Clustering Methods to Empirically Derive Dietary Patterns in the United States

Briana J.K. Stephenson

Harvard T.H. Chan School of Public Health
Department of Biostatistics

August 3, 2020

Outline

Outline

Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

- Multi-center epidemiologic study on cardiometabolic health in Hispanic/Latino populations
- Participants aged 18-74 years of age
- Data collection: Baseline (2008-2011), Visit 2 (2015-2017)
- Field centers: Bronx, Chicago, Miami, San Diego
- Hispanic/Latino origin: Cuban, Puerto Rican, Dominican, Mexican, Central/South American

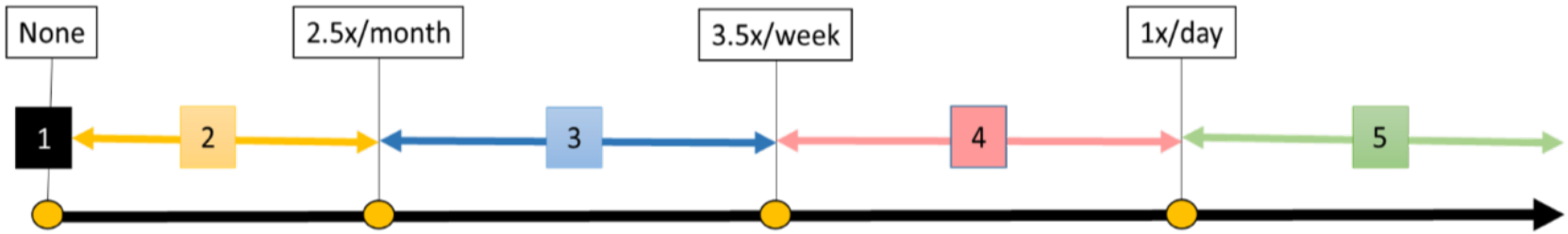
Why do we care about diet?

- Poor diet is the leading cause of morbidity and mortality in the United States
- Dietary intake is a modifiable exposure

Outline

HCHS/SOL Dietary Data: Food Propensity Questionnaire

- Queried consumption of 129 foods and beverages in the past year
- captures both episodically and frequently consumed foods
- 12 frequency consumption response levels collapsed to five categories (servings per day)



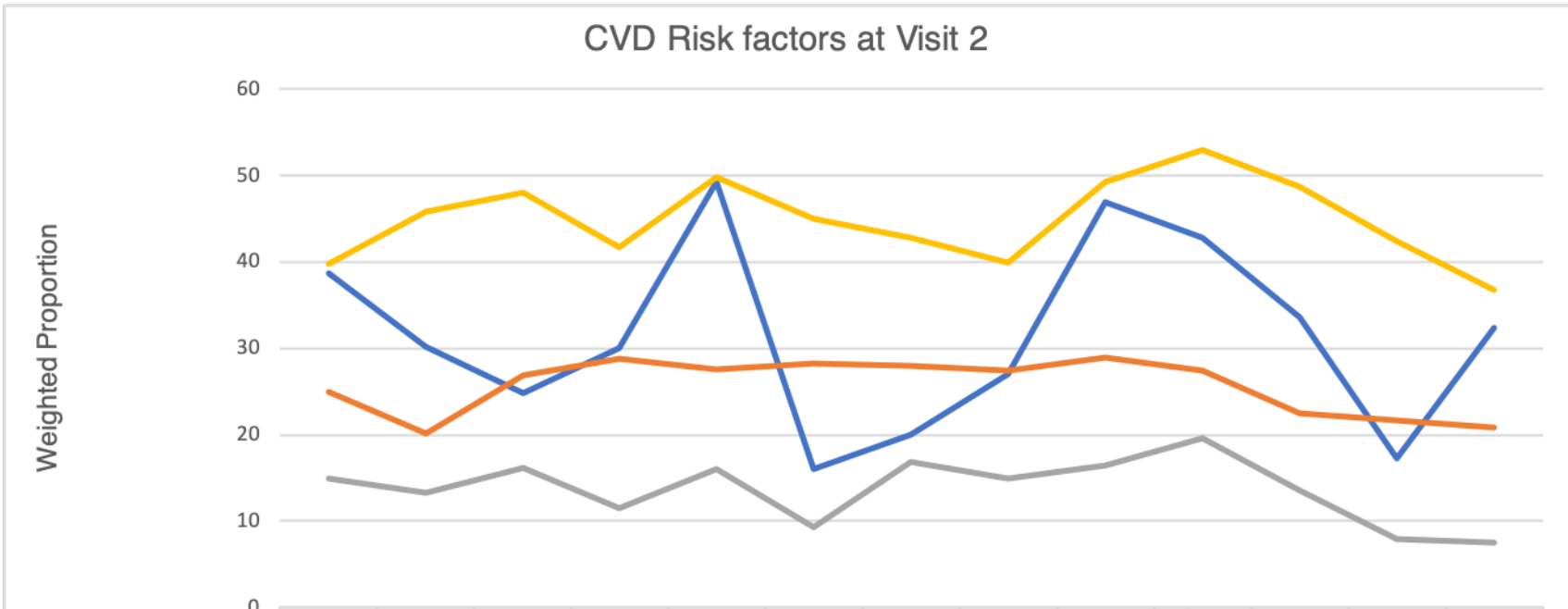
- Collapsed food variables
  - Binary: no consumption, any consumption (15)
  - Tertiary: no consumption, monthly, weekly/daily (48)
  - Quarternary: no consumption, monthly, weekly, daily (51)
  - non-collapsed (10)

