Biostatistics (2022), **0**, 0, *pp.* 1–**??**
doi:10.1093/biostatistics/output

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.

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

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*, 2017*a*). 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

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*, 2020*b*; 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-

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*, 2020*a*). 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*, 2020*b*; 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 $\boldsymbol{y_{i\cdot}} = (y_{i1}, \ldots, y_{ip})$ 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 π_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

$$Pr(\boldsymbol{y_{i\cdot}}|\theta, \pi, z_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \prod_{c=1}^{d_j} \theta_{jc|k}^{1(y_{ij}=c|z_i=k)}. \tag{2.1}$$

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, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, follows a Dirichlet prior and posterior with hyperparameters for each latent class defined as $(\alpha_1, \ldots, \alpha_K)$:

$$\begin{aligned} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim Dir(\alpha_1, \ldots, \alpha_K) \\ (\pi_1, \ldots, \pi_K | \mathbf{y}_{\cdot\cdot}, \mathbf{z}_{\cdot}) &\sim Dir\left(\alpha_1 + \sum_{i=1}^{n} 1(z_i = 1), \ldots, \alpha_K + \sum_{i=1}^{n} 1(z_i = K)\right), \end{aligned} \tag{2.2}$$

where $\mathbf{y}_{\cdot\cdot} = (\mathbf{y_1}_{\cdot}, \ldots, \mathbf{y_n}_{\cdot})$ and $\boldsymbol{z}_{\cdot} = (z_1, \ldots, z_n)$. 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 $\left\{\boldsymbol{\theta_{j\cdot|k)}} = (\theta_{j1|k}, \ldots, \theta_{jd_j|k})\right\}_{j=1}^{p} \sim Dir(\gamma_1, \ldots, \gamma_{d_j})$ for all $k \in \{1, 2, \ldots, K\}$ is also fit with a non-informative flat Dirichlet prior with a constant γ hyperparameter ($\gamma_1 = \gamma_2 = \gamma_{d_j} = \gamma$).

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

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, κ, where $\kappa = \frac{\sum_i w_i}{n}$, with n denoting the study sample size. With this newly defined κ and with $\boldsymbol{w}_{\cdot} = (w_1, \ldots, w_n)$, the conditional posterior of the probability of assignment vector, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, updates based on the weighted number of participants assigned to each pattern:

$$(\pi_1, \ldots, \pi_K | \boldsymbol{y}_{\cdot\cdot}, \boldsymbol{z}_{\cdot}, \boldsymbol{w}_{\cdot}) \sim Dir\left(\alpha_1 + \frac{1}{\kappa}\sum_{i=1}^{n} w_i \times 1(z_i = 1), \ldots, \alpha_K + \frac{1}{\kappa}\sum_{i=1}^{n} w_i \times 1(z_i = K)\right). \tag{2.3}$$

Similarly, for the consumption level distribution of each dietary pattern, $f(\boldsymbol{y_{i\cdot}}|z_i = k) = \prod_{j=1}^{p}\prod_{c=1}^{d_j}\theta_{jc|k}^{1(y_{ij}=c|z_i=k)}$, $k \in \{1, \ldots, K\}$, 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 $k \in \{1, \ldots, K\}$,

$$\boldsymbol{\theta_{j\cdot|k}} \sim Dir\left(\gamma + \frac{1}{\kappa}\sum_{i:z_i=k} w_i \times 1(y_{ij} = 1), \ldots, \gamma + \frac{1}{\kappa}\sum_{i:z_i=k} w_i \times 1(y_{ij} = d_j)\right). \tag{2.4}$$

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 $K_{true} = 3$ patterns exist in the population with probability distribution $\boldsymbol{\pi_{true}} = (0.1, 0.3, 0.6)$. 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

	Pattern 1	**Pattern 2**	**Pattern 3**	**Total**
Subpopulation 1	25	900	75	1000
Subpopulation 2	0	175	1825	2000
Subpopulation 3	25	425	1050	1500
Subpopulation 4	450	0	50	500
Total	500	1500	3000	5000

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 ($MSE(\hat{\boldsymbol{\pi}}_{\boldsymbol{A}}) = 1.3 \times 10^{-4}, MSE(\hat{\boldsymbol{\pi}}_{\boldsymbol{B}}) = 0.002$) in both simulation cases.

