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

Briana J.K. Stephenson

Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

E-mail: bstephenson@hsph.harvard.edu

Francesca Dominici

Harvard T.H. Chan School of Public Health, Department of Biostatistics, Boston, Massachusetts, USA

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 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.

Keywords: 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 about 500,000 deaths, with 84% of those deaths due to cardiovascular disease (CVD)(Mokdad et al., 2018; Roth et al., 2018). The impact of poor diets on CVD risk disproportionately impacts low-income and racial minority populations (Brown et al., 2018; Daviglus et al., 2012; Zhang et al., 2018; Wang et al., 2014; Bahr, 2007; Fahlman et al., 2010). Understanding the dietary consumption behaviors that contribute to poor health in these target populations may help in tailoring interventions and appropriate resources to improve their nutritional health.

Through the implementation of complex survey designs and targeted recruitment strategies, researchers are able to obtain 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 et al., 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 et al., 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 et al., 2018; Bowman et al., 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 et al., 2017; WirfÄLt and Jeffery, 1997; Reedy et al., 2010). Yet, analysis often results in only explaining a fraction of dietary consumption habits present in the data.

Latent class or finite mixture models 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.

Latent class models (LCM) 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 et al., 2007; Linzer and Lewis, 2011; Muthén and Shedden, 1999). Bayesian approaches are also available through an R package (BayesLCA), but is limited to binary consumption responses (White and Murphy, 2014). These models are reflective of the study data being applied. Therefore, a larger demographic present in the study can easily mask dietary habits of smaller-sized demographics that may deviate from the majority habit. Sampling weights present in complex survey designs can correct for this by enhancing the study population to mimic that of the larger general population, allowing for underrepresented and oversampled groups to maintain identifiability. None of these standard packages 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 the frequentist framework with

matrix inversion and computational demand issues when handling dietary intake data that is often 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 et al., 2020b; Mattei et al., 2016).

Patterson et al. (2002) and Vermunt and Magidson (2007) have implemented survey-weighted approaches to latent class models, generated under a frequentist framework. Yet, this can incur a huge computational burden to analysis when the number of subjects and food items increase, as well as the complexity of polytomous response patterns from each food. Dietary intake data that is often sparse, resulting in zero-inflation presents convergence issues for food items that are rarely or occasionally consumed. Additionally, traditional latent class models operated under the assumption of a known number of latent classes (or patterns) to fit the model, which in practice is seldom known (Nylund et al., 2007). This results in multiple model testing and fitting based on stability, clinical interpretability, and biological interactions among diet components (Padmadas et al., 2006; Sotres-Alvarez et al., 2010; Harrington et al., 2014; Keshteli et al., 2015).

1.4. Potential Solutions in Bayesian Nonparametrics

Bayesian nonparametrics offers a more efficient solution that is able to:

- accommodate the complex high dimensionality of dietary intake data
- handle an infinitely growing population
- reduce multiple model testing in determining the appropriate number of patterns
- preserve model stability in the presence of sparsely consumed foods
- integrate prior information with observed data.

These features improve parameter estimation and subsequent population inference (Hjort et al., 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 the complex survey design, and even fewer have applied for use in nutritional studies.

The reduction of multiple testing required in frequentist settings to determine the appropriate number of latent groups has been applied using Bayesian nonparametric approaches with applications primarily in genetics and bioinformatics(Bhattacharya and Dunson, 2011; Runcie and Mukherjee, 2013; Gao et al., 2016; Marttinen et al., 2014; Pelleg, 2000). The application of survey data in Bayesian nonparametrics has centered

most of the focus on generating inference from derived population-based estimates (Si et al., 2015; Savitsky and Toth, 2016; Gunawan et al., 2020). Bayesian nonparametric mixture models have been applied in nutritional settings but has not required sampling weights into the parameter estimation of the model (Fahey et al., 2007; De Vito et al., 2019; Stephenson et al., 2020a). Stephenson et al. (2020b) applied a Bayesian nonparametric mixture model to diet survey data, but the weights were applied after parameters were estimated from the sampling algorithm. Kunihama et al. (2016) et al. 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 but added the following contributions: 1) implemented the model using an overfitted finite mixture, which is asymptotically similar to the Dirichlet Process mixture model; 2) extended and integrated the works of Kunihama et al. (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 adults in the United States from 2011-2018; 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 describes the National Health and Nutrition Examination Survey. Section 4 presents results of the method applied to the National Health and Nutrition Examination Survey. Section 5 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 within a survey sample that share common behaviors amongst a set of observed nominal variables (Van Havre et al., 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 an overwhelming number of latent classes, empty classes are able to drop out of the model during Markov chain Monte Carlo Gibbs sampling algorithm and nonempty classes will remain separating the participants into one of the derived latent classes (patterns). Let $y_{i\cdot} = (y_{i1}, \ldots, y_{ip})$ denote the set of p observed food items consumed at an observed level. The overfitted structure is also asymptotically equivalent to the Dirichlet process model allowing additional flexibility within a Bayesian nonparametric framework (Van Havre et al., 2015). Let π_k denote the probability of a subject being assigned dietary pattern k. 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 membership to latent cluster k. Finally, let K denote the maximum number of dietary patterns fitted to the population, and z_i denote the latent cluster assignment of an individual $i \in (1, \ldots, n)$ from the sampled

population to one of the dietary patterns $k \in (1, \ldots, K)$. The subject-specific likelihood for categorical dietary data, under the overfitted latent class model, is defined as

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

The likelihood shares the same structure as the traditional latent class analysis model, but K is fixed to a large number that asymptotically simulates an infinite mixture model (Van Havre et al., 2015).

In a typical Bayesian estimation of these parameters, the probability parameters are dependent on the number of observed individuals classified to a given cluster or consumption category. For example, exploiting the convenience of conjugacy, the probability vector, $\pi. = (\pi_1, \ldots, \pi_K)$, is estimated with a Dirichlet distribution, such that

$$\begin{aligned} \pi = (\pi_1, \pi_2, \ldots, \pi_K) &\sim Dir(\alpha_1, \ldots, \alpha_K) \\ (\pi_1, \ldots, \pi_K|y_{i\cdot}, z_i) &\sim Dir(\alpha_1 + \sum_{i=1}^{n} 1(z_i = 1), \ldots, \alpha_K + \sum_{i=1}^{n} 1(z_i = K)) \end{aligned} \tag{2}$$

where $(\alpha_1, \ldots, \alpha_K)$ are hyperparameters for each respective latent cluster. With no prior knowledge on the number of clusters one would assume a flat Dirichlet prior, where $\alpha_1 = \alpha_2 = \ldots = \alpha_K = \alpha$. This hyperparameter moderates the rate of nonempty cluster growth. The smaller the hyperparameter the slower the nonempty clusters will form. Similarly, $\{\theta_{j\cdot|k)} = (\theta_{j1|k}, \ldots, \theta_{jd_j|k}\}_{j=1}^{p}$ for all $k \in (1, 2, \ldots, K)$. As both parameters are categorical distributions, we would use a Dirichlet prior distribution for estimation of these probability parameters. Let α denote the hyperparameter of $\pi = (\pi_1, \ldots, \pi_K)$ and γ denote the hyperparameter of $\theta_{j\cdot|k} = (\theta_{j1|k}, \ldots, \theta_{jd|k})$.