HCHS/SOL Dietary Data: Dietary Recall

Two 24-hour dietary recalls

- quantify amount of daily consumption
- 16,000+ foods $\rightarrow$ 165 food variables $\rightarrow$ 52 food groups
- summarized as servings per day
  - no consumption
  - ≤ 0.5 servings/day
  - $0.5 \leq 1$ serving/day
  - $1 - 2$ servings/day
  - more than 2 servings/day

HCHS/SOL Motivation



	BX-Dom.	BX-C.Amer.	CHI-C.Amer.	MIA-C.Amer.	MIA-Cuban	BX-Mexican	CHI-Mexican	SD-Mexican	BX-Puerto Rican	CHI-Puerto Rican	BX-S.Amer.	CHI-S.Amer.	MIA-S.Amer.
Hypertension	38.75	30.16	24.8	29.97	49.26	16.07	20.07	27.08	46.88	42.74	33.64	17.32	32.37
Obese	24.92	20.09	26.91	28.81	27.51	28.29	27.99	27.43	28.99	27.38	22.49	21.69	20.82
Diabetes	14.92	13.3	16.2	11.46	15.98	9.36	16.9	14.9	16.38	19.66	13.53	7.93	7.52
Hypercholesterol	39.83	45.81	48.09	41.65	49.82	45.05	42.74	39.92	49.21	52.96	48.72	42.34	36.71

HCHS/SOL Study population

- 12,738 HCHS/SOL participants with completed FPQ
- 9 Subpopulations: field center and Hispanic/Latino background ($n_s \geq 300$)

	Bronx	Chicago	Miami	San Diego
Dominican	**1071**	18	54	2
Central American	164	**300**	**836**	42
Cuban	32	15	**1937**	6
Mexican	153	**1785**	28	**3002**
Puerto Rican	**1371**	**599**	65	23
South American	145	252	**402**	38
More than One	137	69	87	74

Robust Profile Clustering (RPC)

Breaks apart global clustering assumption from mixture model
Builds upon the local partition process (Dunson, 2009)

$$f(y_{i\cdot}|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j}^{p} f(y_{ij}|\theta_k)$$

GLOBAL
Variables shared with overall population

LOCAL
Variables shared with subpopulation

K_0: number of global clusters
K_s: number of local clusters
p: number of exposure variables

$$f(y_{i\cdot}|\theta_0, \theta_1) = \sum_{k=1}^{K_0} \pi_k \prod_{j, G_{ij}=1}^{p} f(y_{ij}|\theta_{0k}) \prod_{j, G_{ij}=0}^{p} \left[\sum_{l=1}^{K_s} \lambda_l f(y_{ij}|\theta_{1l}^{(s)}, s_i = s) \right]$$

RPC Components

RPC is processed through three probability models (Stephenson et al., 2020b)

1. Global cluster membership index, $C_i \in (1, \ldots, K)$

$$\Pr(C_i = h) = \pi_h \sim Mult(\alpha + \sum_{i=1}^{n} 1(C_i = 1), \ldots \alpha + \sum_{i=1}^{n} 1(C_i = K)$$

2. Local cluster membership index, $L_{ij} \in (1, \ldots, K)$

$$\Pr(L_{ij} = l | s_i = s) = \lambda_l^{(s)}$$
$$\lambda_l^{(s)} \sim Mult(\alpha + \sum_{i \in s,j} 1(L_{ij} = 1 | s_i = s), \ldots, \sum_{i,j} 1(L_{ij} = K | s_i = s))$$

3. Global allocation indicator, G_{ij}

$$\Pr(G_{ij} = 1 | s_i = s) = \nu_j^{(s)} \sim BetaBern(a_\beta, b_\beta, \beta^{(s)})$$

Simulation Study: Am I Global or local?