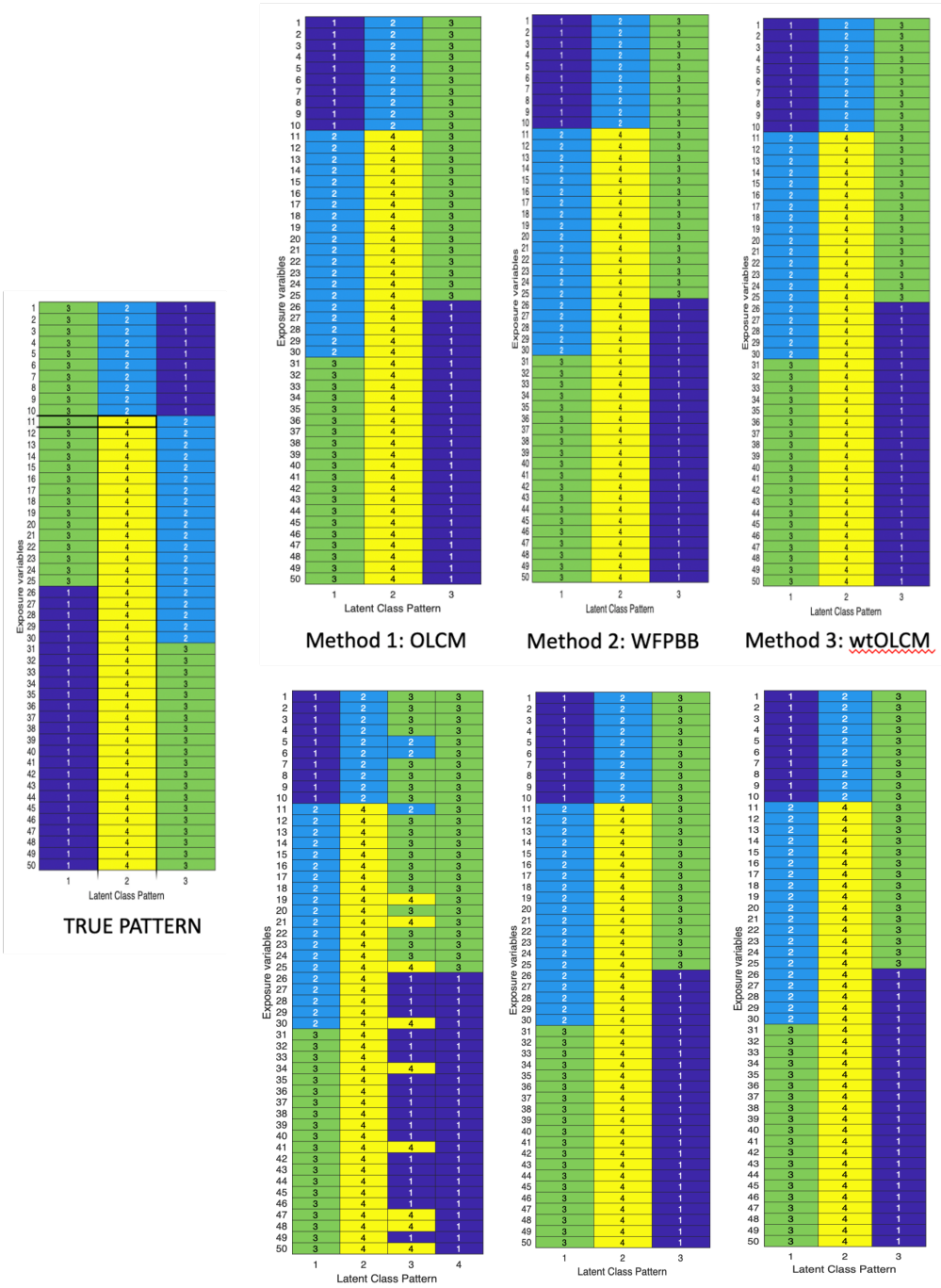

Fig. 1. Modal consumption patterns identified from respective models compared to truth. Method 1: overfitted latent class model, ignoring weights; Method 2: weighted finite population Bayesian bootstrap; Method 3: weighted overfitted latent class model. Top indicates pattern under simulation case A. 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 $\hat{\boldsymbol{\pi}}$ for the three approaches, summarized across 100 simulated datasets

Method	**MSE($\hat{\boldsymbol{\pi}}_{\boldsymbol{A}}$)**	**MSE($\hat{\boldsymbol{\pi}}_{\boldsymbol{B}}$)**	**Runtime (mins)**
Unweighted OLCM	1.49×10^{-2}	.016	18.3
WFPBB	4.31×10^{-4}	.006	19.2, parallelized across 200 cores
wtOLCM	1.32×10^{-4}	.002	20.9