Under a Bayesian framework, the incorporation of survey weights serves as a natural extension to the latent class model. As described in Kunihama et al. (2016), the information used to update each model parameter is enhanced with the survey weight simulating a pseudo-like population that is similar in size and structure to that of the target population. Using a normalization constant the weights are adjusted to sum to the total population size. This allows dietary patterns to form in accordance with the representative target population without consideration for changes that can occur in a given study population's size and composition. Sampling variability should be considered. Otherwise, uncertainty surrounding model estimation could be biased. We instead propose an approach similar to Savitsky and Toth (2016) in achieving the pseudo-like population size, where the weights are normalized with respect to the size of the study population, allowing sampling variability for different sized study populations.

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 population size. With this newly defined κ, we can update the conditional posterior probability vector, $\pi = (\pi_1, \ldots, \pi_K)$ where,

$$(\pi_1, \ldots, \pi_K | y_{i\cdot}, z_i) \sim Dir\left(\alpha_1 + \frac{1}{\kappa}\sum_{i=1}^{n} w_i \times 1(z_i = 1), \ldots, \alpha_K + \frac{1}{\kappa}\sum_{i=1}^{n} w_i \times 1(z_i = K)\right). \tag{3}$$

Similarly, the distribution of each dietary pattern, $f(y_i|z_i = h) = \prod_{j=1}^{p}\prod_{c=1}^{d_j}\theta_{jc|h}^{1(y_i=c|z_i=h)}$ would update based on the pseudo-weighted proportion of participants that share the same behaviors.

$$\theta_{j\cdot|h} \sim Dir_{d_j}\left(\gamma + \frac{1}{\kappa}\sum_{i:z_i=h} w_i \times 1(y_{ij} = c)\right), \text{ for all } j \in (1, \ldots, p), k \in (1, \ldots, K) \tag{4}$$

3. National Health and Nutrition Examination Survey (NHANES)

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 7 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 et al., 2015; usd, 2020; Bowman et al., 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, 2017-2018. Consumption levels were derived by segmenting the data into no consumption (none=0%) and tertiles of positive consumption (Liu et al., 2019; Sotres-Alvarez et al., 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 et al., 2020).

Demographically, the adult participants included for study (Table 1), reflected a majority Non-Hispanic White population (64.0±1.7%). Gender was relatively even amongst the overall population, but indicated a slightly higher representation ($54.5 \pm 0.8\%$) for women amongst low-income adults. The study population favors a slightly younger age group (20-34 years), with those aged 65 and over the smallest of representations. The low-income group reported a slightly less healthier AHEI-2015 dietary score and a lower

Table 1. 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

Framingham 10 year risk score, indicating a greater probability of a CVD outcome in the next 10 years.

4. Application to NHANES Survey 2011-2018

4.1. Statistical Analysis

We considered two different applications of the weighted overfitted latent class model. Model 1 was fit on the entire NHANES adult population and model 2 was fit with NHANES adult low-income population. The normalization constant for each respective model was calculated based on the sum of the sampled weights and the total sampled population included for analysis ($n_{gen} = 19607, n_{low} = 7561$). These two estimates yielded a normalization constant of $\kappa_{gen} = 1.38 \times 10^4$ and $\kappa_{low} = 9.79 \times 10^3$, respectively.

Each model was overfitted with 50 clusters. 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 cluster assignment, π, and food item probability of consumption given assignment to specific cluster, θ. Random permutation sampler was performed to encourage mixing. Dietary weights were calibrated and normalized for inclusion in analysis. A conservative hyperprior was used

to moderate the rate of cluster growth, where $\alpha = \frac{1}{K}$, where K is the number of classes overfitted.

A common consequence in mixture modeling is label switching. To resolve label switching, hierarchical clustering was performed on a similarity matrix of 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 are defined as any cluster with a posterior probability weight greater than 0.05. Dietary patterns were defined by identifying the mode level of consumption (e.g. the consumption level corresponding to the level with the highest posterior median probability for a given food item), as well as the foods most likely to be consumed at a given level within each dietary pattern.

All data included for this study and codes to reproduce the derived dataset and perform subsequent analysis is made available on a public GitHub repository: `http://www.github.com/bjks10/NHANES_wtofm`. Derived dataset from NHANES was generated in SAS 9.4. Statistical analysis and figures were performed in MATLAB 2021a. Table summaries were generated in R version 4.0.2.

4.2. Results

The weighted overfitted latent class model identified six nonempty clusters in the total adult population (model 1) and five nonempty clusters in the low-income adult population. Figure 1 illustrates the posterior mean estimates of the probability of high consumption or no consumption given membership to a given dietary profile. From this figure, we can see which foods were strongly favored to be consumed at a high level. While some foods follow similar patterns (e.g. poultry, eggs, organ meat, seafood, alcohol) amongst all dietary profiles. Some profiles distinctly stood out from the other profiles. For example, profile 3 strongly favors consuming tomatoes and legumes at a high consumption level. Dietary profile 4 favored a high consumption of refined grains, cured meats, cheese, fats, oils, and added sugar. Profiles 1 and 5 had the lowest probabilities for consumption of cheese, oils, and fats at a high consumption level. Dietary profile 1 had the highest probabilities of no consumption of most fruits and vegetables. The very low probabilities of no consumption across all profiles for refined grains, oils, solid fats, and added sugar imply a general nonzero consumption by all US adults.

Figure 2 illustrates the posterior mean estimates of the probability of foods being consumed at the high level or not at all for the low-income adult population. Similar to what we saw in model 1, dietary profile 4 distinctly had a high probability of high consumption of legumes. Profiles 2 and 3 had the highest probability of foods being consumed at a high consumption level for potatoes, cured meats, oils, solid fat, added sugar. Consistent with model 1, refined grains, oils, solid fat and added sugar favor a general positive consumption by all adults. Profiles 1 and 3 favored a high probability of no consumption of fruits and vegetables.

