Identifying Dietary Consumption Patterns from Survey Data: A Bayesian Nonparametric Latent Class Model

BRIANA J.K. STEPHENSON*, STEPHANIE M. WU, FRANCESCA DOMINICI

Harvard T.H. Chan School of Public Health, Department of Biostatistics, 665 Huntington Avenue, Boston, MA, United States
bstephenson@hsph.harvard.edu

SUMMARY

Dietary intake is one of the largest contributing factors to cardiovascular health in the United States. Amongst low-income adults, the impact is even more devastating. Dietary assessments, such as 24-hour recalls, provide snapshots of dietary habits in a study population. Questions remain on how generalizable those snapshots are in nationally representative survey data, where certain subgroups are sampled disproportionately to comprehensively examine the population. Many of the models that derive dietary patterns account for study design by incorporating the sampling weights to the derived model parameter estimates post hoc. We propose a Bayesian overfitted latent class model that accounts for survey design and sampling variability to derive dietary patterns in adults aged 20 and older. We compare these results with a subset of the population, adults considered low-income (at or below the 130% poverty income threshold), to understand if and how these patterns generalize in a smaller subpopulation. Using dietary intake

*To whom correspondence should be addressed.
data from the National Health and Nutrition Examination Surveys, we identified six dietary patterns in the US adult population. These differed in consumption features found in the five dietary patterns derived in low-income adults. Reproducible code/data are provided on GitHub to encourage further research and application in this area.

*Key words:* latent class model, dietary patterns, NHANES, survey design, Bayesian nonparametrics

1. **Introduction**

1.1 **Motivation**

The impact of poor diet has continually devastated the United States, accounting for over 500,000 deaths annually, with 84% of those deaths due to cardiovascular disease (CVD) (Mokdad *and others*, 2018; Roth *and others*, 2018). The negative health impacts of poor diet disproportionately affects low-income and racial minority populations (Brown *and others*, 2018; Daviglus *and others*, 2012; Zhang *and others*, 2018; Wang *and others*, 2014; Bahr, 2007; Fahlman *and others*, 2010). Understanding the dietary consumption behaviors that contribute to poor health in these target populations may help in tailoring interventions and resources to improve their nutritional health.

Through the implementation of complex survey designs and targeted recruitment strategies, researchers are able to obtain more representative population samples in an effort to better understand populations of greatest interest. Consequently, survey sampling methodologies have been developed to improve population-based estimates and generate appropriate inference based on the sampled data.

While low-income and racial minority adults are a population at greatest risk of poor diet and subsequently poorer health outcomes, they are often underrepresented in survey studies (Tourangeau *and others*, 2014). In an effort to achieve a more nationally representative sample, surveys such as the National Health and Nutrition Examination Survey (NHANES) have
corrected for this underrepresentation by oversampling demographics of greater public health interest to improve the accuracy and reliability of national-based estimates of health outcomes and exposures (Zipf and others, 2013). Unfortunately, most of the current statistical approaches for deriving dietary patterns from survey data do not incorporate the weights during estimation, which could lead to biased and inconsistent data-driven patterns for population demographics at greatest risk.

1.2 Challenges in Dietary Pattern Analysis

The high volume and variation of foods consumed by the study population can at times be cumbersome in dietary pattern analysis. Dimension reduction techniques are often employed either directly in the model using *a posteriori* approaches, where analysis is focused on a subset of foods that share strong similarities, as well as *a priori* approaches where similar food items are collapsed into major food groups for subsequent analysis (Schwedhelm and others, 2018; Bowman and others, 2017a). The most common *a posteriori* approaches applied to dietary assessment data are factor and cluster (e.g. K-means) analysis. These techniques rely on continuous, normally distributed data, where foods that share similar variation are grouped together (Sauvageot and others, 2017; Wirfält and Jeffery, 1997; Reedy and others, 2010).

Latent class models (LCM) are able to incorporate the full set of food items or groups to more comprehensively analyze different consumption patterns shared within the study population. In this model, the shared consumption habits for each food item or group are clustered together, as well as the subjects that share the respective clustered consumption habits.

LCMs are made available on commonly used statistical software such as SAS (Proc LCA) and R (poLCA) and offer parameter estimation of latent class model parameters via frequentist algorithms (e.g. Expectation-Maximization and Newton-Raphson) (Lanza and others, 2007; Linzer and Lewis, 2011; Muthén and Shedden, 1999). Bayesian approaches are also available
B. J. K. Stephenson and others

through an R package (BayesLCA), but it is limited to binary consumption responses (White and Murphy, 2014). These models are reflective of the study data being applied. Demographics that dominate a study population can easily mask dietary habits of smaller-sized demographics, which may deviate from the majority habit. This is evident in research studies where certain groups are historically undersampled, and consequently underrepresented in the study sample. Study designs have strived to correct for this through the implementation of sampling weights into the study design which allow undersampled participants to be upweighted and therefore identifiable in the study analysis. None of these standard packages listed previously, currently incorporate sampling weights directly into the estimation procedures. Mplus is one of the few statistical softwares available to adjust for complex survey design, but is limited under a frequentist framework, which can present issues with matrix inversion and computational demand when handling high-dimensional exposure data, such as diet, which can be large and sparse for rarely consumed food items (Muthén and Muthén, 2017).