Figure 1: Case 1: Global

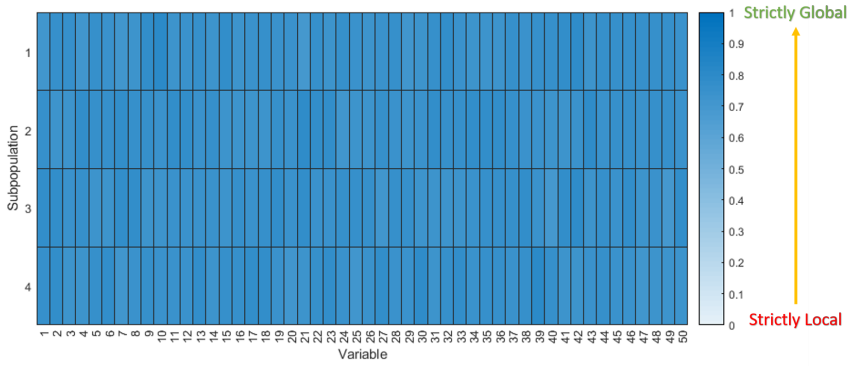


Figure 2: Case 2: Local

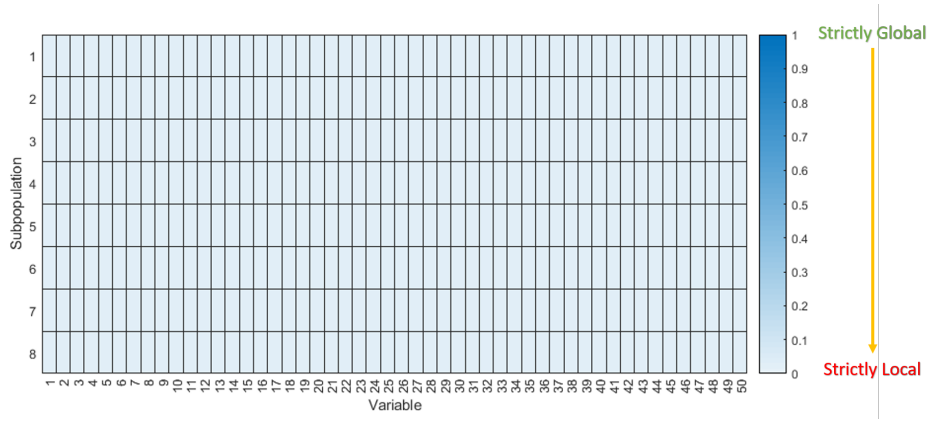
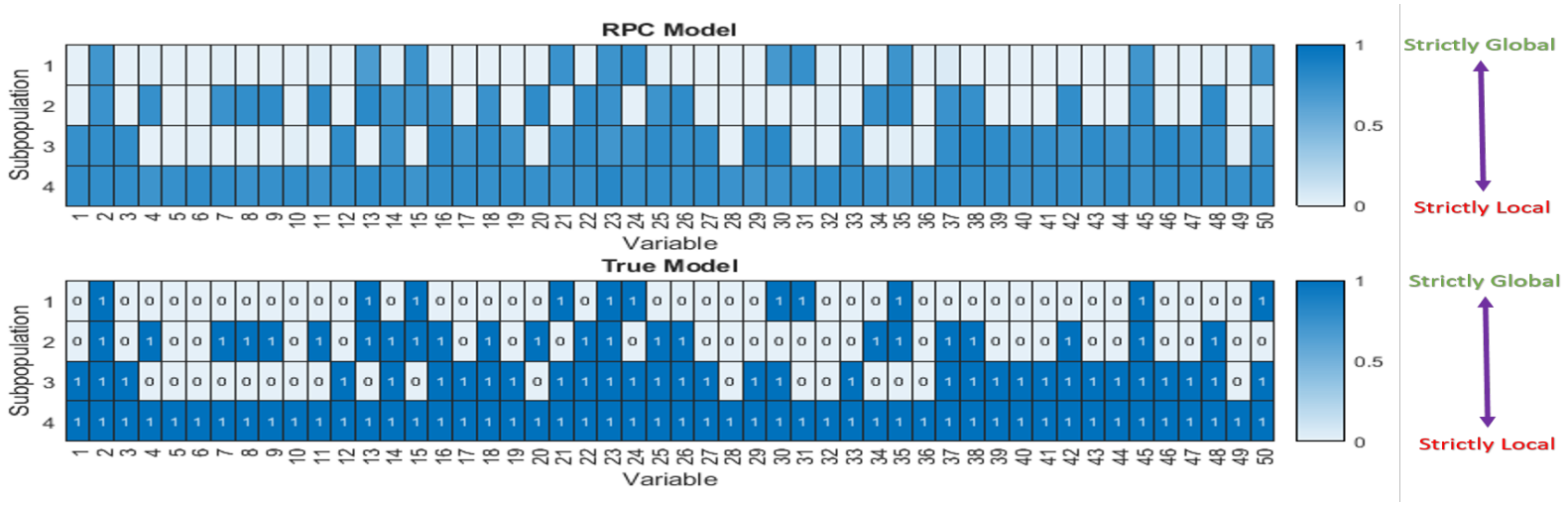