Comparing more closely the posterior modes of consumption for each dietary profile, we are able to identify similarities and differences between the two models (Figure 3). Thirteen foods shared a mode of no consumption across both models: citrus/melon/berries, fruit juice, dark green vegetables, other red/orange vegetables, other

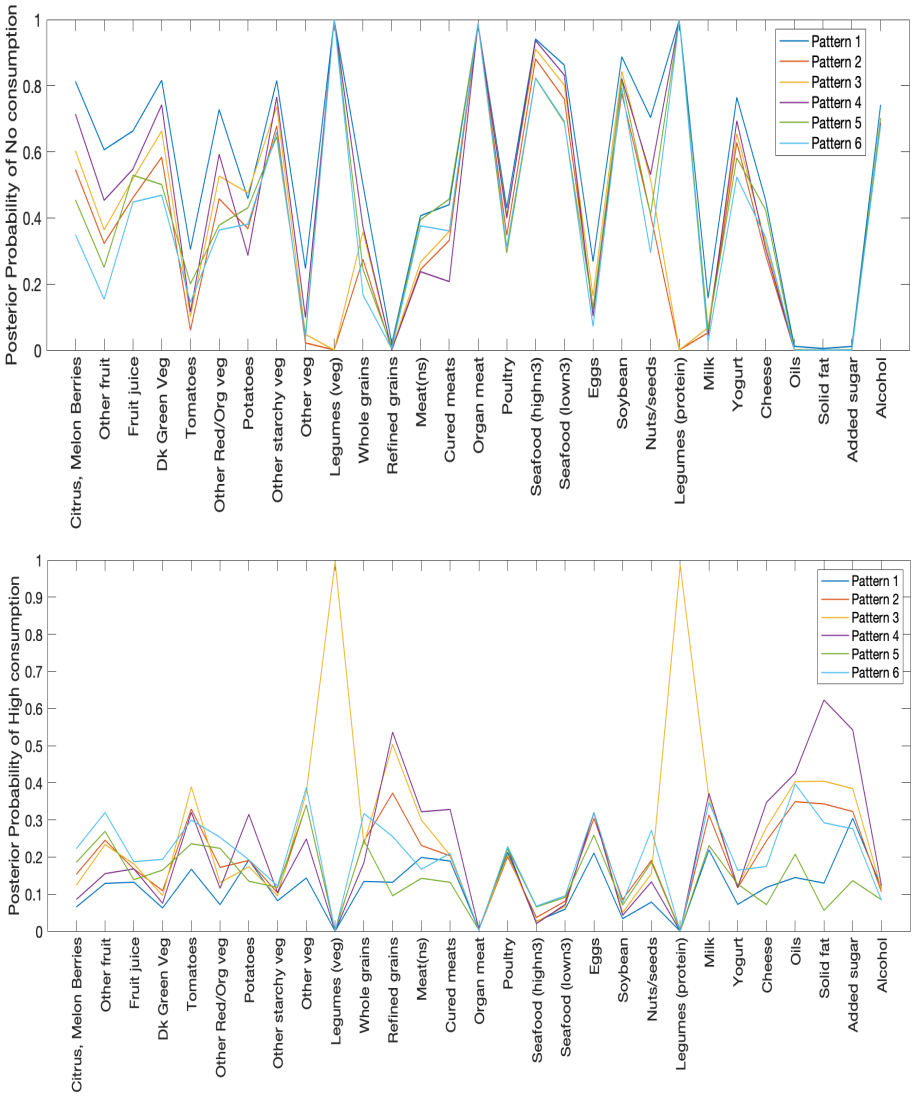

Fig. 1. Overall Adult population - Model 1: (top) Posterior mean probability of high consumption of a given food item given membership into a specified profile; (bottom) Posterior mean probability of no consumption of a given food item given membership into a specified profile