1.3 Challenges in Survey Data Analysis

Model generation for complex survey data has fallen under two main approaches: (1) generate a pseudo-like population from the observed study data via a combination of bootstrapping and resampling techniques (Savitsky and Toth, 2016; Rao and Thomas, 1988; Skinner and Wakefield, 2017); (2) generate model parameter estimates first and correct for them using the sampling weights post hoc for population-based estimates and inference (Vermunt, 2002; Vermunt and Magidson, 2007; Stephenson and others, 2020b; Mattei and others, 2016).

Patterson and others (2002) and Vermunt and Magidson (2007) have implemented survey-weighted approaches to latent class models, generated under a frequentist framework. As previously mentioned, this can incur a huge computational burden during analysis when the number of subjects and food items increase, as well as when polytomous response patterns from each
food become increasingly complex. Dietary intake data that is often sparse (i.e., zero-inflated) also presents convergence issues for food items rarely or occasionally consumed. Additionally, standard latent class models operate under the assumption of a known number of latent classes (or patterns) to fit the model. However, that information is seldom known in practice, and as a result, multiple models are tested and fitted, with post-hoc model selection based on stability, clinical interpretability, and biological interactions among diet components (Nylund and others, 2007; Padmadas and others, 2006; Sotres-Alvarez and others, 2010; Harrington and others, 2014; Keshteli and others, 2015).

### 1.4 Potential Solutions in Bayesian Nonparametrics

Bayesian nonparametrics offers a more efficient solution that is able to (1) accommodate the complex high dimensionality of dietary intake data, (2) handle large-sized populations, such as the United States, (3) reduce multiple model testing and fitting to determine the appropriate number of patterns, (4) preserve model stability in the presence of sparsely consumed foods, and (5) integrate prior information with observed data. These features improve parameter estimation and subsequent population inference (Hjort and others, 2010; Liu, 2008). In spite of all of the benefits available for applications to large population-representative survey studies, few methods are currently offered that can address complex survey design, and even fewer have been applied for use in nutritional studies. Multiple testing required in frequentist settings to determine the appropriate number of latent groups can be reduced by using Bayesian nonparametric approaches, currently applied primarily in genetics and bioinformatics (Bhattacharya and Dunson, 2011; Runcie and Mukherjee, 2013; Gao and others, 2016; Marttinen and others, 2014; Pelleg, 2000). Bayesian nonparametric survey data applications center mostly on generating inference for derived population-based estimates (Si and others, 2015; Savitsky and Toth, 2016; Gunawan and others, 2020). Bayesian nonparametric mixture models have been applied in nutritional set-
B. J. K. Stephenson and others

tings, but implementation accounting for survey design in model estimation has not been fully explored (Fahey and others, 2007; De Vito and others, 2019; Stephenson and others, 2020a). Bayesian nonparametric mixture models have been applied to diet survey data, but the survey weights were applied after parameters were estimated from the sampling algorithm (Stephenson and others, 2020b; Stephenson and Willett, 2022; De Vito and others, 2022). Kunihama and others (2016) used a Dirichlet Process mixture model to introduce a sampling algorithm that can incorporate survey weights directly into the estimation of a Bayesian nonparametric mixture model. However, with a focus primarily on generating a pseudo-like population, it did not take into account sampling variability present in nationally-representative surveys.

In an effort to better examine nationally representative dietary patterns, we have built upon this framework and added the following contributions: (1) implemented an overfitted finite mixture model, which is asymptotically similar to the Dirichlet Process mixture model; (2) extended and integrated the works of Kunihama and others (2016) and Savitsky and Toth (2016) to generate population-based estimates that also adjust for sampling variability in the survey design; (3) demonstrated the utility of this approach by applying this model to publicly available national survey data to derive nationally representative dietary consumption patterns of low-income adults in the United States from 2011-2018; and (4) provided publicly available reproducible code for researchers to apply this technique on future national dietary survey data.

We organize this paper as follows: Section 2 describes our proposed weighted overfitted latent class model. Section 3 compares our model with current model-based approaches for survey data via a simulation study. Section 4 describes the National Health and Nutrition Examination Survey. Section 5 presents results of the method applied to the National Health and Nutrition Examination Survey. Section 6 discusses next steps and future directions.
2. Weighted Overfitted Latent Class Model

A weighted overfitted latent class model is a Bayesian nonparametric technique that can be used to identify subgroups or clusters within a survey sample that share common behaviors amongst a set of observed nominal variables (Van Havre and others, 2015). It can be seen as an extension of the latent class model, which typically requires multiple fits and post hoc testing to determine the appropriate number of latent classes or patterns. The overfitted latent class model removes this redundancy by overfitting the model with a large number of latent classes (or clusters) and allowing a data-driven approach to choosing the number of latent clusters. Empty clusters are able to drop out of the model during the Markov chain Monte Carlo Gibbs sampling algorithm, and nonempty clusters remain. Each participant is assigned to one of the derived clusters, corresponding to a dietary pattern. The overfitted structure is also asymptotically equivalent to the Dirichlet Process model, allowing additional flexibility within a Bayesian nonparametric framework (Van Havre and others, 2015).