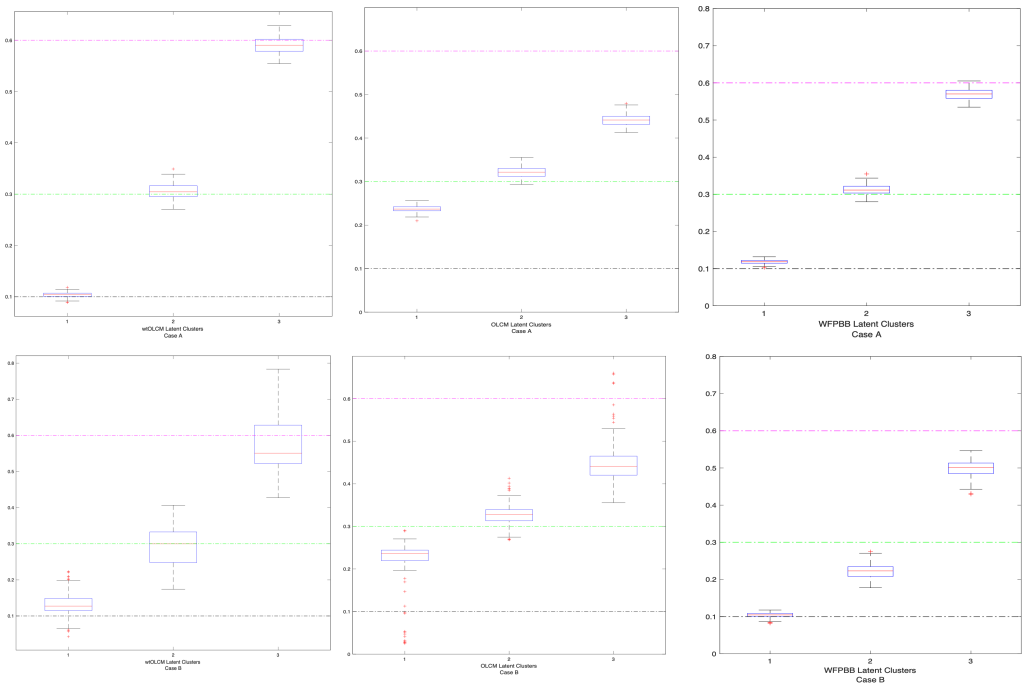

Fig. 2. Predicted population prevalence from unweighted and weighted estimation approaches. Expected prevalence for each respective cluster is 0.1 (black), 0.3 (green), 0.6 (magenta)

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

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, 2017*b*, 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

	Overall			Low-income		
Demographics	N	%	(SE)	N	%	(SE)
Race/Ethnicity						
Mexican	2632	9.2	(1.0)	1353	15.4	(1.7)
Other Hispanic	2026	6.3	(0.6)	950	9.7	(1.0)
Non-Hispanic White	7480	64.0	(1.7)	2390	47.6	(2.3)
Non-Hispanic Black	4471	11.3	(1.0)	1963	17.4	(1.5)
Non-Hispanic Asian	2282	5.8	(0.5)	609	5.6	(0.7)
Mixed/Other	716	3.4	(0.3)	296	4.3	(0.5)
Gender						
Male	10077	48.8	(0.5)	3517	45.5	(0.8)
Female	9530	51.2	(0.5)	4044	54.5	(0.8)
Age						
20-34 Years	4991	28.9	(0.8)	2129	35.7	(1.5)
35-49 Years	4801	25.5	(0.7)	1767	24.3	(0.9)
50-64 Years	5246	26.9	(0.6)	1976	23.9	(0.9)
65+ Years	4569	18.8	(0.6)	1689	16.1	(0.8)
AHEI2015 Score	14865	51.6	(0.3)	5823	49.2	(0.3)
Framingham 10YR Score	18226	8.1	(0.1)	6904	7.7	(0.2)
CVD Risk factors						
Hypertension	8260	32.6	0.9	2446	32.3	1.0
Obesity	8198	39.0	0.9	2589	41.4	1.2
Diabetes	2010	9.3	0.4	643	10.6	0.6
High Cholesterol	9906	75.1	0.8	2874	72.4	1.0
Smoker	3487	15.1	0.7	1605	24.4	1.4

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, $\boldsymbol{\pi}$, and the food item probability of consumption given assignment to specific latent class, $\boldsymbol{\theta_{j\cdot|k}}$, $j \in \{1, \ldots, p\},\ k \in \{1, \ldots, K\}$. 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