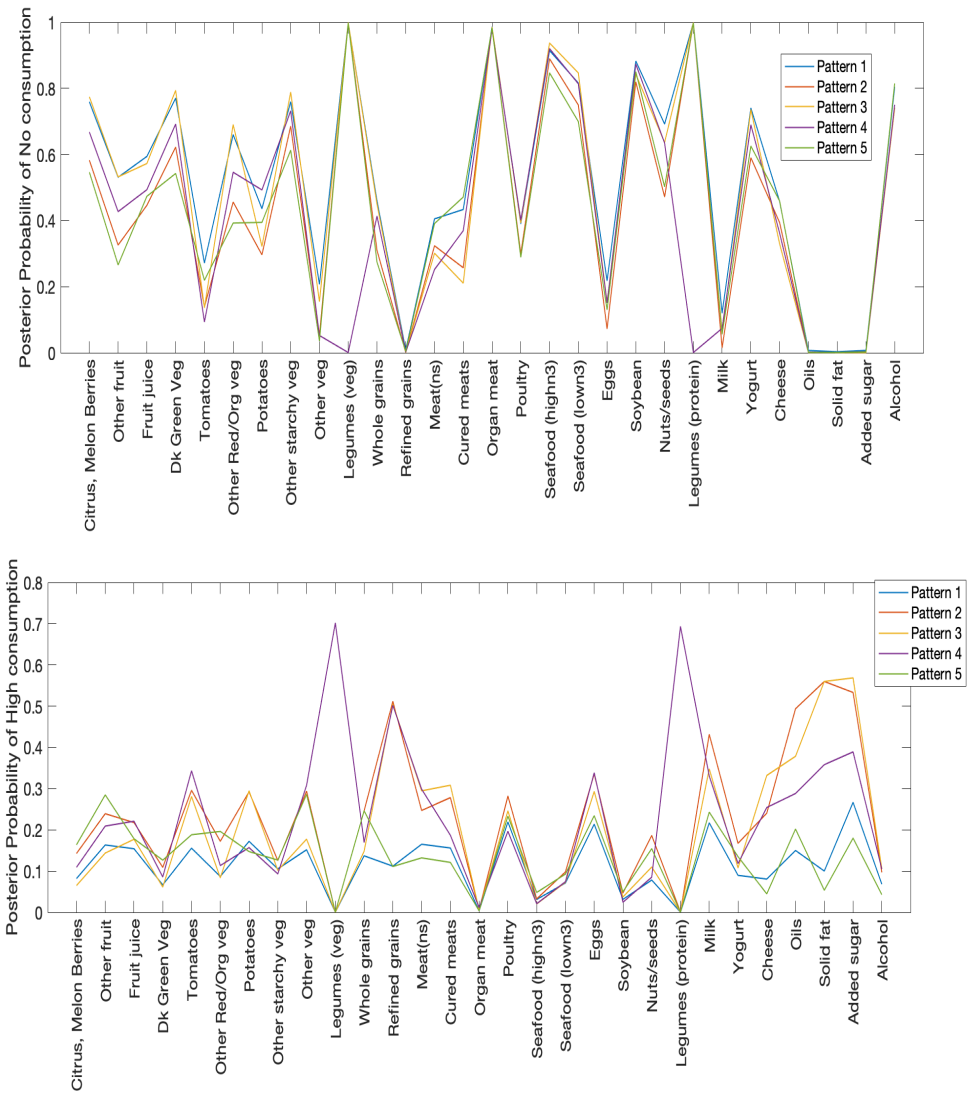

Fig. 2. Low-income Adult population - Model 2: (top) Posterior mean probability of high consumption of a given food item given membership into a specified profile; (bottom) Posterior mean probability of no consumption of a given food item given membership into a specified profile

starchy vegetables, organ meat, poultry, seafood (high-n3), seafood (low-n3), soybean, nuts/seeds, yogurt, alcohol. A shared mode of low consumption was present in model 2 (low-income adult) for other vegetables, but variation of different consumption modes was found in model 1 (overall adult). The first pattern from model 1, which we will denote M1P1, mimics that of M2P1 and closely resembles M2P5. M2P5 differed in consumption modes for only 2 foods, other fruit (high consumption) and milk (medium consumption), compared to pattern 1 which favored a non-consumption mode under both models. M1P3 and M2P4 also share similar consumption modes, but has a lower consumption mode of other vegetables and oils in model 2. While consumption modes were shared amongst the profiles in each of the models. Model 1 had a high consumption of potatoes (M1P4) and medium/high consumption for whole grain (M1P5, M1P6) that was not reflected in any of the 5 dietary profiles of model 2.

Table 2 provides a summary of the demographics for participants assigned to each dietary profile of model 1. Amongst the general adult population, participants assigned to profile 3 had the highest AHEI-2015 scores (57.3 ± 0.5). This is the only profile that reflected a modal high consumption of legumes (Figure 3). Participants assigned to profile 4 had the lowest AHEI2015 score (43.3 ± 0.2). This profile favored a modal high consumption of refined grains, potatoes, cured meats, cheese, oils, fats, and sugars. Demographically, male participants were more heavily represented in these two profiles. Female participants were more heavily represented in the remaining profiles. Amongst CVD risk factors, we see that profile 4 also had the highest proportion of participants with obesity. Profile 1, which had modes of only none or low consumption of foods had the highest proportion of current smokers. A more detailed illustration at the differences in distribution of consumption of all levels of the 29 food items is provided in Supplementary Materials.

Table 3 provides a summary of the demographics for participants assigned to each dietary profile of model 2. Amongst the low-income adult population, participants assigned to profile 5 had the highest average AHEI2015 score (57.4 ± 0.6). Similar to what we saw in model 1, profile 5 favored a high consumption of other fruit, but a low consumption of refined grains and no consumption of meats. Profile 3 had the lowest average AHEI2015 score (41.1 ± 0.3). Similar to model 1, this profile favored a high consumption of refined grains, cured meats, eggs, cheese, fats, oils, and sugars. Also consistent with model 1, demographically, we see that those in M2P5 was overwhelmingly represented by male participants. However, M2P3 was represented mostly by female participants. Non-Hispanic White participants held the majority of all profiles in model 2. However, profile 4 which uniquely favored a high consumption of legumes was the only profile where minority participants were more represented than non-Hispanic White participants.

In a structure such as a latent class or finite mixture model, participants assigned to a pattern are expected to share consumption habits with all of the participants in each cluster. As a nested cohort, we are able to get a better examination of pattern masking looking at low-income adult population subset. Figure 4 illustrates how the exclusion of adult participants not identified as "low-income" changed with each respective model. With darker hues indicating a high level of concordance, low-income adults assigned to profile 4 in model 2 were dispersed exclusively across profiles 2 and 3 in model 1. A