We define some notation of the standard latent class model, with a sampled population of size \( n \) and \( K \) unique dietary patterns, where each pattern describes the consumption of \( p \) food items. Let \( y_i = (y_{i1}, \ldots, y_{ip}) \) denote the set of \( p \) observed food items. Each observed food item, \( y_{ij} \), is categorical, where \( y_{ij} \in \{1, 2, \ldots, d_j\} \) is individuals \( i \)'s consumption level for food item \( j \). Let \( \pi_k \) denote the probability of assignment to dietary pattern \( k \in \{1, \ldots, K\} \), and \( z_i \) the dietary pattern assignment of individual \( i \in \{1, \ldots, n\} \) from the sampled population. Let \( \theta_{jc|k} \) denote the probability of consuming food item \( j \), at the \( c \in \{1, \ldots, d_j\} \) consumption level, given an individual's assignment to diet pattern \( k \). The subject-specific likelihood is then defined as

\[
Pr(y_i; \theta, \pi, z_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \prod_{c=1}^{d_j} \theta_{jc|k} I(y_{ij}=c | z_i=k).
\]

The likelihood for the overfitted latent class model shares the same structure as that of the standard latent class model shown in (2.1), but since \( K \) is typically not known in practice, it
is fixed to an exceedingly large number that asymptotically simulates an infinite mixture model (Van Havre and others, 2015). Under a Bayesian estimation framework, the model parameters are dependent on the number of observed individuals classified to a given latent class or consumption level. For example, exploiting the convenience of conjugacy, the probability vector, \( \pi = (\pi_1, \ldots, \pi_K) \), follows a Dirichlet prior and posterior with hyperparameters for each latent class defined as \( (\alpha_1, \ldots, \alpha_K) \):

\[
\pi = (\pi_1, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)
\]

\[
(\pi_1, \ldots, \pi_K | y, z.) \sim \text{Dir}\left(\alpha_1 + \sum_{i=1}^{n} 1(z_i = 1), \ldots, \alpha_K + \sum_{i=1}^{n} 1(z_i = K)\right),
\]

(2.2)

where \( y = (y_1, \ldots, y_n) \) and \( z = (z_1, \ldots, z_n) \). With no prior knowledge on the number of classes, we utilize a noninformative, flat Dirichlet prior, where \( \alpha_1 = \alpha_2 = \ldots = \alpha_K = \alpha \). This hyperparameter moderates the rate of growth for nonempty latent classes. The smaller the hyperparameter, the slower nonempty clusters will form. Similarly, we assume no prior knowledge on the consumption pattern of each observed food, such that \( \theta_{j.k} \sim \text{Dir}(\gamma_1, \ldots, \gamma_d) \) for all \( k \in \{1, 2, \ldots, K\} \) is also fit with a non-informative flat Dirichlet prior with a constant \( \gamma \) hyperparameter \( (\gamma_1 = \gamma_2 = \gamma_d = \gamma) \).

Incorporating survey weights in a Bayesian setting serves as a natural extension to the overfitted latent class model. As described in Kunihama and others (2016), information used to update each model parameter is enhanced with weights, simulating a pseudo-like population that is similar in size and structure to the target population. A normalization constant is used to ensure the weights sum to the target population. This enables dietary patterns to form in accordance with the target population, but does not consider changes that can occur in size and composition from one sampled population to another. Sampling variability should be considered in the model, and precision estimates should reflect the sample size rather than the population size. Otherwise, uncertainty surrounding model estimation will be biased. To address this limitation, we instead propose an approach similar to Savitsky and Toth (2016) and normalize the sampling weights to
BNP Latent Class Model for Survey Data

sum to the sample size. This will account for sampling variability while allowing model estimates to generalize better to the target population.

Let $w_i$ denote the sampling weight of study participant $i \in \{1, \ldots, n\}$. We impose a fixed normalization constant, $\kappa$, where $\kappa = \frac{\sum w_i}{n}$, with $n$ denoting the study sample size. With this newly defined $\kappa$ and with $\mathbf{w} = (w_1, \ldots, w_n)$, the conditional posterior of the probability of assignment vector, $\mathbf{\pi} = (\pi_1, \ldots, \pi_K)$, updates based on the weighted number of participants assigned to each pattern:

$$
(\pi_1, \ldots, \pi_K|\mathbf{y}, \mathbf{z}, \mathbf{w}) \sim \text{Dir} \left( \frac{\alpha_1}{\kappa} \sum_{i=1}^{n} w_i \times 1(z_i = 1), \ldots, \frac{\alpha_K}{\kappa} \sum_{i=1}^{n} w_i \times 1(z_i = K) \right). 
$$

(2.3)

Similarly, for the consumption level distribution of each dietary pattern, $f(y_{ij}|z_i = k) = \prod_{j=1}^{d_j} \prod_{c=1}^{d_{jc}} \theta_{j|c|k}^{1(y_{ij}=c|z_i=k)}$, $k \in \{1, \ldots, K\}$, updates for the conditional posteriors of the probabilities of consumption are based on the weighted number of participants that share dietary consumption behaviors. For all $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, K\}$,

$$
\theta_{j|k} \sim \text{Dir} \left( \gamma + \frac{1}{\kappa} \sum_{i:z_i=k} w_i \times 1(y_{ij} = 1), \ldots, \gamma + \frac{1}{\kappa} \sum_{i:z_i=k} w_i \times 1(y_{ij} = d_j) \right). 
$$