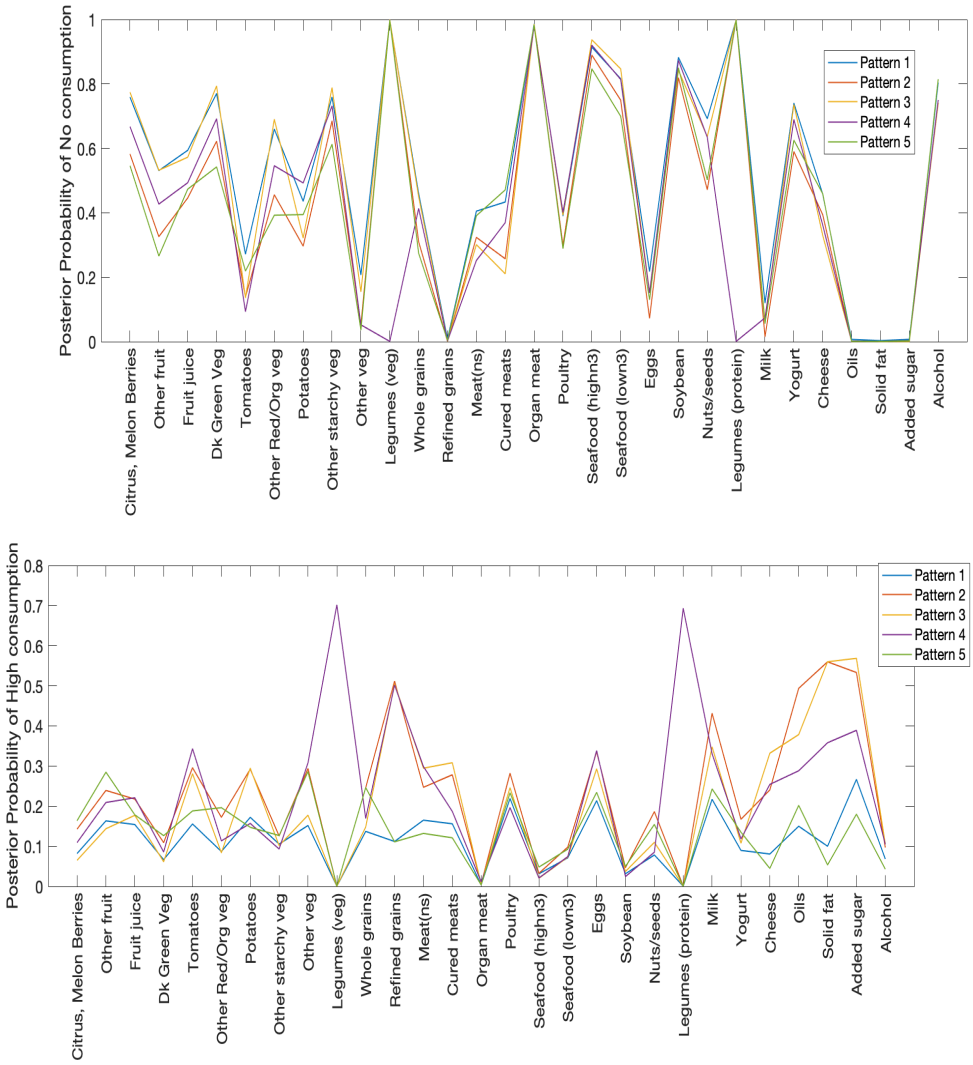

Fig. 3. Low-income Adult population - Model 2: (top) Posterior mean probability of no consumption of a given food item given membership in a specified pattern; (bottom) Posterior mean probability of high consumption of a given food item given membership in a specified pattern

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.

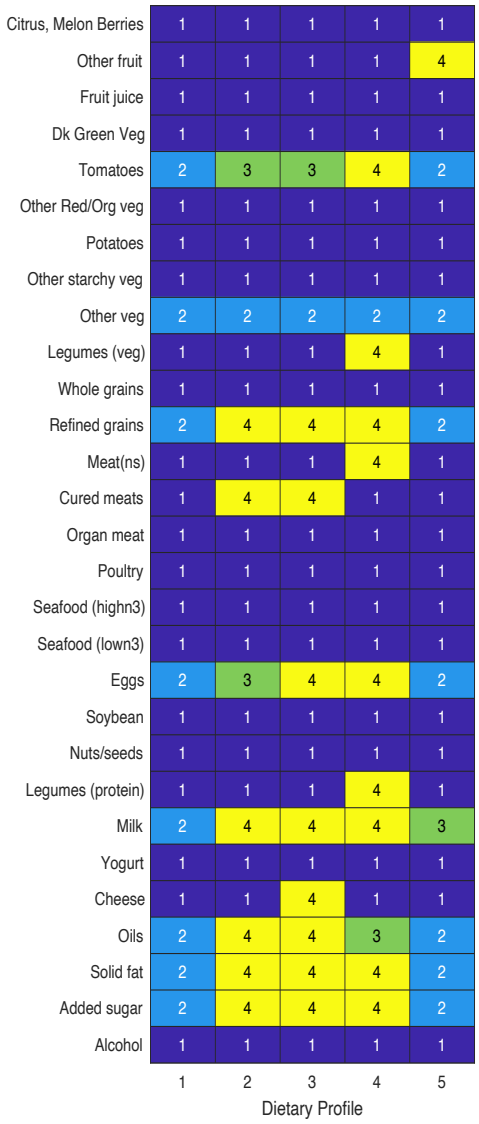

Fig. 4. Posterior mode of consumption pattern of dietary patterns for non-incarcerated adults living at or below the 130% poverty level. Numbers represent levels of consumption: 1= None, 2=Low, 3=Medium, 4=High

Table 4. Demographic distribution of Low-income Dietary patterns

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

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.

