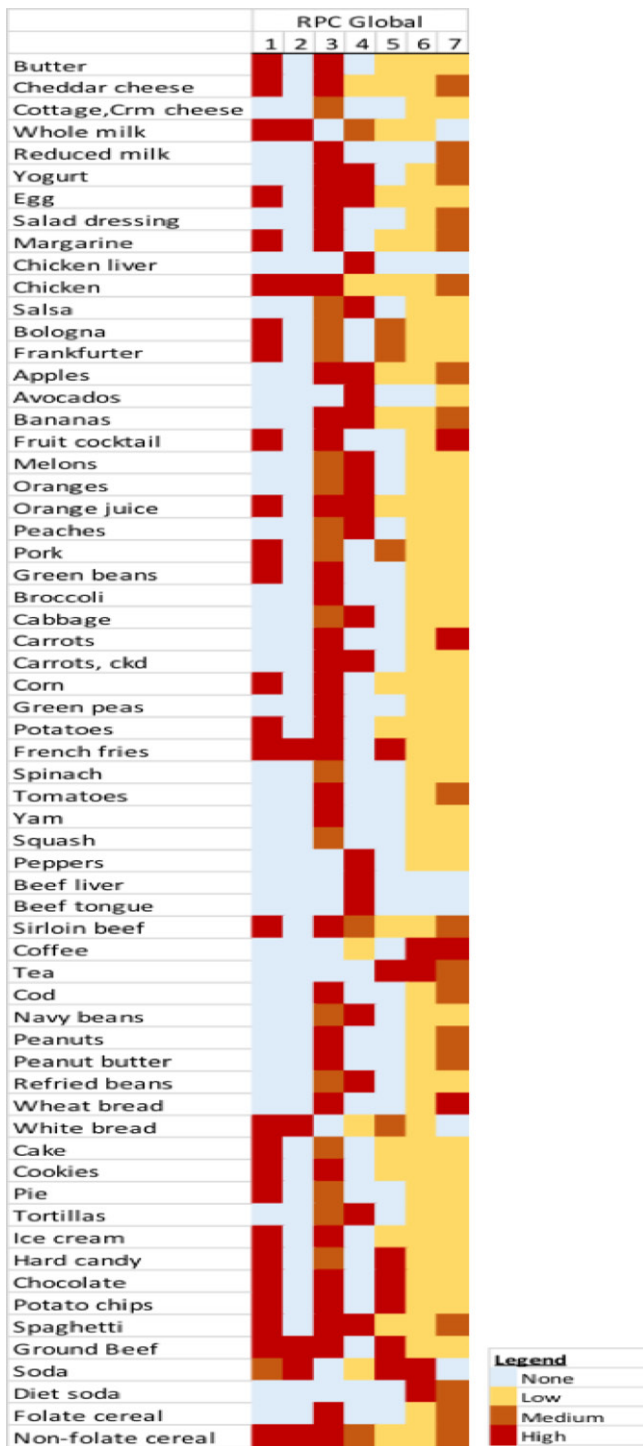
**Figure 1.** Heat map showing modal consumption level of RPC global profiles. Legend provides level of consumption for corresponding color.

This implies that subjects either did consume according to their global profile pattern, or did not consume that food at all. Four foods had a tendency to not deviate to any of the local clusters, and remain in accordance with the global profiles (pork, French fries, potato chips, wheat bread). All other food variables with varying consumption patterns by state, are illustrated in Figure 4. At the local level, foods aggregated to a single cluster

$(\lambda^{(s)} > 0.95)$, indicating less variability of consumption patterns within each subpopulation.

States with higher representation of Latino populations (Texas and California) showed a tendency for higher consumption of foods commonly found in a Latino diet (tortillas, refried beans, salsa, avocados), whereas states such as Massachusetts, New Jersey, and New York showed a tendency toward lower consumption of those foods. Another highly discriminating food was spaghetti, which was consumed at a high level for those subjects in global profile 3. This was the most populous global profile in North Carolina and Utah, yet subjects from those states were likely to deviate in favor of a lower consumption level, respectively. Similarly, subjects from Massachusetts and New Jersey who were not allocated to that global profile (i.e., allocated instead to global profile 2 and global profile 5) still favored a high consumption of spaghetti. Utah subjects had a tendency to deviate in favor of a high consumption of child-friendly fruits (apples, bananas, fruit cocktail) and dairy (egg, reduced fat milk).

### 4.3.2. Comparisons With Other Models

We also analyzed the NBDPS data using the previous methods. In general, these models had an exhaustive number of clusters that shared patterns reflected in both the global and local clusters of the RPC. Using the posterior median estimates, NBDPS participants were classified into the cluster with the highest posterior allocation probability. The mean and 95% credible interval of these probabilities is summarized in Table 2. The LCA-4 model had the strongest allocation probabilities, and the DPM model had the weakest.