(2.4)

3. Simulation Study

3.1 Survey-weighted Approaches

We performed our simulation study under three different approaches for handling survey data. Method 1 serves as our control, a standard overfitted latent class model, where sample weights are ignored. Method 2 provides an alternative approach, a weighted finite population Bayesian bootstrap (WFPBB) (Gunawan and others, 2020; Dong and others, 2014), where pseudo-representative samples are generated by using survey weights to “undo” the unequal sampling scheme and impute a synthetic population, where approximate simple random samples are drawn from this population. This application of the WFPBB method builds on earlier work of pseudo-population
generation through multiple imputation techniques (Raghunathan and others, 2003; Zhou and others, 2016). Implementation details of this method are provided in section ?? Method 3 is our proposed weighted overfitted latent class model (wtOLCM) that extends the work of Kunihama and others (2016) and Savitsky and Toth (2016) where the sample weights are directly incorporated into the sampling algorithm, as detailed in section 2. Our simulation study will evaluate how well these three methods are able to identify the true population prevalence of dietary patterns using the sampled data.

3.2 Simulation Setup

We consider a simulated population of size $N = 5000$. A total of $K_{\text{true}} = 3$ patterns exist in the population with probability distribution $\pi_{\text{true}} = (0.1, 0.3, 0.6)$. Each pattern consists of $p = 50$ categorical variables that can take on values 1, 2, 3 or 4. For case A, the mode was set at 0.85 for the true pattern value of interest, and at 0.05 for the remaining three values. To evaluate under additional noise, case B was performed where the mode was set at 0.55, and 0.15 for all other remaining values. Pattern 1 was defined with a mode at level 3 for the first 25 variables, and a mode at level 1 for the remaining 25 variables. Pattern 2 was defined with a mode at level 2 for the first 10 variables, and a mode at level 4 for the remaining 40 variables. Pattern 3 was defined with a mode at level 1 for the first 10 variables, a mode at level 2 for the next 20 variables, and a mode at level 3 for the remaining 20 variables. Subjects were initially assigned to one of these three patterns, and the subject-specific observed data was simulated by drawing from a multinomial distribution for each of the 50 corresponding variables described above based on the assigned pattern. The total population was comprised of $S = 4$ disproportionate subpopulations containing varied distributions of the three patterns (Table 1). A subset of 100 subjects were randomly selected from each of the simulated subpopulations, totaling $n = 400$ simulated subjects in each sample dataset. A total of 100 simulated datasets were generated for replicability, and
Table 1. Distribution of patterns across simulated population containing four subpopulations and three uniquely distinct patterns

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Pattern 1</th>
<th>Pattern 2</th>
<th>Pattern 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>900</td>
<td>75</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>175</td>
<td>1825</td>
<td>2000</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>425</td>
<td>1050</td>
<td>1500</td>
</tr>
<tr>
<td>4</td>
<td>450</td>
<td>0</td>
<td>50</td>
<td>500</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>500</strong></td>
<td><strong>1500</strong></td>
<td><strong>3000</strong></td>
<td><strong>5000</strong></td>
</tr>
</tbody>
</table>

the three previously described approaches in section 3.1 were run using MATLAB (2021a).

3.3 Simulation Results

Model diagnostics indicated good mixing and successful convergence of model parameters across all three methods. Derived patterns were identified by setting the modal response to be the categorical level of each exposure variable that had the highest posterior probability of consumption. As illustrated in Figure 1, each of the methods successfully identified the true number of patterns ($K = 3$) as well as the modal response patterns in case A. The additional noise incorporated in case B generated additional clusters containing redundancies to the true patterns, but small in size ($\pi_4 < 0.03$). This is a natural occurrence in overfitted mixture models, and can be resolved by cluster merging during postprocessing (Rousseau and Mengersen, 2011; Van Havre and others, 2015; Miller and Harrison, 2013).

Bias and precision in the expected prevalence of the patterns did differ across the three methods, as illustrated in Figure 2. Under the unweighted method, the MSE of the true pattern prevalence in the population was 0.015 and 0.016, respectively. Both the WFPBB and the wtOLCM methods also had an improved estimation of the population prevalence compared to the unweighted case. Among all methods, wtOLCM had better coverage of the true population prevalence ($MSE(\hat{\pi}_A) = 1.3 \times 10^{-4}, MSE(\hat{\pi}_B) = 0.002$) in both simulation cases.
Fig. 1. Modal consumption patterns identified from respective models compared to truth. Method 1: overfitted latent class model, ignoring weights; Method 2: weighted finite population Bayesian bootstrap; Method 3: weighted overfitted latent class model. Top indicates pattern under simulation case A. Bottom indicates pattern under simulation B. The additional noisy cluster is illustrated in method 1, where the size of this pattern had a prevalence of 0.02.
Table 2. Simulation results for $\hat{\pi}$ for the three approaches, summarized across 100 simulated datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE($\hat{\pi}_A$)</th>
<th>MSE($\hat{\pi}_B$)</th>
<th>Runtime (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted OLCM</td>
<td>$1.49 \times 10^{-2}$</td>
<td>.016</td>
<td>18.3</td>
</tr>
<tr>
<td>WFPBB</td>
<td>$4.31 \times 10^{-4}$</td>
<td>.006</td>
<td>19.2, parallelized across 200 cores</td>
</tr>
<tr>
<td>wtOLCM</td>
<td>$1.32 \times 10^{-4}$</td>
<td>.002</td>
<td>20.9</td>
</tr>
</tbody>
</table>