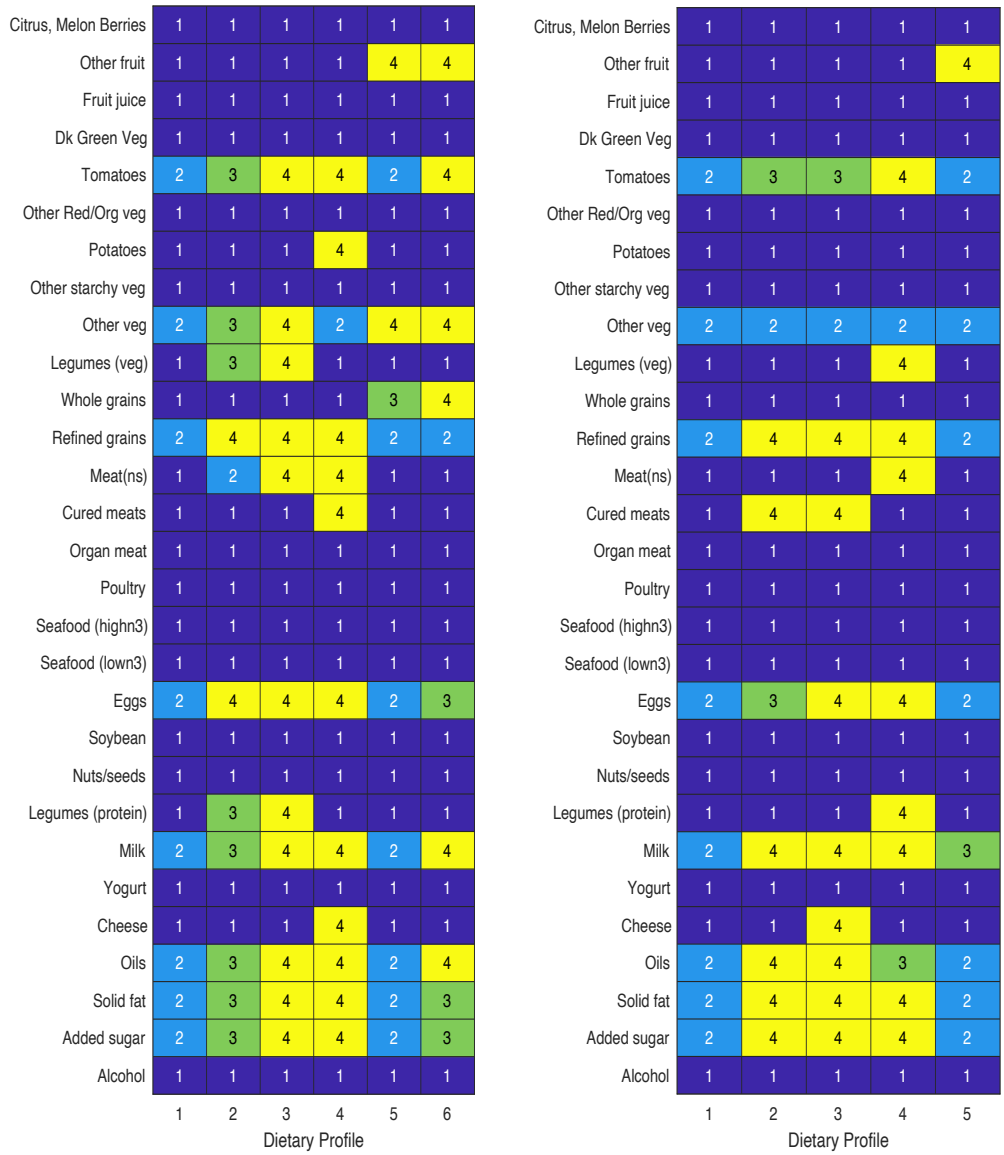

Fig. 3. Posterior mode of consumption pattern of dietary profiles. Numbers represent levels of consumption: 1= None, 2=Low, 3=Medium, 4=High

Table 2. Demographic distribution of Gen Adult Dietary Profiles

	Profile 1		Profile 2		Profile 3		Profile 4		Profile 5		Profile 6	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Overall	16.7	0.5	21.8	0.5	11.5	0.5	21.1	0.5	15.1	0.5	14.7	0.6
Poverty Level												
At or below 130%	34.4	1.6	17.9	1.2	28.5	1.8	25.8	1.4	19.7	1.1	14.2	1.1
Above 130%	65.6	1.6	82.1	1.2	71.5	1.8	74.2	1.4	80.3	1.1	85.8	1.1
Race/Ethnicity												
Mexican	7.2	1.0	9.8	1.2	19.7	2.3	8.3	0.9	7.3	0.9	5.0	0.6
Other Hispanic	5.6	0.8	7.5	0.8	12.2	1.4	4.3	0.5	5.5	0.6	4.2	0.5
Non-Hispanic White	60.5	2.1	64.5	1.9	50.1	3.0	66.9	2.1	64.8	1.9	73.0	1.7
Non-Hispanic Black	17.6	1.6	8.0	0.7	7.2	0.9	14.0	1.5	11.1	1.1	8.8	0.9
Non-Hispanic Asian	4.8	0.5	7.0	0.8	7.8	1.2	2.0	0.2	8.7	0.9	6.0	0.6
Mixed/Other	4.1	0.6	3.2	0.4	2.9	0.7	4.2	0.4	2.6	0.5	3.0	0.6
Gender												
Male	46.5	1.1	44.7	1.2	59.8	1.4	64.4	1.1	32.1	1.3	43.2	1.8
Female	53.5	1.1	55.3	1.2	40.1	1.4	35.5	1.1	67.9	1.3	56.8	1.8
Age Group												
20-34 years	34.7	1.5	27.4	1.2	29.1	1.5	36.3	1.5	20.9	1.4	21.9	1.5
35-49 years	25.4	1.5	28.2	1.2	28.9	1.6	26.3	1.3	22.6	1.4	20.7	1.2
50-64 years	24.5	1.2	26.8	1.2	26.1	1.7	23.7	1.1	30.9	1.5	30.7	2.0
65+ years	15.4	1.0	17.6	1.0	16.0	1.1	13.7	0.9	25.6	1.1	26.7	1.6
HEI 2015 Score	44.7	0.4	54.8	0.4	57.3	0.5	43.3	0.2	57.0	0.4	56.7	0.4
Framingham 10YR Risk	7.6	0.2	7.5	0.2	8.1	0.3	8.2	0.3	8.6	0.2	9.1	0.3
CVD Risk factor												
Hypertension	31.8	1.0	30.7	1.3	31.2	1.8	31.1	1.7	37.0	1.4	34.9	1.7
Hypercholesteremia	74.3	2.0	75.6	1.3	76.0	1.8	72.4	1.7	77.7	1.4	76.0	1.7
Obesity	41.3	1.5	36.3	1.9	35.9	2.3	45.0	1.5	37.6	1.7	35.4	2.0
Diabetes	9.9	1.1	8.5	0.8	9.0	0.8	10.0	0.9	10.2	0.9	8.2	0.9
Smoker	27.4	1.8	10.8	1.1	12.0	1.5	20.6	1.3	11.1	1.5	7.0	0.9

Table 3. Demographic distribution of Low-income Dietary Profiles

	Profile 1		Profile 2		Profile 3		Profile 4		Profile 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

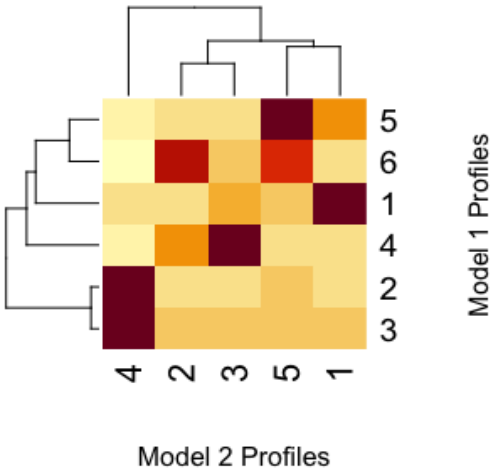

Fig. 4. Heatmap comparing the classification of dietary profile assignment for low-income adults in both models. Model 1 = overall adult population, Model 2 = low-income adult population.

majority of low-income participants assigned to profile 1 in model 1 remained in profile 1 in model 2. This is consistent with the similar consumption patterns we saw across the two models. Yet, a smaller proportion were still dispersed into profile 2,3, and 5 of model 2. Profile 4 in model 1 shared similarities with profile 3 in model 2, which is where a majority of the low-income adults indicated concordance, but shared representation in other profiles as well. Profiles 2 and 3 of model 2 were the most similar in consumption, but most participants assigned to profile 2 were found in M1P6, whereas profile 3 assigned participants were mostly in M1P4.