Despite strong allocation probabilities, overgeneralization was a major limitation of the LCA-4 model. Over 20 foods were identified as favoring no consumption across all four classes. For example, pie favored no consumption for all cluster patterns (Appendix C, Figure C.1). Yet, the RPC model showed that this tendency for nonconsumption was only prominent in California, Massachusetts, New Jersey, and Utah (Figure 4). The remaining subpopulations favored corresponding global profile patterns which indicated a nonconsumption in only three of the seven profiles. Both the OFM and DPM models derived an excessive number of clusters ($K_{OFM} = 15$, $K_{DPM} = 16$). In the DPM model, only one derived cluster was larger than 5% in size. Like the LCA-4 model, the OFM modal cluster pattern indicated a general nonconsumption of nine foods that was not reflected in the RPC model. For example, the consumption of yams by the state of New York is masked in the OFM model. This consumption is relevant to New York participants that belong to global profiles 3, 6, or 7 that favor a high or low consumption of yams. The LPP$_2$ was the only model that allows the global clustering assumption to be relaxed with global and local clustering. A total of nine nonempty clusters were derived in the LPP$_2$, but none of the foods in the dataset favored a deviation away from the global profile patterns ($\nu_j > 0.99$ : for all $j \in 1, \ldots, p$). By allowing deviations by the known subpopulation, the RPC was able to clearly identify and distinguish deviating behaviors from the global profile trends.
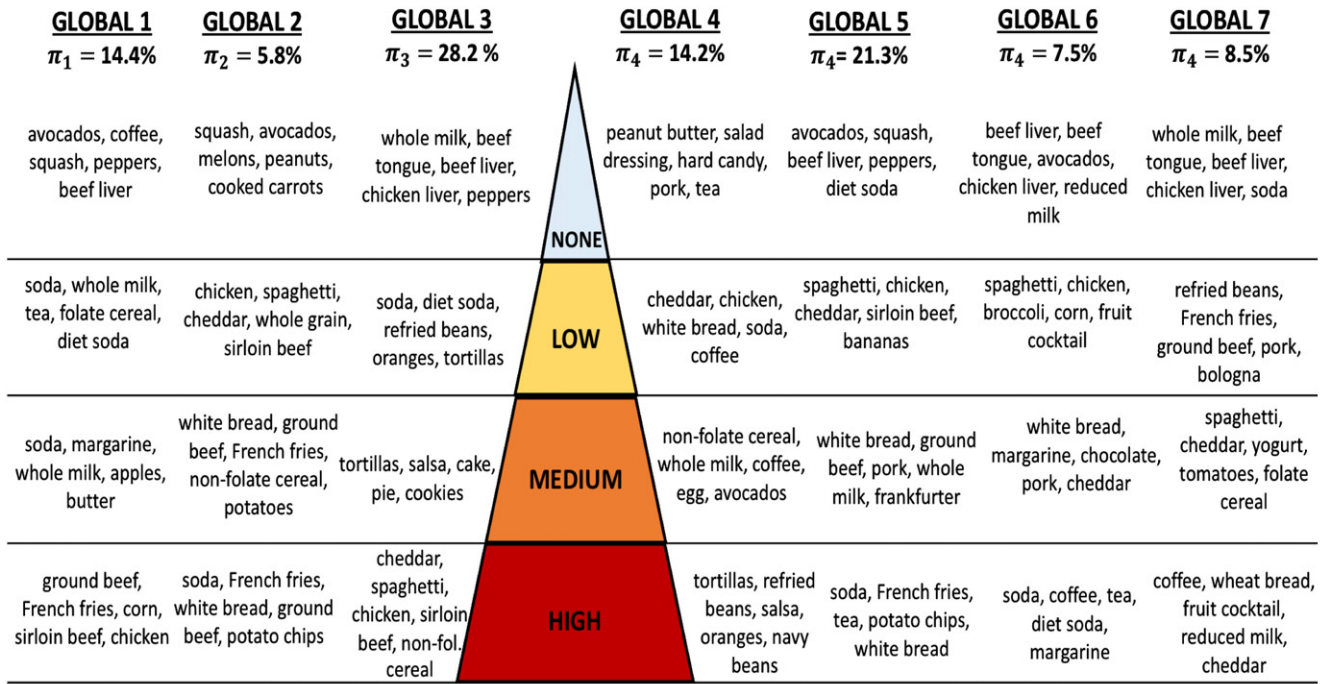
| GLOBAL 1<br>$\pi_1 = 14.4\%$ | GLOBAL 2<br>$\pi_2 = 5.8\%$ | GLOBAL 3<br>$\pi_3 = 28.2\%$ | GLOBAL 4<br>$\pi_4 = 14.2\%$ | GLOBAL 5<br>$\pi_4 = 21.3\%$ | GLOBAL 6<br>$\pi_4 = 7.5\%$ | GLOBAL 7<br>$\pi_4 = 8.5\%$ |
|---|---|---|---|---|---|---|
| avocados, coffee, squash, peppers, beef liver | squash, avocados, melons, peanuts, cooked carrots | whole milk, beef tongue, beef liver, chicken liver, peppers | **NONE** — peanut butter, salad dressing, hard candy, pork, tea | avocados, squash, beef liver, peppers, diet soda | beef liver, beef tongue, avocados, chicken liver, reduced milk | whole milk, beef tongue, beef liver, chicken liver, soda |
| soda, whole milk, tea, folate cereal, diet soda | chicken, spaghetti, cheddar, whole grain, sirloin beef | soda, diet soda, refried beans, oranges, tortillas | **LOW** — cheddar, chicken, white bread, soda, coffee | spaghetti, chicken, cheddar, sirloin beef, bananas | spaghetti, chicken, broccoli, corn, fruit cocktail | refried beans, French fries, ground beef, pork, bologna |
| soda, margarine, whole milk, apples, butter | white bread, ground beef, French fries, non-folate cereal, potatoes | tortillas, salsa, cake, pie, cookies | **MEDIUM** — non-folate cereal, whole milk, coffee, egg, avocados | white bread, ground beef, pork, whole milk, frankfurter | white bread, margarine, chocolate, pork, cheddar | spaghetti, cheddar, yogurt, tomatoes, folate cereal |
| ground beef, French fries, corn, sirloin beef, chicken | soda, French fries, white bread, ground beef, potato chips | cheddar, spaghetti, chicken, sirloin beef, non-fol. cereal | **HIGH** — tortillas, refried beans, salsa, oranges, navy beans | soda, French fries, tea, potato chips, white bread | soda, coffee, tea, diet soda, margarine | coffee, wheat bread, fruit cocktail, reduced milk, cheddar |

**Figure 2.** Foods most likely to be consumed at each consumption level of global profiles. Level of consumption decreases as you move from bottom to top.
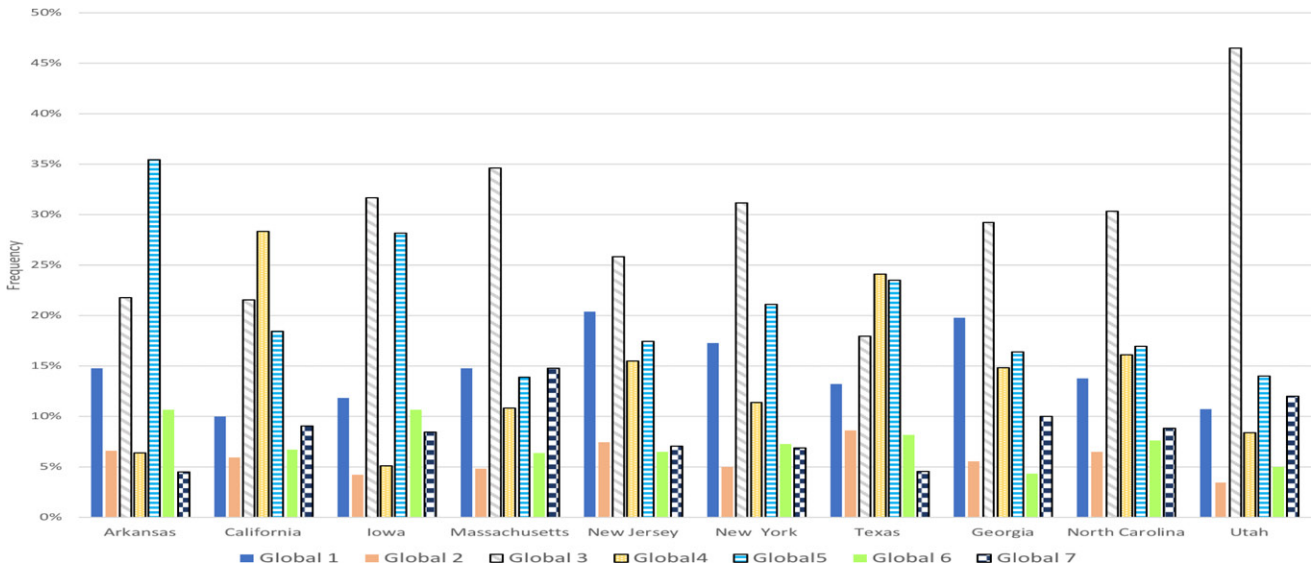


**Figure 3.** Frequency distribution of global profiles by subpopulation (state) for the global profiles identified in the model. Global profiles are displayed in numerical order from left to right, with corresponding colors indicated in the legend below bar plots.

## 5. Discussion

The RPC method provides a convenient and informative population-based model that is able to adapt and account for potential deviations occurring within subpopulations. It creates a practical solution to the global clustering problem by allowing certain variables to assume a distribution separate from their assigned overall cluster. This separate distribution is unique and can be shared among participants belonging to the same subpopulation regardless of the overall cluster assigned to those participants. A logistic regression model could be placed on relevant probabilities as desired to include covariates.