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

Amanatidis, Soumela, Mackerras, Dorothy and Simpson, Judy M. (2001). Comparison of two frequency questionnaires for quantifying fruit and vegetable intake. *Public health nutrition* **4**(2), 233–239.

Bahr, Peter Riley. (2007). Race and nutrition: an investigation of black-white differences in health-related nutritional behaviours. *Sociology of health & illness* **29**(6), 831–856.

Bhattacharya, Anirban and Dunson, David B. (2011). Sparse bayesian infinite factor models. *Biometrika*, 291–306.

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

(2017*a*). Food patterns equivalents intakes by americans: What we eat in america, nhanes 2003–2004 and 2013–2014. *Food Surveys Research Group*.

Bowman, SA, Clemens, JC, Friday, JE, Lynch, KL and Moshfegh, AJ. (2017*b*). Food patterns equivalents database 2013–2014: methodology and user guide. *United States Department of Agriculture: Beltsville, MD, USA*.

Bowman, SA, Clemens, JC, Friday, JE, Thoeric, RC and Moshfegh, AJ. (2016). Food patterns equivalents database 2011-2012: Methodology and user guide. *US Department of Agriculture Agricultural Research Service Web site. http://www. ars. usda. gov/SP2UserFiles/Place/80400530/pdf/fped/FPED_1112. pdf. Accessed July* **15**.

Bowman, SA, Clemens, JC, Shimizu, M, Friday, JE and Moshfegh, AJ. (2018). Food patterns equivalents database 2015–2016: methodology and user guide. *US Department of Agriculture*.

Briefel, Ronette R, Sempos, Christopeher T, Mcdowell, Margaret A, Chien, Sandy and Alaimo, Katherine. (1997). Dietary methods research in the third national health and nutrition examination survey: underreporting of energy intake. *The American journal of clinical nutrition* **65**(4), 1203S–1209S.

Brown, Arleen F, Liang, Li-Jung, Vassar, Stefanie D, Escarce, Jose J, Merkin, Sharon Stein, Cheng, Eric, Richards, Adam, Seeman, Teresa and Longstreth Jr, WT. (2018). Trends in racial/ethnic and nativity disparities in cardiovascular health among adults without prevalent cardiovascular disease in the united states, 1988 to 2014. *Annals of internal medicine* **168**(8), 541–549.

Chen, Te-Ching, Clark, Jason, Riddles, Minsun K, Mohadjer, Leyla K and Fakhouri, Tala HI. (2020). National health and nutrition examination survey, 2015- 2018:

sample design and estimation procedures. *American journal of epidemiology* **177**(11), 1279–1288.

Colditz, Graham A, Philpott, Sydney E and Hankinson, Susan E. (2016). The impact of the nurses' health study on population health: prevention, translation, and control. *American journal of public health* **106**(9), 1540–1545.

Committee, Dietary Guidelines Advisory *and others*. (2015). *Dietary guidelines for Americans 2015-2020*. Government Printing Office.

Daviglus, Martha L, Talavera, Gregory A, Avilés-Santa, M Larissa, Allison, Matthew, Cai, Jianwen, Criqui, Michael H, Gellman, Marc, Giachello, Aida L, Gouskova, Natalia, Kaplan, Robert C *and others*. (2012). Prevalence of major cardiovascular risk factors and cardiovascular diseases among hispanic/latino individuals of diverse backgrounds in the united states. *Jama* **308**(17), 1775–1784.

De Vito, R, Lee, Yuan Chin Amy, Parpinel, M, Serraino, D, Olshan, Andrew Fergus, Zevallos, Jose Pedro, Levi, F, Zhang, Zhuo Feng, Morgenstern, H, Garavello, W *and others*. (2019). Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology (Cambridge, Mass.)* **30**(1), 93.

De Vito, Roberta, Stephenson, Briana, Sotres-Alvarez, Daniela, Siega-Riz, Anna-Maria, Mattei, Josiemer, Parpinel, Maria, Peters, Brandilyn A, Bainter, Sierra A, Daviglus, Martha L, Van Horn, Linda *and others*. (2022). Shared and ethnic background site-specific dietary patterns in the hispanic community health study/study of latinos (hchs/sol). *medRxiv*.