Figure 3: Case 3: Hybrid



Outline

RPC global profiles (Stephenson et al., 2020c)

GLOBAL PROFILE 1

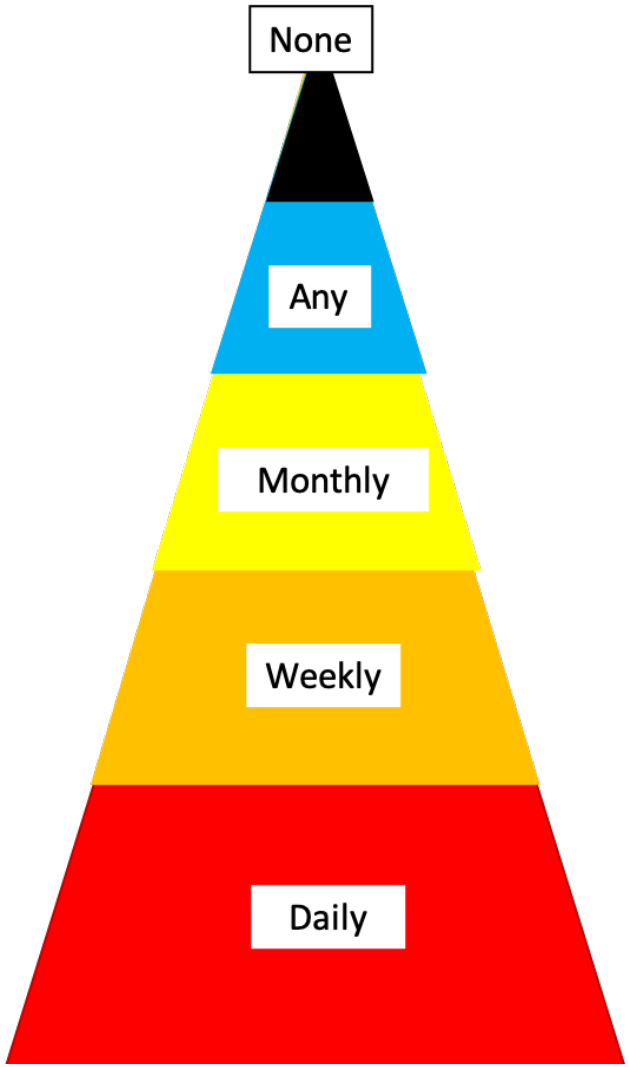
1. Fruit pie
2. Pizza (meat)
3. Sweet muffins
4. Fruit crisp
5. Pie (not fruit)