The overfitted mixture framework of the RPC is sensitive to the selection of Dirichlet hyperpriors. Smaller Dirichlet hyperprior values can slow down the rate of cluster growth (Rousseau and Mengersen 2011). Our selection of $\frac{1}{K}$ as the hyperprior provided a modest rate of growth for our large sample population size. Other more informative or even nonsymmetric hyperpriors can be used as an alternative, as deemed appropriate. Additionally, the overfitted finite mixture model is still prone to redundancies in the cluster profile patterns. This often can require an additional step of post-processing to examine the final profile results and remove redundant patterns as they are expressed in the sampler. An alternative would be to use a
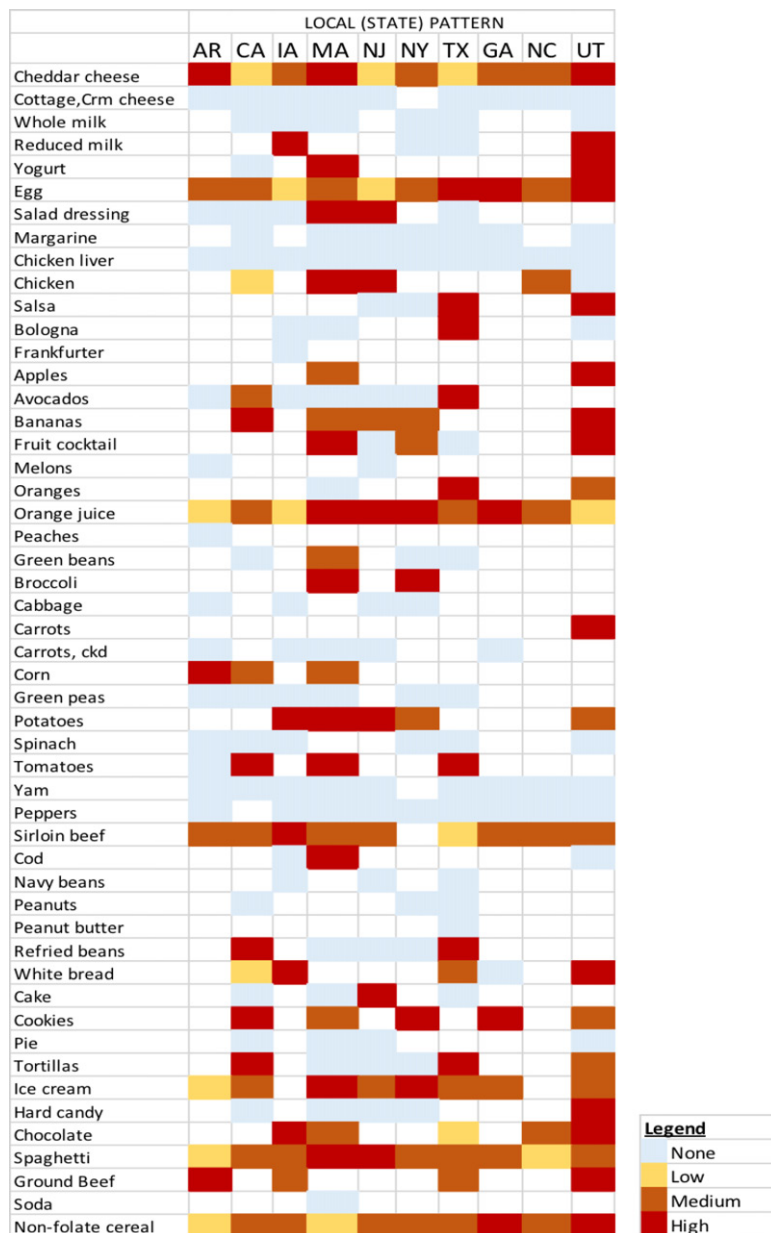
**Figure 4.** Heat map illustrating foods with a tendency to deviate from the global profile ($v_j^{(s)} < 0.5$), by subpopulation. Legend provides level of consumption for corresponding color.

hierarchical hybrid Dirichlet that can handle discrete data, if a more complex approach is desired. This extension is a different potential direction that to our knowledge has not yet been pursued.

Nevertheless, RPC allows researchers to characterize behaviors attributable to a general population or isolated within a subpopulation. In the application using NBDPS dietary data, numerous regional food trends were identified using the RPC model that were not apparent using other approaches.

**Table 2.** Summary comparing number of nonempty clusters and mean posterior probability of NBDPS assigned cluster 95% credible interval of maximum posterior probability of subject allocation to preferred cluster.

| | Nonempty clusters ($K_0$) | Posterior probability of assigned cluster Mean (95% credible interval) |
|---|---|---|
| tLCA | 4 | 0.999 (0.56, 1.00) |
| DPM | 16 | 0.934 (0.44, 1.00) |
| oFMM | 15 | 0.974 (0.49, 1.00) |
| LPP$_2$ | 9 | 0.964 (0.44, 1.00) |
| RPC | 7 | 0.999 (0.52, 1.00) |

## Appendix A: Simulation Study Details

Global and local cluster patterns were generated by applying a probability of 0.7 to the expected true response level and a probability of 0.1 to all other response categories. Let $r^*$ denote the expected response level for a given global cluster. If $r^* = 3$ for variable $j = 1$, then $(\theta_{011,1}, \ldots, \theta_{011,d}) = (0.1, 0.1, 0.7, 0.1)$. A table of expected location of $r^*$ mode for each global cluster profile is provided below.

## Case 1: Strictly Global

Each subject $i$ belongs to one of $S = 4$ subpopulations($s_i \in 1, 2, 3, 4$). Within that subpopulation, each subject is assigned to one of 3 global profiles ($C_i \in 1, 2, 3$). Probability of variable deviation is preset to $v_j^{(s)} = 1$, such that $G_{ij} = 1$ for all subjects $i \in (1, \ldots, n)$, and all variables $j \in (1, \ldots, p)$. Expected global profile pattern, defined by modal response, is provided in Table A.1.

1Subject-level observed data was generated as follows:

---
**Algorithm 1** Simulation case 1.

---
1: Generate $y_{ij} | C_i, s_i \sim \text{Multinomial}(\theta_{0jC_i,1}, \ldots, \theta_{0jC_i,d})$

---

## Case 2: Strictly Local

Each subject $i$ belongs to one of $S = 8$ subpopulations ($s_i \in 1, 2, \ldots, 8$). Each subpopulation contains one local cluster. Probability of variable deviation is preset to $v_j^{(s)} = 0$, such that $G_{ij} = 0$ for all subjects $i \in (1, \ldots, n)$, and all variables $j \in (1, \ldots, p)$. Expected local profile pattern, defined by modal response, is provided in Table A.2.

Subject-level observed data was generated as follows:

---
**Algorithm 2** Simulation case 2.