Fig. 2. Predicted population prevalence from unweighted and weighted estimation approaches. Expected prevalence for each respective cluster is 0.1 (black), 0.3 (green), 0.6 (magenta)

4. NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES) DATA

The National Health and Nutrition Examination Survey (NHANES) is a population-based survey designed to assess the health and nutritional status of adults and children in the United States. The survey samples at least 9,000 people across various socioeconomic status (SES) levels each year residing in 15 randomly selected counties in the United States. Starting in 2011, NHANES created more granularity to the race/ethnicity variable, separating Mexican-American from Other Hispanic participants, as well as adding an identifier for Non-Hispanic Asian. For the scope of
B. J. K. Stephenson and others

this study, we limited analysis to survey cycles containing the seven race/ethnicity groups, and adults aged 20 and over. Low-income participants were identified as those reporting at or below the 130% poverty income level.

Dietary intake was collected via the ‘What We Eat in America’ survey component of NHANES. Food items and beverages were consumed and recorded via two 24-hour recalls. Nutrients comprising these reported food/beverage items were calculated using the Food and Nutrition Database for Dietary Studies (FNDDS) and then converted into food pattern equivalents per 100 g of consumption based on the Dietary Guidelines for Americans (Committee and others, 2015; U.S. Department of Agriculture and Promotion, 2020; Bowman and others, 2016, 2017b, 2018).

Dietary consumption data were summarized as 29 food groups and pooled across four NHANES survey cycles: 2011-2012, 2013-2014, 2015-2016, and 2017-2018. Consumption levels were derived by segmenting the data into no consumption (none=0%) and tertiles of positive consumption (Liu and others, 2019; Sotres-Alvarez and others, 2013). NHANES dietary weights were adjusted for the pooled survey years in accordance with protocols outlined in NHANES analytic guidelines (National Center for Health Statistics and Surveys, 2018; Chen and others, 2020).

Demographic information of the low-income adult participants collected in NHANES are detailed in Table 3. The low-income sampled population reflected a demographic with the larger proportion of participants identifying as non-Hispanic White (47.6%), female (54.5%), and between 20-34 years old (35.7%). This sampled population reported an Alternative Healthy Eating Index (AHEI-2015) score of 49.2 out of 100, which is less than the overall national average of 58 out of 100 (U.S. Department of Agriculture and Promotion, 2022). The mean Framingham 10-year risk score indicated a low risk of a CVD outcome occurring in the next ten years ($FRS = 7.7$).
Table 3. NHANES 2011–2018 adult participant demographics

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Overall</th>
<th>Low-income</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>% (SE)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexican</td>
<td>2632</td>
<td>9.2 (1.0)</td>
</tr>
<tr>
<td>Other Hispanic</td>
<td>2026</td>
<td>6.3 (0.6)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>7480</td>
<td>64.0 (1.7)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>4471</td>
<td>11.3 (1.0)</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>2282</td>
<td>5.8 (0.5)</td>
</tr>
<tr>
<td>Mixed/Other</td>
<td>716</td>
<td>3.4 (0.3)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>10077</td>
<td>48.8 (0.5)</td>
</tr>
<tr>
<td>Female</td>
<td>9530</td>
<td>51.2 (0.5)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-34 Years</td>
<td>4991</td>
<td>28.9 (0.8)</td>
</tr>
<tr>
<td>35-49 Years</td>
<td>4801</td>
<td>25.5 (0.7)</td>
</tr>
<tr>
<td>50-64 Years</td>
<td>5246</td>
<td>26.9 (0.6)</td>
</tr>
<tr>
<td>65+ Years</td>
<td>4569</td>
<td>18.8 (0.6)</td>
</tr>
<tr>
<td>AHEI2015 Score</td>
<td>14865</td>
<td>51.6 (0.3)</td>
</tr>
<tr>
<td>Framingham 10YR Score</td>
<td>18226</td>
<td>8.1 (0.1)</td>
</tr>
<tr>
<td>CVD Risk factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>8260</td>
<td>32.6 (0.9)</td>
</tr>
<tr>
<td>Obesity</td>
<td>8198</td>
<td>39.0 (0.9)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2010</td>
<td>9.3 (0.4)</td>
</tr>
<tr>
<td>High Cholesterol</td>
<td>9906</td>
<td>75.1 (0.8)</td>
</tr>
<tr>
<td>Smoker</td>
<td>3487</td>
<td>15.1 (0.7)</td>
</tr>
</tbody>
</table>