Dong, Qi, Elliott, Michael R and Raghunathan, Trivellore E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey methodology* **40**(1), 29.

Fahey, Michael T, Thane, Christopher W, Bramwell, Gemma D and Coward, W Andy. (2007). Conditional gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(1), 149–166.

Fahlman, Mariane M, McCaughtry, Nate, Martin, Jeffrey and Shen, Bo. (2010). Racial and socioeconomic disparities in nutrition behaviors: targeted interventions needed. *Journal of nutrition education and behavior* **42**(1), 10–16.

Frühwirth-Schnatter, Sylvia. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**(453), 194–209.

Gao, Chuan, McDowell, Ian C, Zhao, Shiwen, Brown, Christopher D and Engelhardt, Barbara E. (2016). Context specific and differential gene co-expression networks via bayesian biclustering. *PLoS computational biology* **12**(7), e1004791.

Gunawan, David, Panagiotelis, Anastasios, Griffiths, William and Chotikapanich, Duangkamon. (2020). Bayesian weighted inference from surveys. *Australian & New Zealand Journal of Statistics* **62**(1), 71–94.

Haraldsdóttir, J. (1993, Oct). Minimizing error in the field: quality control in dietary surveys. *Eur J Clin Nutr* **47 Suppl 2**, S19–24.

Harrington, Janas M, Dahly, Darren L, Fitzgerald, Anthony P, Gilthorpe, Mark S and Perry, Ivan J. (2014). Capturing changes in dietary patterns among older adults: a latent class analysis of an ageing irish cohort. *Public health nutrition* **17**(12), 2674–2686.

Hirvonen, T, Männistö, S, Roos, E and Pietinen, P. (1997). Increasing prevalence of

underreporting does not necessarily distort dietary surveys. *European journal of clinical nutrition* **51**(5), 297–301.

Hjort, Nils Lid, Holmes, Chris, Müller, Peter and Walker, Stephen G. (2010). *Bayesian nonparametrics*, Volume 28. Cambridge University Press.

Keshteli, Ammar Hassanzadeh, Feizi, Awat, Esmaillzadeh, Ahmad, Zaribaf, Fatemeh, Feinle-Bisset, Christine, Talley, Nicholas J and Adibi, Peyman. (2015). Patterns of dietary behaviours identified by latent class analysis are associated with chronic uninvestigated dyspepsia. *British Journal of Nutrition* **113**(5), 803–812.

Klesges, Robert C, Eck, Linda H and Ray, JoAnne W. (1995). Who underreports dietary intake in a dietary recall? evidence from the second national health and nutrition examination survey. *Journal of consulting and clinical psychology* **63**(3), 438.

Krebs, Charles J. (1989). *Ecological Methodology*. Harper Collins Publishers.

Kunihama, Tsuyoshi, Herring, AH, Halpern, CT and Dunson, DB. (2016). Nonparametric bayes modeling with sample survey weights. *Statistics & probability letters* **113**, 41–48.

Lanza, Stephanie T, Collins, Linda M, Lemmon, David R and Schafer, Joseph L. (2007). Proc lca: A sas procedure for latent class analysis. *Structural equation modeling: a multidisciplinary journal* **14**(4), 671–694.

Linzer, Drew A and Lewis, Jeffrey B. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of statistical software* **42**(1), 1–29.

Liu, Jun S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

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

Bankole A and Chai, Haitao. (2019). Statistical analysis of zero-inflated nonnegative continuous data: a review. *Statistical Science* **34**(2), 253–279.

Marttinen, Pekka, Pirinen, Matti, Sarin, Antti-Pekka, Gillberg, Jussi, Kettunen, Johannes, Surakka, Ida, Kangas, Antti J, Soininen, Pasi, O'Reilly, Paul, Kaakinen, Marika *and others*. (2014). Assessing multivariate gene-metabolome associations with rare variants using bayesian reduced rank regression. *Bioinformatics* **30**(14), 2026–2034.

Mattei, Josiemer, Sotres-Alvarez, Daniela, Daviglus, Martha L, Gallo, Linda C, Gellman, Marc, Hu, Frank B, Tucker, Katherine L, Willett, Walter C, Siega-Riz, Anna Maria, Van Horn, Linda *and others*. (2016). Diet quality and its association with cardiometabolic risk factors vary by hispanic and latino ethnic background in the hispanic community health study/study of latinos. *The Journal of nutrition* **146**(10), 2035–2044.

Medvedovic, Mario and Sivaganesan, Siva. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**(9), 1194–1206.