---
Generate $y_{ij} | L_{ij}, s_i \sim \text{Multinomial}(\theta_{1jL_{ij},1}^{(s_i)}, \ldots, \theta_{1jL_{ij},d}^{(s_i)})$

---

## Case 3: Hybrid

Each subject $i$ belongs to one of $S = 4$ subpopulations ($s_i \in 1, 2, 3, 4$). Within each subpopulation, each subject is assigned to one of two local clusters, $L_{ij} = l \in (1, 2)$ and one of three global clusters, $C_i = k \in (1, 2, 3)$. Expected global profile patterns, defined by modal response, are provided in Table A.1, when $G_{ij} = 1$. Expected local profile patterns, defined by modal response, are provided in Table A.2, when $G_{ij} = 0$. Probability of allocation to global or local was fixed for all variables within a subpopulation:

Subpopulation 1: $v_j^{(1)} = 0.25$ for all $j \in (1, \ldots, p)$, local profiles 1 and 2 (Table A.2)

Subpopulation 2: $v_j^{(2)} = 0.50$ for all $j \in (1, \ldots, p)$, local profiles 3 and 4 (Table A.2)

Subpopulation 3: $v_j^{(3)} = 0.75$ for all $j \in (1, \ldots, p)$, local profiles 5 and 6 (Table A.2)

Subpopulation 4: $v_j^{(4)} = 1.00$ for all $j \in (1, \ldots, p)$, local profiles 7 and 8 (Table A.2)

---
**Algorithm 3** Simulation case 3.

---
Generate $G_{ij} | s_i \sim \text{Bernoulli}(v_j^{(s_i)})$
  **if** $G_{ij} = 1$ **then**
3:    Generate $y_{ij} | C_i \sim \text{Multinomial}(\theta_{0jC_i,1}, \ldots, \theta_{0jC_i,d})$
  **else if** $G_{ij} = 0$ **then**
    Generate $y_{ij} | L_{ij}, s_i \sim \text{Multinomial}(\theta_{1jL_{ij},1}^{(s_i)}, \ldots, \theta_{1jL_{ij},d}^{(s_i)})$
6: **end if**

---

## Case 4: Null Set

Each subject $i$ was assigned to one of $S = 4$ subpopulations ($s_i \in 1, 2, 3, 4$). Subject-level observed data was randomly assigned from a discrete uniform distribution containing $d = 4$ possible response levels.

---
**Algorithm 4** Simulation case 4.

---
Generate $y_{ij} \sim \text{Uniform}(1, \ldots, d)$

---

## Case 5: Mock NBDPS

Each subject $i$ was assigned to one of $S = 10$ subpopulations ($s_i \in 1, 2, \ldots, 10$). Within each subpopulation, each subject is assigned to a global profile, labelled $C_i$, and each subpopulation contained at least one of 28 unique profiles containing a combination of global and local variables. Subject-level observed data was generated as referenced. Subpopulations 1–3, and 8 contained 1200 subjects, with 400 subjects equally assigned to one of three global profiles. Each of these subpopulations contained 2 local profiles with subjects evenly assigned to each profile (600 subjects per local profile for subpopulations 1–3; 400 subjects per local profile for subpopulation 8). Subpopulations 4–6, and 9 contained 1200 subjects with 400 subjects equally assigned to one of the three global profiles (400 subjects per global profile). Subpopulation 10 contained 1600 subjects, where 400 subjects contained one global profile and no local profile and 1200 subjects with one local profile and no global profile. Expected global profile patterns, defined by modal response, are provided in Table A.1, when $G_{ij} = 0$. Expected local profile patterns, defined by modal response, are provided in Table A.3, when $G_{ij} = 1$.

---
**Algorithm 5** Simulation case 5.

---
Subpopulation 1: Generate $y_{ij}$ as in Algorithm 3, where $v_j^{(s)} = 0.75$.
Subpopulation 2: Generate $y_{ij}$ as in Algorithm 3, where $v_j^{(s)} = 0.50$.
Subpopulation 3: Generate $y_{ij}$ as in Algorithm 3, where $v_j^{(s)} = 0.25$.
Subpopulation 4–6: Generate $y_{ij}$ as in Algorithm 1.
5: Subpopulation 7: Generate $y_{ij}$ as in Algorithm 2.
Subpopulation 8: Generate $y_{ij} | C_i = 1$ as in Algorithm 2 and $y_{ij} | C_i = 2, 3$ as in Algorithm 3, where $v_j^{(s)} = 0.9$.
Subpopulation 9: Generate $y_{ij}$ as in Algorithm 3, where $v_j^{(s)} = 0.1$.
Subpopulation 10: Generate $y_{ij} | C_i = 1$ as in Algorithm 1 and $y_{ij} | C_i = 2$ as in Algorithm 2.

---

## Case 6: Violations of Beta-Bernoulli assumption

Each subject $i$ belongs to one of $S = 4$ subpopulations ($s_i \in 1, 2, 3, 4$). Within each subpopulation, each subject is assigned to one of two local clusters, $L_{ij} = l \in (1, 2)$ and one of three global clusters, $C_i = k \in (1, 2, 3)$. Expected global profile patterns, defined by modal response, are provided in Table A.1, when $G_{ij} = 1$. Expected local profile patterns, defined by modal response, are designed as in Case 3, with the 8 local profile patterns distributed across the four subpopulations.

**Table A.1.** Location of mode for each respective global cluster where maximal probability ($\theta_{0jC_i,r^*} = 0.7$).

| $j$ | Global 1 | Global 2 | Global 3 | $j$ | Global 1 | Global 2 | Global 3 |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 2 | 1 | 26 | 1 | 4 | 2 |
| 2 | 3 | 2 | 1 | 27 | 1 | 4 | 2 |
| 3 | 3 | 2 | 1 | 28 | 1 | 4 | 2 |
| 4 | 3 | 2 | 1 | 29 | 1 | 4 | 2 |
| 5 | 3 | 2 | 1 | 30 | 1 | 4 | 2 |
| 6 | 3 | 2 | 1 | 31 | 1 | 4 | 3 |
| 7 | 3 | 2 | 1 | 32 | 1 | 4 | 3 |
| 8 | 3 | 2 | 1 | 33 | 1 | 4 | 3 |
| 9 | 3 | 2 | 1 | 34 | 1 | 4 | 3 |
| 10 | 3 | 2 | 1 | 35 | 1 | 4 | 3 |
| 11 | 3 | 4 | 2 | 36 | 1 | 4 | 3 |
| 12 | 3 | 4 | 2 | 37 | 1 | 4 | 3 |
| 13 | 3 | 4 | 2 | 38 | 1 | 4 | 3 |
| 14 | 3 | 4 | 2 | 39 | 1 | 4 | 3 |
| 15 | 3 | 4 | 2 | 40 | 1 | 4 | 3 |
| 16 | 3 | 4 | 2 | 41 | 1 | 4 | 3 |
| 17 | 3 | 4 | 2 | 42 | 1 | 4 | 3 |
| 18 | 3 | 4 | 2 | 43 | 1 | 4 | 3 |
| 19 | 3 | 4 | 2 | 44 | 1 | 4 | 3 |
| 20 | 3 | 4 | 2 | 45 | 1 | 4 | 3 |
| 21 | 3 | 4 | 2 | 46 | 1 | 4 | 3 |
| 22 | 3 | 4 | 2 | 47 | 1 | 4 | 3 |
| 23 | 3 | 4 | 2 | 48 | 1 | 4 | 3 |
| 24 | 3 | 4 | 2 | 49 | 1 | 4 | 3 |
| 25 | 3 | 4 | 2 | 50 | 1 | 4 | 3 |