5. Discussion

The overfitted approach for survey data proposed in this paper, extends from the traditional latent class model, and adapts a current 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 of US adults collected in the National Health and Nutrition Examination Survey 2011-2018. Our model identified six dietary patterns reflective of the adult population, but only five dietary patterns reflective of the low-income adult population. The five patterns derived from the low-income adult population differed from those derived in the general adult population. This reflects the different sampled populations. Demographics can often influence dietary habits and behaviors. Yet, when trying to describe an entire population, dominating demographics can overpower the information gleaned from each cluster. For example, once adults not considered low-income were removed from the sample, we were able to identify additional food items that differed from the general adult population model because they did not share the same priority or modes of consumption. For example, the similarities of the first profile of model 1 with profile 5 of model 1 were similar but shared a higher modal consumption of other fruit and milk.

This method builds its strength on its generalizability and use in nationally representative dietary surveys, but also highlights the drawback of overgeneralization. With dominating demographics driving the distribution of patterns reflected in the population, if certain subpopulations are important to understand dietary differences or similarities from the dominating demographic, those subpopulations should be conducted in a separate analysis or a more advanced method should be implemented that is able to identify dietary patterns while also accounting for those population differences.

While the utility of this model has effectively demonstrated its use in dietary intake data, we must also acknowledge that the data is severely limited in its reliance on self-reporting. Several nutrition studies have found that prudent foods like vegetables and fruits are often overreported and less prudent foods like fats and oils are frequently underreported (Haraldsdóttir, 1993; Amanatidis et al., 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 et al., 2009; Briefel et al., 1997; Klesges et al., 1995; Hirvonen et al., 1997; Tooze et al., 2004).

While methods, such as doubly labeled water and biomarkers for select nutrients, are available to validate dietary assessment tools. These instruments were 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 should still be deemed relatively reliable(Tooze et al., 2004; Yuan et al., 2017). Another limitation of dietary recalls is the inability to capture day-to-day variation. The dietary patterns are based on one or two days of dietary records, which may or may not reflect the participant's 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, these assessments are often costly and therefore 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 could glean 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.

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.

References

(2020) Us department of agriculture: Choosemyplate. URL: `https://www.myplate.gov`.

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

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

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

Bowman, S., Clemens, J., Friday, J., Lynch, K., LaComb, R. and Moshfegh, A. (2017a) Food patterns equivalents intakes by americans: What we eat in america, nhanes 2003–2004 and 2013–2014. *Food Surveys Research Group.*

Bowman, S., Clemens, J., Friday, J., Lynch, K. and Moshfegh, A. (2017b) Food patterns equivalents database 2013–2014: methodology and user guide. *United States Department of Agriculture: Beltsville, MD, USA.*

Bowman, S., Clemens, J., Friday, J., Thoeric, R. and Moshfegh, A. (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, S., Clemens, J., Shimizu, M., Friday, J. and Moshfegh, A. (2018) Food patterns equivalents database 2015–2016: methodology and user guide. *US Department of Agriculture.*

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

Brown, A. F., Liang, L.-J., Vassar, S. D., Escarce, J. J., Merkin, S. S., Cheng, E., Richards, A., Seeman, T. and Longstreth Jr, W. (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**, 541–549.

Chen, T.-C., Clark, J., Riddles, M. K., Mohadjer, L. K. and Fakhouri, T. H. (2020) National health and nutrition examination survey, 2015- 2018: sample design and estimation procedures. *American journal of epidemiology*, **177**, 1279–1288.

Committee, D. G. A. et al. (2015) *Dietary guidelines for Americans 2015-2020.* Government Printing Office.

Daviglus, M. L., Talavera, G. A., Avilés-Santa, M. L., Allison, M., Cai, J., Criqui, M. H., Gellman, M., Giachello, A. L., Gouskova, N., Kaplan, R. C. et al. (2012) Prevalence of major cardiovascular risk factors and cardiovascular diseases among hispanic/latino individuals of diverse backgrounds in the united states. *Jama*, **308**, 1775–1784.

De Vito, R., Lee, Y. C. A., Parpinel, M., Serraino, D., Olshan, A. F., Zevallos, J. P., Levi, F., Zhang, Z. F., Morgenstern, H., Garavello, W. et al. (2019) Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology (Cambridge, Mass.)*, **30**, 93.

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

Fahlman, M. M., McCaughtry, N., Martin, J. and Shen, B. (2010) Racial and socioeconomic disparities in nutrition behaviors: targeted interventions needed. *Journal of nutrition education and behavior*, **42**, 10–16.

Gao, C., McDowell, I. C., Zhao, S., Brown, C. D. and Engelhardt, B. E. (2016) Context specific and differential gene co-expression networks via bayesian biclustering. *PLoS computational biology*, **12**, e1004791.

Gunawan, D., Panagiotelis, A., Griffiths, W. and Chotikapanich, D. (2020) Bayesian weighted inference from surveys. *Australian & New Zealand Journal of Statistics*, **62**, 71–94.

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

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

National Center for Health Statistics, D. o. t. N. H. and Surveys, N. E. (2018) National health and nutrition examination survey: Analytic guidelines, 2011-2014 and 2015-2016. *Tech. rep.*, Centers for Disease Control.

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**, 297–301.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010) *Bayesian nonparametrics*, vol. 28. Cambridge University Press.

Keshteli, A. H., Feizi, A., Esmaillzadeh, A., Zaribaf, F., Feinle-Bisset, C., Talley, N. J. and Adibi, P. (2015) Patterns of dietary behaviours identified by latent class analysis are associated with chronic uninvestigated dyspepsia. *British Journal of Nutrition*, **113**, 803–812.

Klesges, R. C., Eck, L. H. and Ray, J. 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**, 438.

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

Kunihama, T., Herring, A., Halpern, C. and Dunson, D. (2016) Nonparametric bayes modeling with sample survey weights. *Statistics & probability letters*, **113**, 41–48.

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

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

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

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A. and Chai, H. (2019) Statistical analysis of zero-inflated nonnegative continuous data: a review. *Statistical Science*, **34**, 253–279.

Marttinen, P., Pirinen, M., Sarin, A.-P., Gillberg, J., Kettunen, J., Surakka, I., Kangas, A. J., Soininen, P., O'Reilly, P., Kaakinen, M. et al. (2014) Assessing multivariate gene-metabolome associations with rare variants using bayesian reduced rank regression. *Bioinformatics*, **30**, 2026–2034.

Mattei, J., Sotres-Alvarez, D., Daviglus, M. L., Gallo, L. C., Gellman, M., Hu, F. B., Tucker, K. L., Willett, W. C., Siega-Riz, A. M., Van Horn, L. et al. (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**, 2035–2044.