1. Artificial sweetener
2. Corn oil
3. Olive oil
4. Other oil
5. Canola/rapeseed oil

1. Eggs
2. Other soups
3. Cheese
4. Avocado
5. Foods w/ added oil

1. Nopal
2. Carrots
3. Broccoli
4. Other fish/seafood
5. Summer squash

1. Coffee
2. Cooked dried beans
3. Bananas
4. Fresh tomatoes
5. Rice: not whole grain



GLOBAL PROFILE 2

1. Fruit pie
2. Diet fruit drinks
3. Pie (not fruit)
4. Other oil
5. Meal replacement

1. Olive oil
2. Non-dairy creamer
3. Canola/rapeseed oil
4. Corn oil
5. Pizza (meat)

1. Cake
2. Ribs
3. Beef hamburger
4. Potato salad
5. Hot dogs

1. Fresh tomatoes
2. Bananas
3. Rice: not whole grain
4. Cheese
5. Eggs

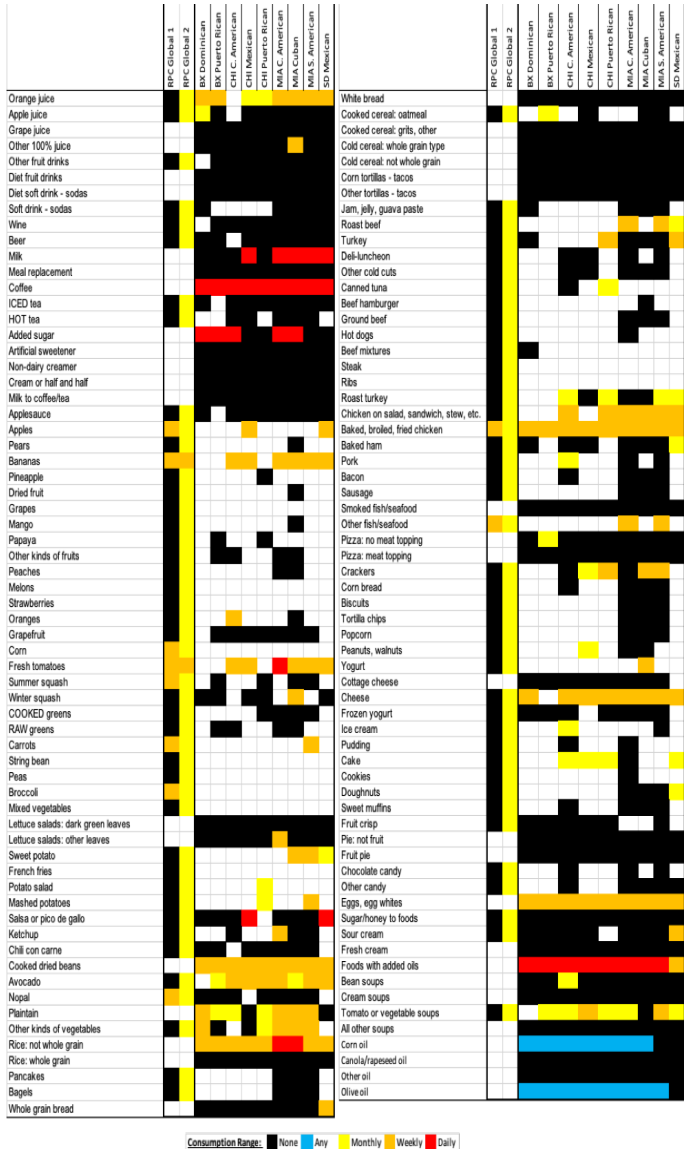
1. Coffee
2. Milk to coffee/tea
3. Added sugar
4. Foods with added oil
5. Milk.

RPC Global Profile: Frequency Distribution

		Global 1	Global 2
Bronx	Dominican	46%	**54%**
	Puerto Rican	42%	**58%**
Chicago	Central American	48%	**52%**
	Mexican	42%	**58%**
	Puerto Rican	34%	**66%**
Miami	Central American	**79%**	21%
	Cuban	**75%**	25%
	South American	**68%**	32%
San Diego	Mexican	26%	**74%**
Age	18-44 yrs	45%	**56%**
	45+ yrs	**55%**	45%
Gender	Female	48%	**52%**
	Male	50%	50%
Income	$< \$30,000$	**54%**	46%
	$\geq \$30,000$	37%	**63%**
	Not reported	**66%**	34%
Education	Less than HS	**55%**	45%
	High School/GED	47%	**53%**
	More than HS/GED	46%	**54%**

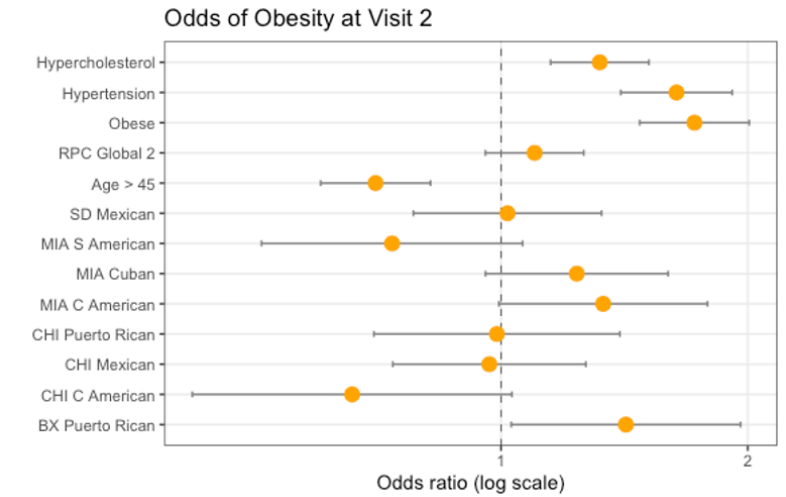
HCHS/SOL Results: Global/Local patterns ($\nu_j^{(s)} \geq 0.45$)