Miller, Jeffrey W and Harrison, Matthew T. (2013). A simple example of dirichlet process mixture inconsistency for the number of components. *Advances in neural information processing systems* **26**.

Mokdad, Ali H, Ballestros, Katherine, Echko, Michelle, Glenn, Scott, Olsen, Helen E, Mullany, Erin, Lee, Alex, Khan, Abdur Rahman, Ahmadi, Alireza, Ferrari, Alize J *and others*. (2018). The state of us health, 1990-2016: burden of diseases, injuries, and risk factors among us states. *Jama* **319**(14), 1444–1472.

Muthén, Bengt and Shedden, Kerby. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics* **55**(2), 463–469.

Muthén, Linda K and Muthén, Bengt. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

National Center for Health Statistics, Division of the National Health and Surveys, Nutrition Examination. (2018). National health and nutrition examination survey: Analytic guidelines, 2011-2014 and 2015-2016. *Technical Report*, Centers for Disease Control.

Nylund, Karen L, Asparouhov, Tihomir and Muthén, Bengt O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling: A multidisciplinary Journal* **14**(4), 535–569.

Ohlhorst, Sarah D, Russell, Robert, Bier, Dennis, Klurfeld, David M, Li, Zhaoping, Mein, Jonathan R, Milner, John, Ross, A Catharine, Stover, Patrick and Konopka, Emily. (2013). Nutrition research to affect food and a healthy life span. *The Journal of nutrition* **143**(8), 1349–1354.

Padmadas, Sabu S, Dias, José G and Willekens, Frans J. (2006). Disentangling women's responses on complex dietary intake patterns from an indian cross-sectional survey: a latent class analysis. *Public Health Nutrition* **9**(2), 204–211.

Patterson, Blossom H, Dayton, C Mitchell and Graubard, Barry I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association* **97**(459), 721–741.

Pelleg, D. (2000). Extending k-means with efficient estimation of the number of clusters in icml. In: *Proceedings of the 17th international conference on machine learning*. pp. 277–281.

Poslusna, Kamila, Ruprich, Jiri, de Vries, Jeanne HM, Jakubikova, Marie and van't Veer, Pieter. (2009). Misreporting of energy and micronutrient intake estimated by food

records and 24 hour recalls, control and adjustment methods in practice. *British Journal of Nutrition* **101**(S2), S73–S85.

Raghunathan, Trivellore E, Reiter, Jerome P and Rubin, Donald B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics* **19**(1), 1.

Rao, JNK and Thomas, D Roland. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 213–269.

Reedy, Jill, Wirfält, Elisabet, Flood, Andrew, Mitrou, Panagiota N, Krebs-Smith, Susan M, Kipnis, Victor, Midthune, Douglas, Leitzmann, Michael, Hollenbeck, Albert, Schatzkin, Arthur *and others*. (2010). Comparing 3 dietary pattern methods—cluster analysis, factor analysis, and index analysis—with colorectal cancer risk: the nih–aarp diet and health study. *American journal of epidemiology* **171**(4), 479–487.

Roth, Gregory A, Johnson, Catherine O, Abate, Kalkidan Hassen, Abd-Allah, Foad, Ahmed, Muktar, Alam, Khurshid, Alam, Tahiya, Alvis-Guzman, Nelson, Ansari, Hossein, Ärnlöv, Johan *and others*. (2018). The burden of cardiovascular diseases among us states, 1990-2016. *JAMA cardiology* **3**(5), 375–389.

Rousseau, Judith and Mengersen, Kerrie. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 689–710.

Runcie, Daniel E and Mukherjee, Sayan. (2013). Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics* **194**(3), 753–767.

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

BERT, ADELIN *and others*. (2017). Stability-based validation of dietary patterns obtained by cluster analysis. *Nutrition journal* **16**(1), 1–13.

SAVITSKY, TERRANCE D AND TOTH, DANIELL. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics* **10**(1), 1677–1708.

SCHWEDHELM, CAROLINA, IQBAL, KHALID, KNÜPPEL, SVEN, SCHWINGSHACKL, LUKAS AND BOEING, HEINER. (2018, 02). Contribution to the understanding of how principal component analysis–derived dietary patterns emerge from habitual data on food consumption. *The American Journal of Clinical Nutrition* **107**(2), 227–235.

SI, YAJUAN, PILLAI, NATESH S AND GELMAN, ANDREW. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis* **10**(3), 605–625.

SKINNER, CHRIS AND WAKEFIELD, JON. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science* **32**(2), 165–175.