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

Mokdad, A. H., Ballestros, K., Echko, M., Glenn, S., Olsen, H. E., Mullany, E., Lee, A., Khan, A. R., Ahmadi, A., Ferrari, A. J. et al. (2018) The state of us health, 1990-2016: burden of diseases, injuries, and risk factors among us states. *Jama*, **319**, 1444–1472.

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

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

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

Padmadas, S. S., Dias, J. G. and Willekens, F. 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**, 204–211.

Patterson, B. H., Dayton, C. M. and Graubard, B. I. (2002) Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association*, **97**, 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*, 277–281.

Poslusna, K., Ruprich, J., de Vries, J. H., Jakubikova, M. and van't Veer, P. (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**, S73–S85.

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

Reedy, J., Wirfält, E., Flood, A., Mitrou, P. N., Krebs-Smith, S. M., Kipnis, V., Midthune, D., Leitzmann, M., Hollenbeck, A., Schatzkin, A. et al. (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**, 479–487.

Roth, G. A., Johnson, C. O., Abate, K. H., Abd-Allah, F., Ahmed, M., Alam, K., Alam, T., Alvis-Guzman, N., Ansari, H., Ärnlöv, J. et al. (2018) The burden of cardiovascular diseases among us states, 1990-2016. *JAMA cardiology*, **3**, 375–389.

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

Sauvageot, N., Schritz, A., Leite, S., Alkerwi, A., Stranges, S., Zannad, F., Streel, S., Hoge, A., Donneau, A.-F., Albert, A. et al. (2017) Stability-based validation of dietary patterns obtained by cluster analysis. *Nutrition journal*, **16**, 1–13.

Savitsky, T. D. and Toth, D. (2016) Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, **10**, 1677–1708.

Schwedhelm, C., Iqbal, K., Knüppel, S., Schwingshackl, L. and Boeing, H. (2018) 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**, 227–235. URL: `https://doi.org/10.1093/ajcn/nqx027`.

Si, Y., Pillai, N. S. and Gelman, A. (2015) Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, **10**, 605–625.

Skinner, C. and Wakefield, J. (2017) Introduction to the design and analysis of complex survey data. *Statistical Science*, **32**, 165–175.

Sotres-Alvarez, D., Herring, A. H. and Siega-Riz, A. M. (2010) Latent class analysis is useful to classify pregnant women into dietary patterns. *The Journal of nutrition*, **140**, 2253–2259.

Sotres-Alvarez, D., Siega-Riz, A. M., Herring, A. H., Carmichael, S. L., Feldkamp, M. L., Hobbs, C. A., Olshan, A. F. and Study, N. B. D. P. (2013) Maternal dietary patterns are associated with risk of neural tube and congenital heart defects. *American journal of epidemiology*, **177**, 1279–1288.

Stephenson, B. J., Herring, A. H. and Olshan, A. (2020a) Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association*, **115**, 521–537.

Stephenson, B. J., Sotres-Alvarez, D., Siega-Riz, A.-M., Mossavar-Rahmani, Y., Daviglus, M. L., Van Horn, L., Herring, A. H. and Cai, J. (2020b) Empirically derived dietary patterns using robust profile clustering in the hispanic community health study/study of latinos. *The Journal of nutrition*, **150**, 2825–2834.

Tooze, J. A., Subar, A. F., Thompson, F. E., Troiano, R., Schatzkin, A. and Kipnis, V. (2004) Psychosocial predictors of energy underreporting in a large doubly labeled water study. *The American journal of clinical nutrition*, **79**, 795–804.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M. and Bates, N. (2014) *Hard-to-survey populations*. Cambridge University Press.

Van Havre, Z., White, N., Rousseau, J. and Mengersen, K. (2015) Overfitting bayesian mixture models with an unknown number of components. *PloS one*, **10**, e0131739.

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

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

Wang, D. D., Leung, C. W., Li, Y., Ding, E. L., Chiuve, S. E., Hu, F. B. and Willett, W. C. (2014) Trends in dietary quality among adults in the united states, 1999 through 2010. *JAMA internal medicine*, **174**, 1587–1595.

White, A. and Murphy, T. B. (2014) Bayeslca: An r package for bayesian latent class analysis. *Journal of Statiscal Software*, **61**, 1–28.

WirfÄLt, A. E. and Jeffery, R. 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**, 272–279.

Yuan, C., Spiegelman, D., Rimm, E. B., Rosner, B. A., Stampfer, M. J., Barnett, J. B., Chavarro, J. E., Subar, A. F., Sampson, L. K. and Willett, W. C. (2017) Validity of a dietary questionnaire assessed by comparison with multiple weighed dietary records or 24-hour recalls. *American journal of epidemiology*, **185**, 570–584.

Zhang, F. F., Liu, J., Rehm, C. D., Wilde, P., Mande, J. R. and Mozaffarian, D. (2018) Trends and disparities in diet quality among us adults by supplemental nutrition assistance program participation status. *JAMA network open*, **1**, e180237–e180237.

Zipf, G., Chiappa, M., Porter, K. S., Ostchega, Y., Lewis, B. G. and Dostal, J. (2013) Health and nutrition examination survey plan and operations, 1999-2010. *Tech. rep.*, National Center for Health Statistics.