Probability of deviation to global or local level for all variables within a subpopulation was drawn from a different subpopulation:

---

**Algorithm 6** Simulation case 6.

> **if** $s_i = 1$ **then**
>> Randomly draw $v_j^{(1)} \sim \text{Beta}(2, 1)$ for all $j \in (1, \ldots, p)$
>
> **else if** $s_i = 2$ **then**
>> Randomly draw $v_j^{(2)} \sim U(0, 1)$ for all $j \in (1, \ldots, p)$
>
> **else if** $s_i = 3$ **then**
> 6:   Randomly draw $v_j^{(3)} \sim P(N(0, 1))$ for all $j \in (1, \ldots, p)$
>
> **else if** $s_i = 4$ **then**
>> Randomly draw $v_j^{(3)}$ from a standard Cauchy distribution for all $j \in (1, \ldots, p)$
>
> **end if**
> **if** $G_{ij} = 1$ **then**
>> Generate $y_{ij}|C_i \sim \text{Multinomial}(\theta_{0jC_i,1}, \ldots, \theta_{0jC_i,d})$
>
> 12: **else if** $G_{ij} = 0$ **then**
>> Generate $y_{ij}|L_{ij}, s_i \sim \text{Multinomial}(\theta_{1jL_{ij},1}^{(s_i)}, \ldots, \theta_{1jL_{ij},d}^{(s_i)})$
>
> **end if**

---

### *Case 7: Continuous Data*

Each subject $i$ belongs to one of $S = 2$ subpopulations ($s_i \in 1, 2$) of size 1500 subjects each, describing $p = 30$ variables. Within each subpopulation, each subject is assigned to one of two local clusters, $L_{ij} = l \in (1, 2)$ and one of two global clusters, $C_i = k \in (1, 2)$. Allocation to global or local level for all variables within a subpopulation was preset such that following variables were allocated from the local density from subpopulation 1: vars (6, 7, 16, 17, 26, 27); subpopulation

2: vars (5, 8, 15, 18, 25, 28). Subject-level observed data for global and local profiles are drawn from a normal distribution with respective means, $\mu_0, \mu_1^{(s)}$. Standard deviations remained uniform depending on the case: (a) $\sigma = 0.1$, (b) $\sigma = 1.0$, (c) $\sigma = 3.0$.

---

**Algorithm 7** Simulation case 7.

> Generate $G_{ij}|s_i \sim \text{Bernoulli}(v_j^{(s_i)})$
> **if** $G_{ij} = 1$ **then**
>> Generate $y_{ij}|C_i \sim N(\mu_0, \sigma)$
>
> **else if** $G_{ij} = 0$ **then**
>> Generate $y_{ij}|L_{ij}, s_i \sim N(\mu_1^{(s)}, \sigma)$
>
> **end if**

---

## Appendix B: Supporting Figures From Mock NBDPS Simulation Study (Section 3.2)

Using one of the replicates from the mock NBDPS simulations (Section 3.2), Figure B.1 provides a comparative illustration of variables that were identified as deviating from the global cluster pattern by each subpopulation. The figure on the left shows the variables that deviated as a result of the RPC model and the figure on the right illustrates the true deviated variables. Discrepancies in the predicted and actual results were found in subpopulation 7. This subpopulation contained patterns where all variables deviated from the global pattern and contained two unique local clusters. The RPC model misspecified the completely local pattern as a global pattern for subpopulation 7. Most of the deviation was identified as deviating to local for subpopulation 10. However, some locally allocated variables overlapped with the globally expected value and were therefore considered global in those cases.

**Table A.2.** Location of maximal probability for each variable in respective local cluster where maximal probability ($\theta^{(s)}_{1jL_{ij},r^*} = 0.7$).

| j | Local 1 | Local 2 | Local 3 | Local 4 | Local 5 | Local 6 | Local 7 | Local 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 2 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 3 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 4 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 5 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 6 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 7 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 8 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 9 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 10 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 11 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 12 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 13 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 14 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 15 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 16 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 17 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 18 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 19 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 20 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 21 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 22 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 23 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 24 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 25 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 26 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 27 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 28 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 29 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 30 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 31 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 32 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 33 | 1 | 1 | 2 | 2 | 3 |  | 4 | 4 |
| 34 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 35 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 36 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 37 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 38 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 39 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 40 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 41 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 42 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 43 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 44 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 45 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 46 | 1 | 2 | 2 | 4 | 3 | 1 | 4 | 3 |
| 47 | 1 | 3 | 2 | 1 | 3 | 4 | 4 | 2 |
| 48 | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 |
| 49 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 50 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |

**Table A.3.** Case 5: Location of maximal probability for each variable in respective local cluster, where $(\theta^{(s)}_{1jL_{ij},r*} = 0.7)$.