SOTRES-ALVAREZ, DANIELA, HERRING, AMY H AND SIEGA-RIZ, ANNA MARIA. (2010). Latent class analysis is useful to classify pregnant women into dietary patterns. *The Journal of nutrition* **140**(12), 2253–2259.

SOTRES-ALVAREZ, DANIELA, SIEGA-RIZ, ANNA MARIA, HERRING, AMY H, CARMICHAEL, SUZAN L, FELDKAMP, MARCIA L, HOBBS, CHARLOTTE A, OLSHAN, ANDREW F AND STUDY, NATIONAL BIRTH DEFECTS PREVENTION. (2013). Maternal dietary patterns are associated with risk of neural tube and congenital heart defects. *American journal of epidemiology* **177**(11), 1279–1288.

STEPHENS, MATTHEW. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 795–809.

Stephenson, Briana JK, Herring, Amy H and Olshan, Andrew. (2020*a*). Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association* **115**(530), 521–537.

Stephenson, Briana JK, Sotres-Alvarez, Daniela, Siega-Riz, Anna-Maria, Mossavar-Rahmani, Yasmin, Daviglus, Martha L, Van Horn, Linda, Herring, Amy H and Cai, Jianwen. (2020*b*). Empirically derived dietary patterns using robust profile clustering in the hispanic community health study/study of latinos. *The Journal of nutrition* **150**(10), 2825–2834.

Stephenson, Briana Joy Kennedy and Willett, Walter C. (2022). Racial/ethnic heterogeneity in diet of low-income adult women in the united states: Results from national health and nutrition examination surveys 2011-2018. *medRxiv*.

Tooze, Janet A, Subar, Amy F, Thompson, Frances E, Troiano, Richard, Schatzkin, Arthur and Kipnis, Victor. (2004). Psychosocial predictors of energy underreporting in a large doubly labeled water study. *The American journal of clinical nutrition* **79**(5), 795–804.

Tourangeau, Roger, Edwards, Brad, Johnson, Timothy P, Wolter, Kirk M and Bates, Nancy. (2014). *Hard-to-survey populations*. Cambridge University Press.

U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. (2020, July). Us department of agriculture: Choosemyplate.

U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. (2022, April). Healthy eating index for americans.

Van Havre, Zoé, White, Nicole, Rousseau, Judith and Mengersen, Kerrie. (2015). Overfitting bayesian mixture models with an unknown number of components. *PloS one* **10**(7), e0131739.

Vermunt, Jeroen K. (2002). Comments on "latent class analysis of complex sampling data" by. *Journal of the American Statistical Association*.

Vermunt, Jeroen K and Magidson, Jay. (2007). Latent class analysis with sampling weights: A maximum-likelihood approach. *Sociological methods & research* **36**(1), 87–111.

Wall, Hilary K, Ritchey, Matthew D, Gillespie, Cathleen, Omura, John D, Jamal, Ahmed and George, Mary G. (2018). Vital signs: prevalence of key cardiovascular disease risk factors for million hearts 2022—united states, 2011–2016. *Morbidity and Mortality Weekly Report* **67**(35), 983.

Wang, Dong D, Leung, Cindy W, Li, Yanping, Ding, Eric L, Chiuve, Stephanie E, Hu, Frank B and Willett, Walter C. (2014). Trends in dietary quality among adults in the united states, 1999 through 2010. *JAMA internal medicine* **174**(10), 1587–1595.

White, Arthur and Murphy, Thomas Brendan. (2014). Bayeslca: An r package for bayesian latent class analysis. *Journal of Statiscal Software* **61**(13), 1–28.

Wirfält, AK Elisabet and Jeffery, Robert W. (1997). Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. *Journal of the American Dietetic Association* **97**(3), 272–279.

Yuan, Changzheng, Spiegelman, Donna, Rimm, Eric B, Rosner, Bernard A, Stampfer, Meir J, Barnett, Junaidah B, Chavarro, Jorge E, Subar, Amy F, Sampson, Laura K and Willett, Walter C. (2017). Validity of a dietary questionnaire assessed by comparison with multiple weighed dietary records or 24-hour recalls. *American journal of epidemiology* **185**(7), 570–584.

Zhang, Fang Fang, Liu, Junxiu, Rehm, Colin D, Wilde, Parke, Mande, Jerold R and Mozaffarian, Dariush. (2018). Trends and disparities in diet quality among us adults

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

Zhou, Hanzhi, Elliott, Michael R and Raghunathan, Trviellore E. (2016). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics* **72**(1), 242–252.

Zipf, George, Chiappa, Michele, Porter, Kathryn S, Ostchega, Yechiam, Lewis, Brenda G and Dostal, Jennifer. (2013). Health and nutrition examination survey plan and operations, 1999-2010. *Technical Report*, National Center for Health Statistics.

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