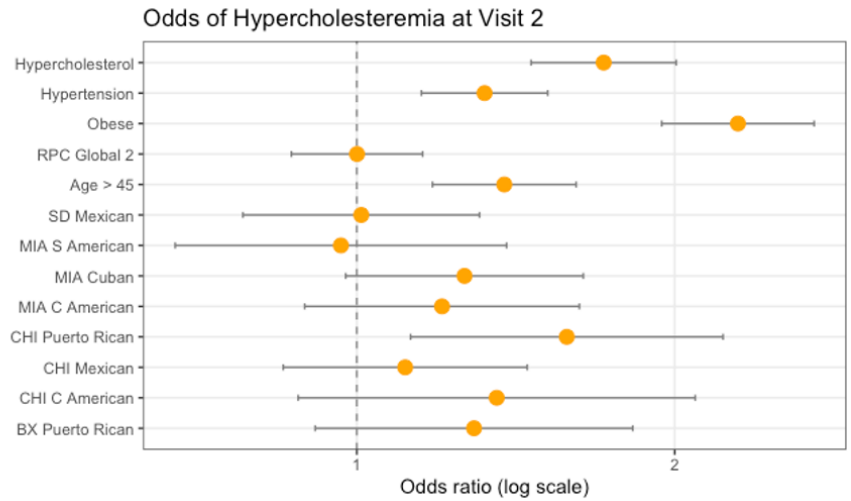
- All: 11 global, 41 local foods
- BX participants: most global variables (Dom=37%, PR=43%)
- MIA participants: least global variables (Cub=10%,
- Salsa: consumed daily - Mexican (CHI,SD)
- Nopal: only consumed by Mexican (CHI,SD)
- White Rice: consumed daily - MIA Cuban and C. American
- Milk: daily consumed - Mexican (CHI,SD), all MIA
- Oils: SD-Mex not consumed; less with foods

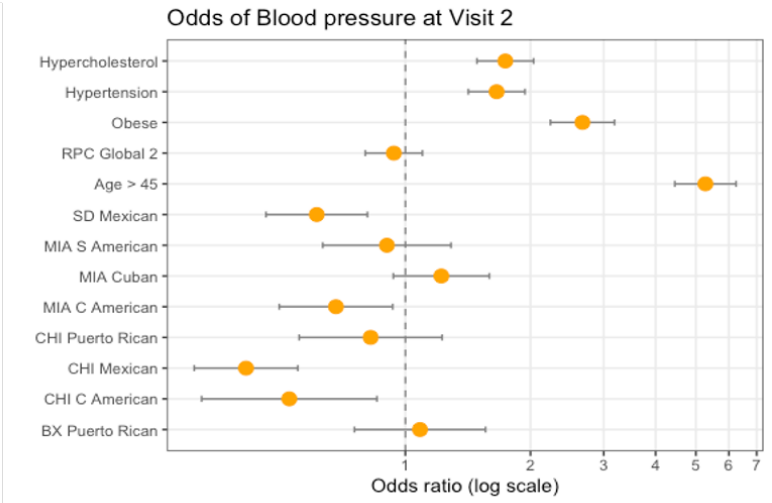


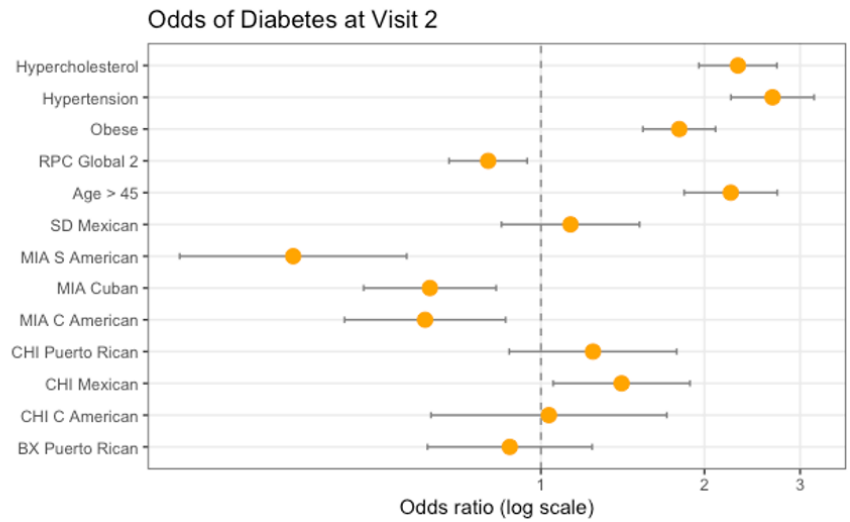
RPC CVD Risk factors at Visit 2

Additional Confounders: US nativity, gender, education, income, depression, alcohol, cigarette use
Referent group: Global Profile 2; SD-Mexican background participants