| | Subpop 1 | | Subpop 2 | | Subpop 3 | | Subpop 7 | | Subpop 8 | | Subpop 9 | | Subpop 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $j$ | Local 1 | Local 2 | Local 1 | Local 2 | Local 1 | Local 2 | Local 1 | Local 2 | Local 1 | Local 2 | Local 1 | Local 2 | Local 1 |
| 1 | 2 | 2 | 3 | 2 | | | 4 | 2 | | | 2 | 2 | 4 |
| 2 | 1 | 4 | 1 | 4 | 2 | 2 | 1 | 4 | | | 3 | 1 | 4 |
| 3 | | | | | 4 | 3 | 1 | 3 | 1 | 2 | | | 3 |
| 4 | 3 | 3 | 3 | 1 | | | 2 | 1 | | | 1 | 1 | 3 |
| 5 | | | | | 1 | 3 | 4 | 2 | | | 4 | 2 | 4 |
| 6 | | | 2 | 2 | 3 | 2 | 3 | 4 | | | 2 | 1 | 2 |
| 7 | | | | | | | 3 | 3 | | | 3 | 1 | 3 |
| 8 | | | 3 | 2 | 2 | 1 | 3 | 4 | | | 2 | 1 | 2 |
| 9 | | | 3 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 4 | 2 |
| 10 | | | 4 | 2 | 4 | 4 | 1 | 1 | | | 3 | 4 | 2 |
| 11 | | | 2 | 2 | | | 4 | 4 | | | 2 | 3 | 3 |
| 12 | | | 4 | 4 | 4 | 3 | 4 | 3 | | | 1 | 4 | 3 |
| 13 | | | 2 | 1 | 4 | 1 | 4 | 2 | | | 2 | 1 | 4 |
| 14 | | | 3 | 2 | 3 | 3 | 1 | 1 | 2 | 4 | 2 | 3 | 4 |
| 15 | | | 3 | 1 | 4 | 3 | 4 | 3 | | | 3 | 2 | 3 |
| 16 | | | 1 | 1 | 1 | 3 | 2 | 3 | | | 2 | 2 | 2 |
| 17 | | | 1 | 1 | 2 | 2 | 3 | 4 | | | 1 | 3 | 3 |
| 18 | | | 4 | 3 | 4 | 1 | 1 | 1 | | | 1 | 4 | 2 |
| 19 | 2 | 4 | 4 | 3 | | | 4 | 4 | | | 4 | 1 | 4 |
| 20 | | | 4 | 1 | 1 | 1 | 3 | 3 | | | | | 2 |
| 21 | | | | | 1 | 1 | 1 | 1 | | | 3 | 1 | 2 |
| 22 | 3 | 4 | | | 4 | 4 | 4 | 3 | | | 3 | 3 | 2 |
| 23 | 1 | 4 | 2 | 4 | 1 | 3 | 3 | 1 | | | 3 | 3 | 2 |
| 24 | | | 3 | 1 | 2 | 3 | 4 | 2 | | | 4 | 4 | 3 |
| 25 | 2 | 2 | | | 1 | 2 | 3 | 3 | | | 4 | 4 | 3 |
| 26 | | | 3 | 4 | 2 | 1 | 3 | 1 | | | 3 | 3 | 4 |
| 27 | | | | | 1 | 4 | 2 | 4 | | | 3 | 3 | 2 |
| 28 | | | | | 3 | 3 | 2 | 3 | | | 4 | 1 | 3 |
| 29 | | | 3 | 4 | 3 | 4 | 4 | 1 | | | 2 | 2 | 3 |
| 30 | | | 4 | 3 | 4 | 1 | 1 | 1 | | | 2 | 4 | 2 |
| 31 | | | | | 4 | 4 | 4 | 1 | 2 | 1 | | | 4 |
| 32 | | | | | 4 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 1 |
| 33 | | | 2 | 2 | 3 | 2 | 2 | 1 | | | 4 | 2 | 4 |
| 34 | | | | | 4 | 1 | 4 | 1 | | | 1 | 1 | 1 |
| 35 | 4 | 4 | | | 2 | 1 | 1 | 1 | 4 | 4 | 2 | 4 | 4 |
| 36 | | | 1 | 2 | 3 | 2 | 2 | 1 | | | 1 | 2 | 4 |
| 37 | | | 1 | 1 | 1 | 2 | 1 | 2 | | | 3 | 4 | 2 |
| 38 | | | 4 | 4 | | | 2 | 2 | | | 4 | 2 | 2 |
| 39 | | | 2 | 1 | 1 | 4 | 4 | 2 | | | 1 | 2 | 2 |
| 40 | | | | | 2 | 1 | 2 | 3 | | | 4 | 2 | 4 |
| 41 | | | 2 | 4 | 2 | 4 | 3 | 3 | | | 3 | 1 | 1 |
| 42 | | | | | 3 | 2 | 3 | 1 | | | 4 | 1 | 3 |
| 43 | 1 | 2 | 4 | 2 | 3 | 1 | 4 | 2 | | | 3 | 1 | 3 |
| 44 | | | | | 1 | 4 | 3 | 3 | | | 3 | 1 | 1 |
| 45 | | | 3 | 4 | | | 4 | 4 | | | | | 4 |
| 46 | | | 2 | 2 | 3 | 4 | 3 | 1 | | | 1 | 2 | 2 |
| 47 | | | | | 1 | 3 | 3 | 4 | | | 3 | 2 | 2 |
| 48 | | | | | 4 | 1 | 3 | 4 | | | 2 | 4 | 2 |
| 49 | 1 | 3 | | | 1 | 1 | 3 | 2 | | | 3 | 1 | 3 |
| 50 | 2 | 3 | 1 | 4 | 4 | 4 | 1 | 3 | | | | | 4 |

**Table A.4.** Simulation set-up for each subcase for continuous set from a $N(\mu,\sigma)$ distribution.

| Case | Global density $N(\mu_0,\sigma)$ | Local density $N(\mu_1^{(s)},\sigma)$ | |
|---|---|---|---|
| 7a: Completely separate | $N(-9,0.1)$ | $1: N(5,0.1)$ | $N(9,0.1)$ |
| | $N(2,0.1)$ | $2: N(-5,0.1)$ | $N(-2,0.1)$ |
| 7b: Partially separate | $N(-9,1)$ | $1: N(5,1)$ | $N(9,1)$ |
| | $N(2,1)$ | $N(-5,1)$ | $N(-2,1)$ |
| 7c: Complete overlap | $N(-9,3)$ | $1: N(5,3)$ | $N(9,3)$ |
| | $N(2,3)$ | $2: N(-5,3)$ | $N(-2,3)$ |

NOTES: Each dataset contains 2 subpopulations with $p = 30$ variables. Each subpopulation contains information from 2 global and 2 local clusters.
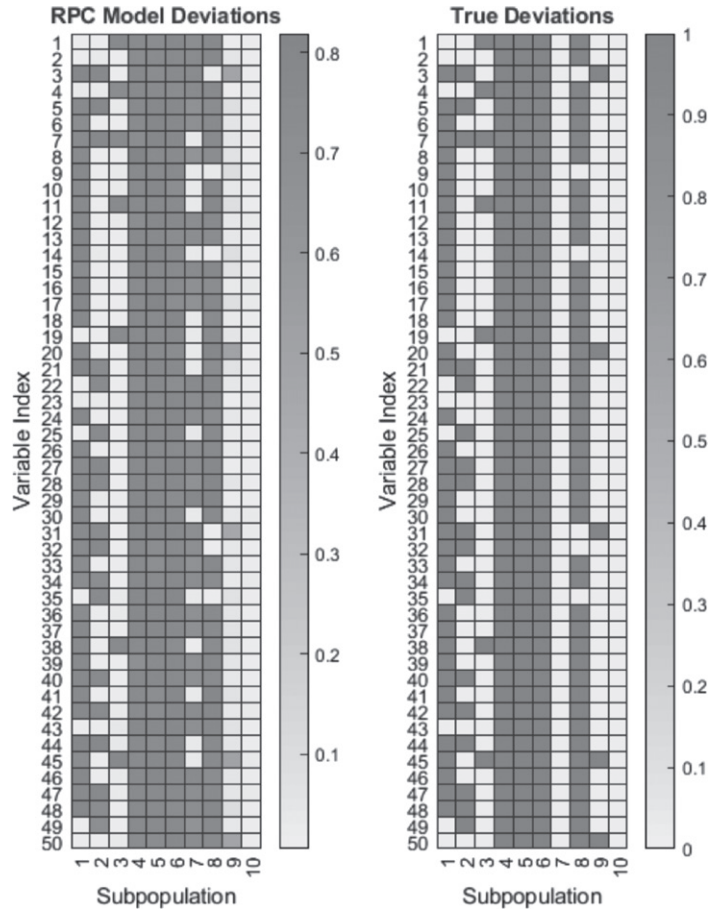
**Figure B.1.** Simulation example comparing probability of a variable being allocated to a global cluster set (gray, $G_{ij} = 1$) and local cluster set (white, $G_{ij} = 0$). Side-by-side figure compares RPC-derived results (left) with true deviation results (right).