6. Supplementary Materials

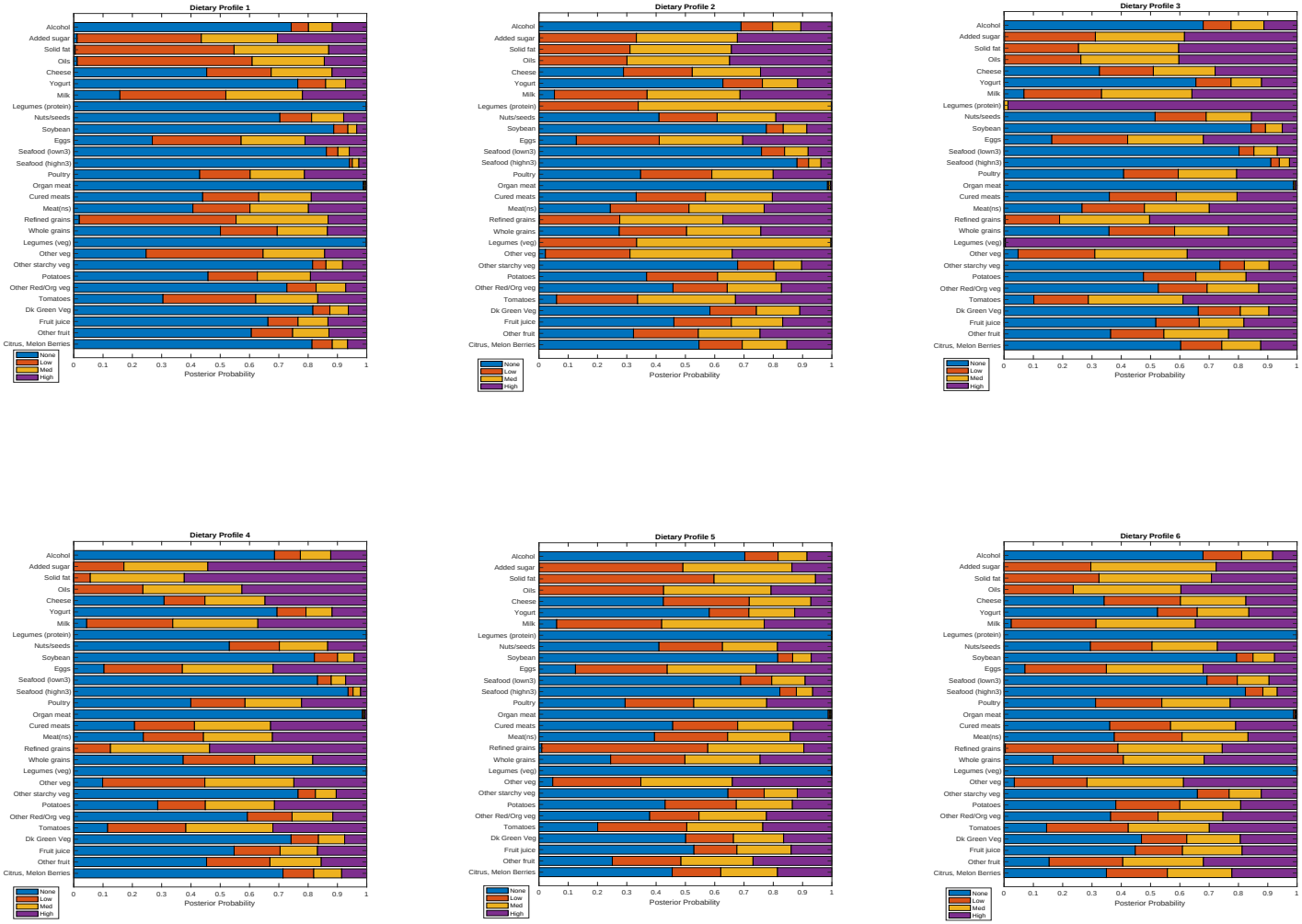

Fig. 5. Distribution of dietary consumption for each dietary profile derived from Model 1 (US adults)

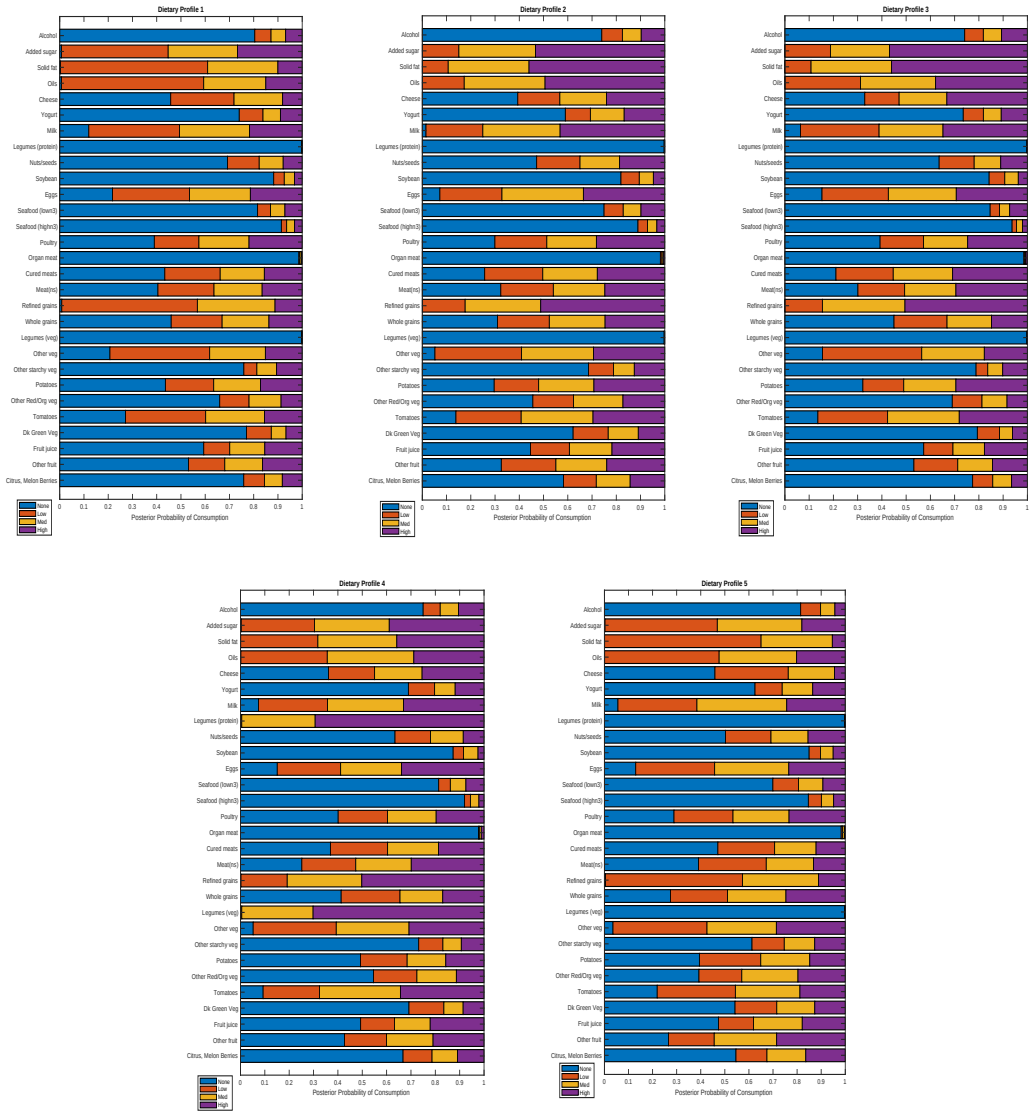

Fig. 6. Distribution of dietary consumption for each dietary profile derived from Model 2 (Low-income adults)