5. NHANES Application of Weighed Overfitted Latent Class Model

5.1 Fitting the Model

For our model, the normalization constant ($\kappa = 9.79 \times 10^3$) was calculated based on the sum of the sampled weights divide by the total sample size ($n = 7561$). We overfit the model with $K = 50$ latent classes. Estimation was performed using a Gibbs sampler of 10,000 iterations after a 15,000 burn-in and a thinning every 5 iterations. Posterior median and 95% credible intervals were derived from the MCMC output results. Flat, symmetric Dirichlet priors were fit with the probability of class assignment, $\pi$, and the food item probability of consumption given assignment to specific latent class, $\theta_{j|k}$, $j \in \{1, \ldots, p\}$, $k \in \{1, \ldots, K\}$. A random permutation sampler was
implemented to encourage mixing (Frühwirth-Schnatter, 2001). Dietary weights were calibrated and normalized for inclusion in analysis. We defined hyperprior $\alpha = \frac{1}{K}$ to conservatively moderate the rate of cluster growth as suggested in Rousseau and Mengersen (2011).

A common consequence in mixture modeling under Bayesian estimation is label switching, where label components swap assignment of individuals while the likelihood remains invariant (Stephens, 2000). We resolved this phenomenon by performing hierarchical clustering on a similarity matrix of size $n \times n$. Matrix elements contained pairwise posterior probabilities of two subjects being clustered together in each MCMC iteration (Krebs, 1989; Medvedovic and Sivaganesan, 2002). Labels were identified based on subjects that remained clustered together through the sampling algorithm. Nonempty clusters were defined as any cluster containing at least 5% of the sampled participants. Dietary patterns were defined by identifying the consumption level corresponding to the highest posterior median probability for each food item in the set.

All data included for this study and code to reproduce the derived dataset and perform subsequent analyses are made available on the author’s GitHub repository: http://www.github.com/bjks10/NHANES_wtofm. Dietary data was originally obtained from the NHANES website (https://wwwn.cdc.gov/nchs/nhanes) and processed in SAS 9.4. Statistical analysis and figures were performed in MATLAB 2021a. Posthoc analysis and table summaries were generated in R version 4.0.2.

5.2 Results

The weighted overfitted latent class model identified five nonempty clusters in the low-income adult population. Figure 3 illustrates the posterior mean estimates of the probability of no consumption or high consumption given membership to a given dietary pattern. From this figure, we can see which foods were strongly favored to be consumed for various patterns. The very low probabilities of no consumption across all patterns for refined grains, oils, solid fats, and added
sugar imply a general nonzero consumption by all low-income adults. Foods such as poultry, seafood, eggs, soybean and alcohol shared similar consumption behaviors across all diet patterns, but other foods differed by pattern. For example, patterns 1 and 5 had the lowest probabilities for consumption of cheese, oils, added sugars, and fats at the high consumption level. Patterns 2 and 3 had the highest probabilities for consumption of refined grains, potatoes, cured meats, oils, solid fat, and added sugar at the high consumption level. Lastly, pattern 4 distinctly had the highest probability of legumes being consumed at the high consumption level. Dietary pattern 1, followed by pattern 3, had the highest probabilities of no consumption of most fruits and vegetables.

Comparing more closely the posterior modes of consumption for each dietary pattern, we note that 15 foods shared a mode of non-consumption (i.e., consumption value of 1) across the five dietary patterns (Figure 4): citrus/melon/berries, fruit juice, dark green vegetables, other red/orange vegetables, potatoes, other starchy vegetables, whole grains, organ meat, poultry, seafood (high-n3), seafood (low-n3), soybean, nuts/seeds, yogurt, and alcohol. Pattern 1 showed strong similarities with Pattern 5. However, pattern 5 had comparatively higher levels of consumption of other fruit and milk. Patterns 2 and 3 also shared similar consumption of foods, with differences noted in the higher level of consumption for eggs and cheese in pattern 3. As previously noted, pattern 4 was the most distinguishable amongst the five patterns, with a high level of consumption favored in tomatoes, legumes (veg and protein), and non-specified meat.

Table 4 provides a summary of the demographics for participants assigned to each dietary pattern. Amongst the low-income adult population, participants assigned to pattern 5 had the highest average HEI-2015 score (57.4 ± 0.6). This pattern favored a high consumption of other fruit, but a low consumption of refined grains and no consumption of meats. Pattern 3 had the lowest average HEI-2015 score (41.1 ± 0.3). This pattern favored a high consumption of refined grains, cured meats, eggs, cheese, fats, oils, and sugars. Demographically, we observe that
Fig. 3. Low-income Adult population - Model 2: (top) Posterior mean probability of no consumption of a given food item given membership in a specified pattern; (bottom) Posterior mean probability of high consumption of a given food item given membership in a specified pattern.

those in pattern 5 were predominantly male adults, whereas those assigned to pattern 3 were predominantly female adults. While non-Hispanic White participants held the majority of all patterns in our model, pattern 4, which uniquely favored a high consumption of legumes, was the only pattern where minority participants had a higher representation.
**Fig. 4.** Posterior mode of consumption pattern of dietary patterns for non-incarcerated adults living at or below the 130% poverty level. Numbers represent levels of consumption: 1= None, 2=Low, 3=Medium, 4=High
Table 4. Demographic distribution of Low-income Dietary patterns