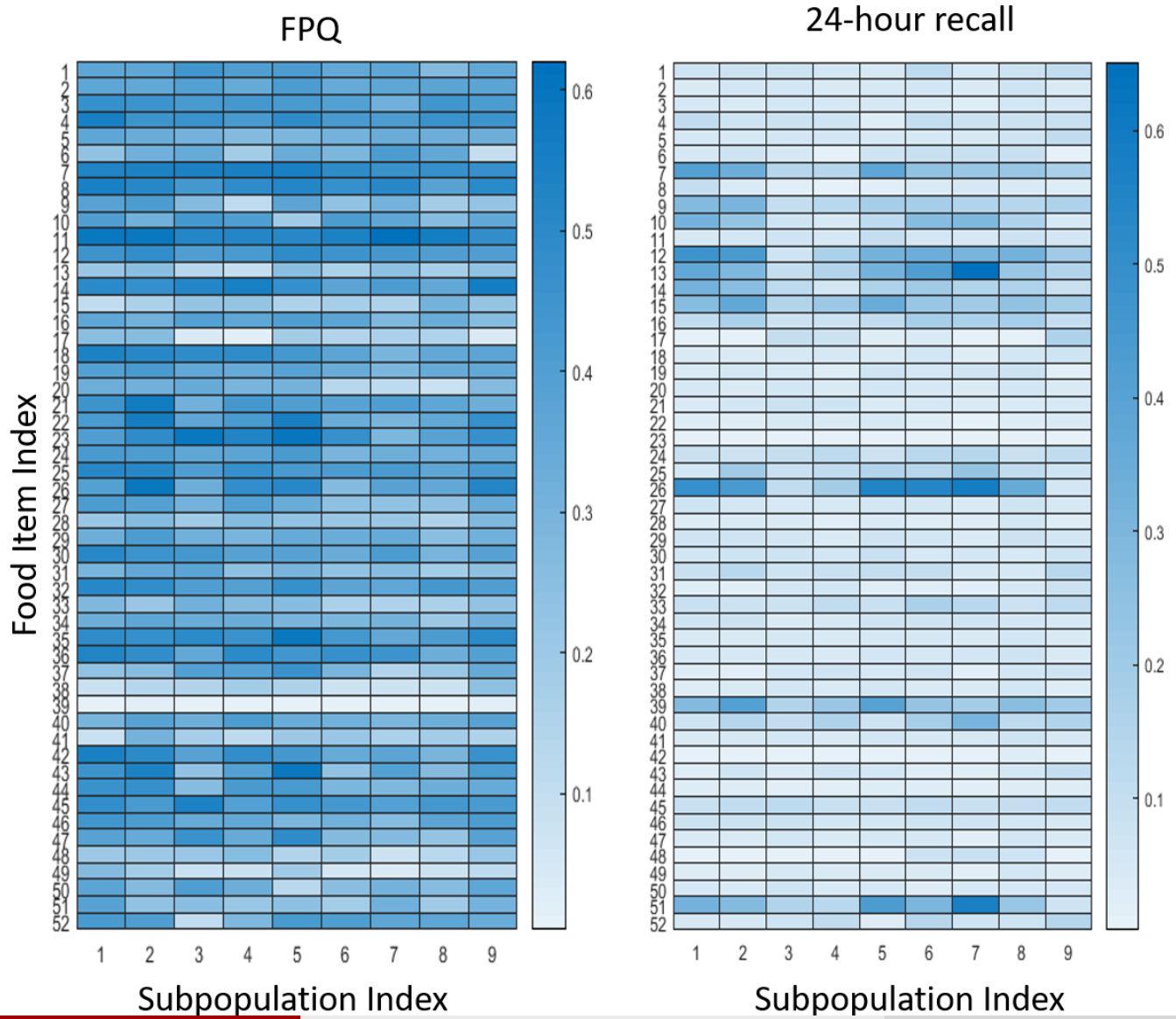




Outline

Application to 24HR recall: 52 food groups

Consumption sparsity drives patterns local



Application to 24HR Recall

Challenges incorporating local profile information into regression model.