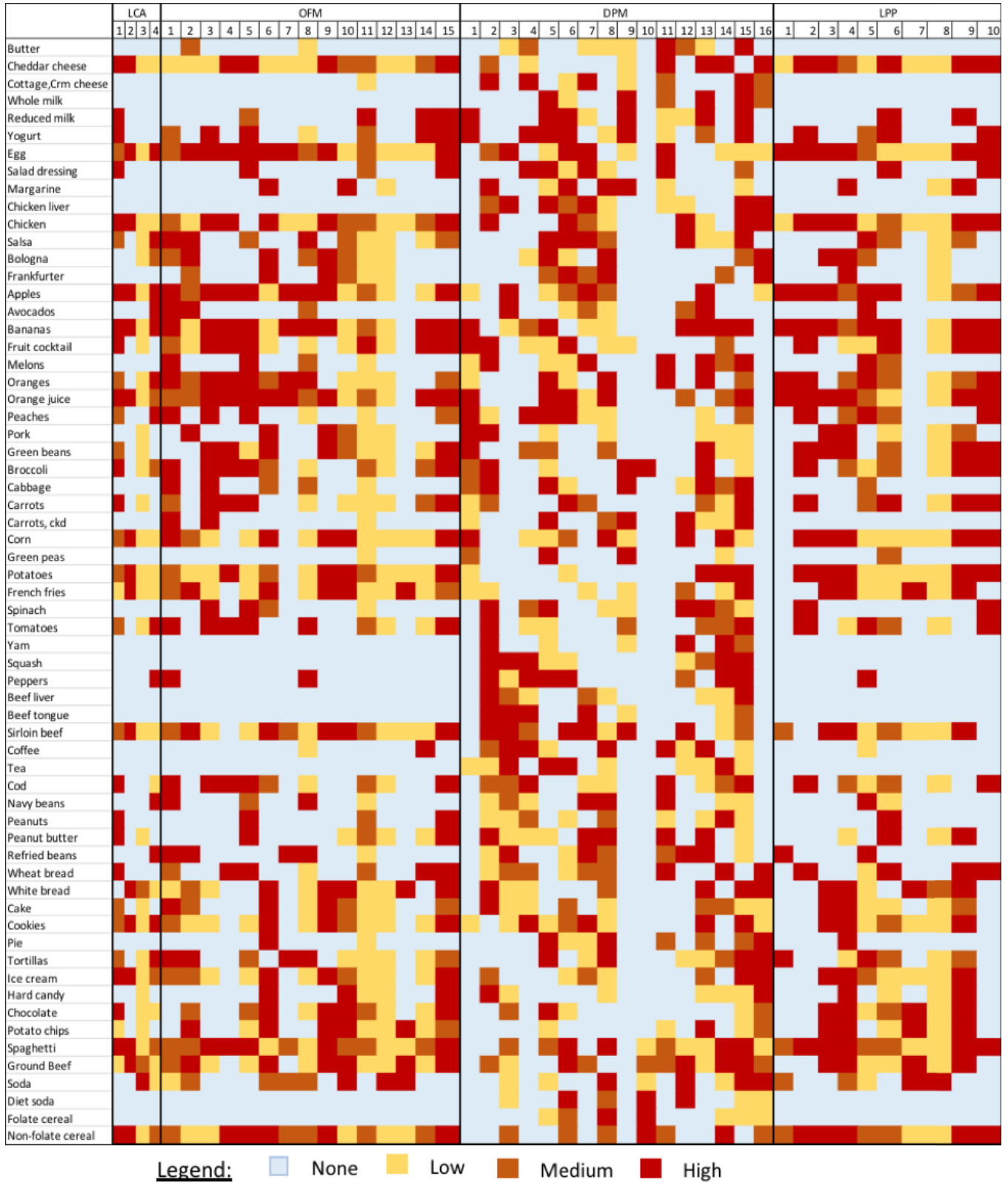## Appendix C: Supporting Figure for NBDPS Previous Methods (Section 4.3.2)



**Figure C.1.** Heat map showing modal consumption level of each of the nonempty clusters derived from the LCA-4, OFM-15, DPM-16, LPP$_2$-10. Legend provides level of consumption for corresponding color.

## Supplementary Materials

An example dataset to illustrate the RPC method all source code, data, and expected output are publicly available at *https://github.com/bjks10/RPC*.

**README file:** Text file overviewing MATLAB functions and files to perform RPC program.
**RPC source code:** MATLAB code to perform the RPC method and supporting files.
**Simulated dataset:** Simulated example dataset to illustrate RPC method.
**Expected Output:** Saved output results from running RPC example set.
**Posterior Computation:** MCMC Gibbs sampler for posterior computation of RPC

1. Update the global component indicators $(G_{ij} \mid s_i = s) \sim \text{Bern}(p_{ij})$, where

$$p_{ij} = \frac{\nu_j^{(s)} \prod_{r=1}^d \Theta_{0jC_i,r}^{\mathbf{1}(y_{ij}=r)}}{\nu_j^{(s)} \prod_{r=1}^d \Theta_{0jC_i,r}^{\mathbf{1}(y_{ij}=r)} + (1-\nu_j^{(s)}) \prod_{r=1}^d (\Theta_{1jL_{ij},r}^{(s)})^{\mathbf{1}(y_{ij}=r)}}$$

for each subject $i \in (1,\ldots,n)$ with respective subpopulation index $s$.

2. Update global cluster index $C_i$, $i = 1,\ldots,n$ from its multinomial distribution where

$$\Pr(C_i = h) = \frac{\pi_h \prod_{j:G_{ij}=1} \prod_{r=1}^d \Theta_{0jh,r}^{\mathbf{1}(y_{ij}=r)}}{\sum_{l=1}^K \pi_l \prod_{j:G_{ij}=1} \prod_{r=1}^d \Theta_{0jl,r}^{\mathbf{1}(y_{ij}=r)}}.$$

3. Update local cluster index $L_{ij}$ for all $i : s_i = s$ and $j = 1,\ldots,p$, repeating for each $s$, from its multinomial distribution conditional on $s_i = s$ where

$$\Pr(L_{ij} = h) = \frac{\lambda_h^{(s)} \prod_{r=1}^d \left(\Theta_{1jh,r}^{(s)}\right)^{\mathbf{1}(y_{ij}=r,G_{ij}=0)}}{\sum_{l=1}^K \lambda_l^{(s)} \prod_{r=1}^d \left(\Theta_{1jl,r}^{(s)}\right)^{\mathbf{1}(y_{ij}=r,G_{ij}=0)}}.$$

4. Update the global clustering weights

$$\pi = (\pi_1,\ldots,\pi_K)$$
$$\sim \text{Dir}\left(\frac{1}{K} + \sum_{i=1}^n \mathbf{1}(C_i = 1),\ldots,\frac{1}{K} + \sum_{i=1}^n \mathbf{1}(C_i = K)\right).$$

5. Update the local clustering weights in subpopulation $s$,

$$\lambda^{(s)} = \left(\lambda_1^{(s)},\ldots,\lambda_K^{(s)}\right)$$
$$\sim \text{Dir}\left(\frac{1}{K} + \sum_{i:s_i=s}\sum_{j=1}^p \mathbf{1}(L_{ij}=1),\ldots,\frac{1}{K}\right.$$
$$\left. + \sum_{i:s_i=s}\sum_{j=1}^p \mathbf{1}(L_{ij}=K)\right).$$