<table>
<thead>
<tr>
<th></th>
<th>Pattern 1</th>
<th>Pattern 2</th>
<th>Pattern 3</th>
<th>Pattern 4</th>
<th>Pattern 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean SE</td>
<td>Mean SE</td>
<td>Mean SE</td>
<td>Mean SE</td>
<td>Mean SE</td>
</tr>
<tr>
<td>Overall</td>
<td>21.7 0.6</td>
<td>11.7 0.6</td>
<td>22.0 0.7</td>
<td>29.3 1.0</td>
<td>15.4 0.9</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexican</td>
<td>10.2 1.7</td>
<td>12.2 1.6</td>
<td>11.0 1.5</td>
<td>26.4 2.9</td>
<td>10.8 1.5</td>
</tr>
<tr>
<td>Other Hispanic</td>
<td>8.4 1.2</td>
<td>7.5 1.2</td>
<td>5.2 0.9</td>
<td>15.7 1.8</td>
<td>7.9 1.0</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>50.5 2.9</td>
<td>54.3 3.4</td>
<td>55.1 2.7</td>
<td>37.0 2.8</td>
<td>47.9 3.2</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>21.2 2.1</td>
<td>18.2 2.1</td>
<td>22.1 2.6</td>
<td>11.1 1.2</td>
<td>16.8 1.8</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>5.4 0.9</td>
<td>4.5 0.8</td>
<td>1.2 0.2</td>
<td>5.8 0.9</td>
<td>12.5 1.8</td>
</tr>
<tr>
<td>Mixed/Other</td>
<td>4.4 1.0</td>
<td>3.3 0.7</td>
<td>5.5 0.7</td>
<td>4.0 0.8</td>
<td>4.2 0.9</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>63.2 1.5</td>
<td>47.2 2.5</td>
<td>38.6 1.5</td>
<td>53.6 1.5</td>
<td>72.2 2.1</td>
</tr>
<tr>
<td>Female</td>
<td>36.8 1.5</td>
<td>52.8 2.5</td>
<td>61.4 1.5</td>
<td>46.4 1.5</td>
<td>27.7 2.1</td>
</tr>
<tr>
<td>Age Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-34 years</td>
<td>33.1 1.9</td>
<td>38.2 2.8</td>
<td>46.6 2.5</td>
<td>35.2 2.1</td>
<td>22.8 2.0</td>
</tr>
<tr>
<td>35-49 years</td>
<td>24.9 1.5</td>
<td>20.5 1.6</td>
<td>25.6 1.5</td>
<td>26.6 1.4</td>
<td>20.0 1.5</td>
</tr>
<tr>
<td>50-64 years</td>
<td>26.1 1.4</td>
<td>22.5 2.0</td>
<td>18.2 1.6</td>
<td>25.3 1.7</td>
<td>27.3 2.1</td>
</tr>
<tr>
<td>65+ years</td>
<td>15.9 1.1</td>
<td>18.8 2.1</td>
<td>9.6 1.0</td>
<td>12.9 1.0</td>
<td>30.0 2.1</td>
</tr>
<tr>
<td>HEI 2015 Score</td>
<td>45.8 0.5</td>
<td>50.3 0.6</td>
<td>41.1 0.3</td>
<td>53.2 0.5</td>
<td>57.4 0.6</td>
</tr>
<tr>
<td>Framingham 10YR Risk</td>
<td>7.7 0.3</td>
<td>7.6 0.4</td>
<td>7.1 0.4</td>
<td>7.4 0.3</td>
<td>9.3 0.5</td>
</tr>
<tr>
<td>CVD Risk factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>34.9 2.1</td>
<td>30.9 3.5</td>
<td>30.9 2.0</td>
<td>30.2 1.7</td>
<td>36.4 2.8</td>
</tr>
<tr>
<td>Hypercholesteremia</td>
<td>76.7 1.9</td>
<td>67.9 4.2</td>
<td>69.4 1.9</td>
<td>75.5 1.9</td>
<td>68.1 2.5</td>
</tr>
<tr>
<td>Obesity</td>
<td>45.7 2.6</td>
<td>42.2 3.8</td>
<td>41.4 2.5</td>
<td>40.1 1.9</td>
<td>38.0 2.7</td>
</tr>
<tr>
<td>Diabetes</td>
<td>12.8 1.0</td>
<td>10.0 1.6</td>
<td>8.2 1.2</td>
<td>9.1 1.1</td>
<td>14.2 1.6</td>
</tr>
<tr>
<td>Smoker</td>
<td>34.5 3.1</td>
<td>18.6 2.5</td>
<td>34.9 3.0</td>
<td>19.2 2.1</td>
<td>11.9 1.8</td>
</tr>
</tbody>
</table>

6. Discussion

The weighted overfitted latent class model for survey data, proposed in this paper, is an extension of the standard latent class model and integrates a Bayesian nonparametric survey-weighted approach to account for sampling variability in its parameter estimation. In this paper, we applied the proposed Bayesian nonparametric latent class model to dietary survey data collected in the 2011-2018 National Health and Nutrition Examination Surveys of US adults living at or below the 130% poverty income level. Our model identified five dietary patterns in this sampled subset. Application of our model to this target population allowed us to leverage survey weights to obtain representative estimates from a smaller, often underrepresented and understudied, subset of the
surveyed participants.