- variable selection
- fixed vs random effects

$$\xi_1 + \sum_{s=2}^{S} \mathbf{1}(s_i = s)\xi_s + \sum_{k=2}^{K_0} \mathbf{1}(C_i = k)\xi_{S+k} + W_i\xi_{dem} + \sum_{j,s} +1(L_{ij} = l | s_i = s)b_{js}?$$

Different Approach

- Abandon a priori food clustering
- Consumption-focused clustering
- Bottom-up approach: cluster from local profiles
  - SNOB: subset nonparametric Bayesian clustering Zuanetti et al. (2019)
  - Scalable Bayesian nonparametric Clustering (Ni et al., 2020)

Outline

Where do we go from here?

What we have

- Reduced model testing
- Flexible and robust interpretability
- Identify global and local patterns jointly
- User-control on rate of cluster formation

What we still need

- dietary intake misreporting
- incorporation of local profile information in regression
- post-processing to remove redundancies
- probabilistic variable partitioning

Further Directions

- Nutrient intake: Bayesian Multi-study factor analysis (in progress)
- Supervised Robust Profile Clustering *(under review)*

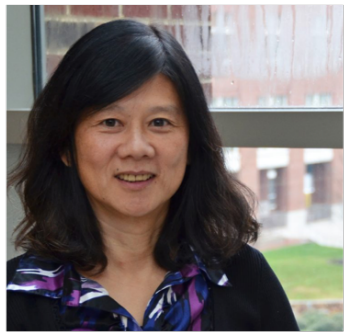
$$\mathcal{L}(x_i, y_i|-) = \left[\sum_{h=1}^{K_0} \pi_h \prod_{j:G_{ij}=1}^{p} \prod_{r=1}^{d_j} \theta_{0jh,r}^{\mathbf{1}(x_{ij}=r)}\right] \Phi(W_i\xi)^{y_i}[1-\Phi(W_i\xi)]^{1-y_i} \left[\sum_{l=1}^{K_s} \lambda_l^{(s_i)} \prod_{r=1}^{d_j} (\theta_{1jl,r}^{(s_i)})^{\mathbf{1}(x_{ij}=r)}\right]$$

- Inclusion of small-sized subpopulations
- Temporal clustering under repeated measures study design
- Bayesian model-based clustering under complex survey design

Acknowledgements

Amy Herring

Daniela Sotres-Alvarez



Jianwen Cai

Anna-Maria Siega-Riz

Yasmin Mossavar-Rahmani

Linda Van Horn

Martha Daviglus

NHLBI: HHSN268201300001

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts HHSN268201300001I / N01-HC-65233 to the University of North Carolina , HHSN268201300004I / N01-HC-65234 to the University of Miami, HHSN268201300002I / N01-HC-65235 to the Albert Einstein College of Medicine, HHSN268201300003I / N01-HC-65236 to the University of Illinois at Chicago (Northwestern University), and HHSN268201300005I / N01-HC- 65237 to San Diego State University. The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute onMinority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, and the NIH Institution-Office of Dietary Supplements.

Any Questions?

Relevant Publications

Daviglus, M. L., Pirzada, A., and Talavera, G. A. (2014). Cardiovascular disease risk factors in the hispanic/latino population: lessons from the hispanic community health study/study of latinos (hchs/sol). *Progress in cardiovascular diseases*, 57(3):230–236.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75(1):337–346.

Dunson, D. B. (2009). Nonparametric bayes local partition models for random effects. *Biometrika*, 96(2):249–262.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of epidemiology*, 20(8):642–649.

Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y., and Ji, Y. (2020). Scalable bayesian nonparametric clustering and classification. *Journal of Computational and Graphical Statistics*, 29(1):53–65.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, 20(8):629–641.

Stephenson, B., Herring, A., and Olshan, A. (2020a). Supervised robust profile clustering. *arXiv preprint arXiv:2007.04509*.

Stephenson, B. J., Herring, A. H., and Olshan, A. (2020b). Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association*, 115(530):521–537.

Stephenson, B. J. K., Sotres-Alvarez, D., Siega-Riz, A.-M., Mossavar-Rahmani, Y., Daviglus, M. L., Van Horn, L., Herring, A. H., and Cai, J. (2020c). Empirically Derived Dietary Patterns Using Robust Profile Clustering in the Hispanic Community Health Study/Study of Latinos. *The Journal of Nutrition*. nxaa208.

Zuanetti, D. A., Müller, P., Zhu, Y., Yang, S., and Ji, Y. (2019). Bayesian nonparametric clustering for large data sets. *Statistics and Computing*, 29(2):203–215.