6. Update the multinomial parameters, where $\eta$ is a flat, symmetric Dirichlet hyperparameter preset at 1

$$\theta_{0jh,\cdot} \sim \text{Dir}\left(\eta + \sum_{i:G_{ij}=1,C_i=h} \mathbf{1}(y_{ij}=1),\ldots,\eta \right.$$
$$\left. + \sum_{i:G_{ij}=1,C_i=h} \mathbf{1}(y_{ij}=d)\right)$$

$$\theta_{1jh,\cdot}^{(s)} \sim \text{Dir}\left(\eta + \sum_{i:G_{ij}=0,L_{ij}=h,s_i=s} \mathbf{1}(y_{ij}=1),\ldots,\eta \right.$$
$$\left. + \sum_{i:G_{ij}=0,L_{ij}=h,s_i=s} \mathbf{1}(y_{ij}=d)\right)$$

7. Update $\nu_j^{(s)} \sim \text{Be}(1 + \sum_{i:s_i=s} G_{ij}, \beta^{(s)} + \sum_{i:s_i=s}(1 - G_{ij}))$.

8. Update Beta-Bernoulli hyperparameter: $\beta^{(s)} \sim \text{Ga}(a + p, b - \sum_{j=1}^p \log(1 - \nu_j^{(s)}))$.

## Acknowledgments

## Funding

## References

Dunson, D. B. (2009), "Nonparametric Bayes Local Partition Models for Random Effects," *Biometrika*, 96, 249–262. [2,3,4]

Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051. [3,4]

Figueiredo, M. A. T., and Jain, A. K. (2002), "Unsupervised Learning of Finite Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 381–396. [2]

Ford, J. D., Elhai, J. D., Connor, D. F., and Frueh, B. C. (2010), "Poly-Victimization and Risk of Posttraumatic, Depressive, and Substance Use Disorders and Involvement in Delinquency in a National Sample of Adolescents," *Journal of Adolescent Health*, 46, 545–552. [2]

Frühwirth-Schnatter, S. (2001), "Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models," *Journal of the American Statistical Association*, 96, 194–209. [3,6]

Haraldsdottir, J. (1993), "Minimizing Error in the Field: Quality Control in Dietary Surveys," *European Journal of Clinical Nutrition*, 47, 19–24. [6]

Hu, J., Reiter, J. P., and Wang, Q. (2018), "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data," *Bayesian Analysis*, 13, 183–200. [2]

Jønneland, A., Overvad, K., Haraldsdóttir, J., Bang, S., Ewertz, M., and Jensen, O. M. (1991), "Validation of a Semiquantitative Food Frequency Questionnaire Developed in Denmark," *International Journal of Epidemiology*, 20, 906–912. [6]

Kant, A. K. (2004), "Dietary Patterns and Health Outcomes," *Journal of the American Dietetic Association*, 104, 615–635. [1]

Keshteli, A. H., Feizi, A., Esmaillzadeh, A., Zaribaf, F., Feinle-Bisset, C., Talley, N. J., and Adibi, P. (2015), "Patterns of Dietary Behaviours Identified by Latent Class Analysis Are Associated With Chronic Uninvestigated Dyspepsia," *British Journal of Nutrition*, 113, 803–812. [1]

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, L. S. (2009), "Modeling Data With Excess Zeros and Measurement Error: Application to Evaluating Relationships Between Episodically Consumed Foods and Health Outcomes," *Biometrics*, 65, 1003–1010. [6]

Krebs, C. J. (1989), "Ecological Methodology," Technical Report, Harper & Row, New York. [4,6]

Lazarsfeld, P., and Henry, N. (1968), *Latent Structure Analysis*, Mifflin: Houghton. [1,3]

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016), "Model-Based Clustering Based on Sparse Finite Gaussian Mixtures," *Statistics and Computing*, 26, 303–324. [4]

Medvedovic, M., and Sivaganesan, S. (2002), "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatics*, 18, 1194–1206. [4,6]

Miller, J. W., and Harrison, M. T. (2013), "A Simple Example of Dirichlet Process Mixture Inconsistency for the Number of Components," in *Advances in Neural Information Processing Systems*, pp. 199–206. [2]

——— (2018), "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356. [2]

Motulsky, A. G. (1989), *Diet and Health: Implications for Reducing Chronic Disease Risk*, Washington, DC: National Academies. [1]

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007), "Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study," *Structural Equation Modeling*, 14, 535–569. [2,4]

Papaspiliopoulos, O., and Roberts, G. O. (2008), "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models," *Biometrika*, 95, 169–186. [4]

Petrone, S., Guindani, M., and Gelfand, A. E. (2009), "Hybrid Dirichlet Mixture Models for Functional Data," *Journal of the Royal Statistical Society*, Series B, 71, 755–782. [2]

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008), "The Nested Dirichlet Process," *Journal of the American Statistical Association*, 103, 1131–1154. [2]

Rousseau, J., and Mengersen, K. (2011), "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models," *Journal of the Royal Statistical Society*, Series B, 73, 689–710. [2,3,4,8]

Silverwood, R. J., Nitsch, D., Pierce, M., Kuh, D., and Mishra, G. D. (2011), "Characterizing Longitudinal Patterns of Physical Activity in Mid-adulthood Using Latent Class Analysis: Results From a Prospective Cohort Study," *American Journal of Epidemiology*, 174, 1406. [2]

Smith, M. S., and Khaled, M. A. (2012), "Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation," *Journal of the American Statistical Association*, 107, 290–303. [2]

Sotres-Alvarez, D., Herring, A. H., and Siega-Riz, A. M. (2010), "Latent Class Analysis Is Useful to Classify Pregnant Women Into Dietary Patterns," *The Journal of Nutrition*, 140, 2253–2259. [1]

Sotres-Alvarez, D., Siega-Riz, A. M., Herring, A. H., Carmichael, S. L., Feldkamp, M. L., Hobbs, C. A., Olshan, A. F., and National Birth Defects Prevention Study (2013), "Maternal Dietary Patterns Are Associated With Risk of Neural Tube and Congenital Heart Defects," *American Journal of Epidemiology*, 177, 1279–1288. [6]

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001), "Comparative Validation of the Block, Willett, and National Cancer Institute Food Frequency Questionnaires the Eating at America's Table Study," *American Journal of Epidemiology*, 154, 1089–1099. [1,2]

Teh, Y. W. (2006), "A Hierarchical Bayesian Language Model Based on Pitman–Yor Processes," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 985–992. [2]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [2]

van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015), "Overfitting Bayesian Mixture Models With an Unknown Number of Components," *PLoS One*, 10, e0131739. [3,6]

Venkaiah, K., Brahmam, G., and Vijayaraghavan, K. (2011), "Application of Factor Analysis to Identify Dietary Patterns and Use of Factor Scores to Study Their Relationship With Nutritional Status of Adult Rural Populations," *Journal of Health, Population, and Nutrition*, 29, 327–338. [1]

Yoon, P. W., Rasmussen, S. A., Lynberg, M., Moore, C., Anderka, M., Carmichael, S., Costa, P., Druschel, C., Hobbs, C., Romitti, P., and Langlois, P. H. (2001), "The National Birth Defects Prevention Study," *Public Health Reports*, 116, 32–40. [6]

Zhang, J., Ghahramani, Z., and Yang, Y. (2004), "A Probabilistic Model for Online Document Clustering With Application to Novelty Detection," *Advances in Neural Information Processing Systems*, 4, 1617–1624. [2]

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L., and Carroll, R. J. (2011), "A New Multivariate Measurement Error Model With Zero-Inflated Dietary Data, and Its Application to Dietary Assessment," *The Annals of Applied Statistics*, 5, 1456–1487. [6]