This method builds its strength on its generalizability and use in nationally representative dietary surveys, yet recognizes the concerns of overgeneralization. Dominating demographics can still influence pattern distribution in a given population. Non-Hispanic White participants have historically dominated surveys and studies that examine diet-disease relationships (Wall and others, 2018; Colditz and others, 2016; Ohlhorst and others, 2013; Fahlman and others, 2010). This is due to the global clustering assumption prevalent in finite mixture models, where all participants are assigned to a single cluster and assume shared consumption behaviors for all foods described in that cluster. An overrepresentation of this subgroup can mask accurate pattern identification for racial/ethnic minorities who may be at greatest risk of chronic disease. This is exemplified in our model through Pattern 4, which had the most distinguishable dietary consumption pattern. Compared to the other five patterns, this contained the smallest proportion of non-Hispanic White adults, but still the largest relative proportion amongst the other racial/ethnic subgroups. If certain subgroups are important to understand nutrition disparities, those subgroups should be studied in a separate analysis or a more advanced method that is able to jointly account for subgroup differences should be implemented. To our knowledge a few advanced methods have been used to better examine subpopulation behavior differences, but incorporating the complex survey design directly into model estimation has not yet been fully explored (De Vito and others, 2019; Stephenson and Willett, 2022).

While the utility of this model has effectively demonstrated its use in diet survey data, we must also acknowledge that the dietary intake data used is limited by its reliance on self-reporting. Several nutrition studies have found that prudent foods like vegetables and fruits are often over-reported and less prudent foods like fats and oils are frequently underreported (Haraldsdóttir, 1993; Amanatidis and others, 2001). These tendencies to misreport have been associated with demographics such as BMI, age, sex, socioeconomic status, as well as other psychosocial and
cognitive factors (Poslusna and others, 2009; Briefel and others, 1997; Klesges and others, 1995; Hirvonen and others, 1997; Tooze and others, 2004).

Methods such as doubly labeled water and biomarkers for select nutrients are available to validate dietary assessment tools, but these instruments are beyond the scope of tools utilized in the National Health and Nutrition Examination Survey. In spite of this limitation, the misreporting rate remains relatively low and the instruments can still be deemed relatively reliable (Tooze and others, 2004; Yuan and others, 2017). Another limitation of dietary recalls is the inability to capture day-to-day variation. As a result, these dietary patterns are based on one or two days of dietary records, which may or may not reflect participants' regular dietary behaviors. Alternative dietary assessments, such as food frequency questionnaires and 7-day daily diet records, are available to capture more episodic and rarely consumed foods. However, more detailed assessments are often costly and seldom widely available in large population-based surveys. Future research can explore ways to integrate these tools, when available, to quantify the unknown variation and uncertainty that comes from misreporting in dietary assessments.

The clustering approach applied in this paper, as well as more traditionally used cluster and factor analysis, are all generated independent of any health outcome. Yet, when looking at exposures from a multi-dimensional perspective, these exposures may be driven by an underlying health outcome, in which case a more supervised approach may yield more useful information to understand how the combination of these exposures (e.g., dietary habits) can drive a known outcome (e.g., cardiometabolic health). Further research is needed to develop supervised clustering methods that address the issue of confounding overgeneralizations and are applicable in population surveys with complex survey designs.
REFERENCES

7. Software

Software in the form of MATLAB code, together with a sample input data set and complete documentation, is available in the GitHub repository https://github.com/bjks10/NHANES_wtofm.

SUPPLEMENTARY MATERIALS

Supplementary material detailing the weighted finite population Bayesian bootstrap (WFPBB) method and figure illustrating the complete pattern distribution of low-income adults are available online at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGEMENTS

The authors are grateful to Walter Willett, DC Rao, and Lei Liu for helpful comments on earlier versions of this work. This study was supported in part by NHLBI grant R25 HL105400 to DC Rao and Victor G. Davila-Roman. Conflict of Interest: None declared.

REFERENCES


BOWMAN, SA, Clemens, JC, Friday, JE, Lynch, KL, LaComb, RP and Moshfegh, AJ.
REFERENCES


Chen, Te-Ching, Clark, Jason, Riddles, Minsun K, Mohadjer, Leyla K and Fakhouri, Tala HI. (2020). National health and nutrition examination survey, 2015-


REFERENCES


**Liu, Lei, Shih, Ya-Chen Tina, Strawderman, Robert L, Zhang, Daowen, Johnson,**
REFERENCES


POSLSNA, KAMILA, RUPRICH, JIRI, DE VRIES, JEANNE HM, JAKUBIKOVA, MARIE AND VAN’T VEER, PIETER. (2009). Misreporting of energy and micronutrient intake estimated by food


Sauvageot, Nicolas, Schritz, Anna, Leite, Sonia, Alkerwi, Al’a, Stranges, Saverio, Zannad, Faiez, Streefl, Sylvie, Hoge, Axelle, Donneau, Anne-Françoise, Al-


REFERENCES


by supplemental nutrition assistance program participation status. *JAMA network open* 1(2), e180237–e180237.


[Received August 1, 20XX; revised October 1, 20XX; accepted for publication November 1, 20XX]