K-Means Clustering for Longitudinal Chemical Mixtures Analysis in LIFECODES Dataset

Hui Sui
Faculty advisor: Rachel Carroll
UNC-Chapel Hill Department of Statistics

Introduction

Phthalate and paraben exposure in pregnant women has previously been linked to adverse health outcomes such as **preterm birth**. Unsupervised clustering methods may serve as useful tools for identifying individuals with shared real-life patterns of chemical exposures. Knowledge of these groupings and their risk of adverse outcomes has the potential to inform targeted public health prevention strategies.

This research applies **k-means clustering** to **identify clusters** of pregnant women with **shared exposure profiles** as defined by levels of the urinary phthalate and parabens using data from a case-control study within the LIFECODES birth cohort.

LIFECODES Dataset

- A birth cohort of women who were planning to deliver at a Boston area hospital between 2006 and 2008
- Four study visits at median 10, 18, 26, and 35 weeks of gestation (only use the first three visits)
- At each study visit, participants provide urine and blood samples
- 427 variables including personal information (age, height, BMI, education, race, etc.) of each subject and chemical measures from each visit
- Sample size is 482 (352 controls and 130 cases)
- Used only the **337 controls** who have completed age, BMI, race and education and the fifteen chemicals of interest

Conclusion

- K-means successfully cluster the 337 pregnant women into **4 clusters** based on their chemical exposures
- **Unevenly** distributed clusters
- **High consistency** in cluster memberships over time
- Moderate high group has more whites, more college graduates, older aged, and lower BMI subjects
- Cluster memberships **not very useful** in predicting oxidative stress biomarkers

Acknowledgement

This research was funded by NSF's grant DMS-1659288 to UNC Wilmington. Special thanks to Dr. Rachel Carroll for allowing us to use and extend her research, and to Dr. Cuixian Chen, Dr. Yishi Wang for helping support us at the UNCW

Reference

- Carroll, White, K. M. M. Z. and Ferguson (2019). Latent classes for meaningful chemical mixtures analyses in epidemiology: An example using phthalate and phenol exposure biomarkers in pregnant women.
- Ferguson, K. K., McElrath, T. F., and Meeker, J. D. (2014). Environmental phthalate exposure and preterm birth. JAMA pediatrics, 168(1):61–67.
- Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C., et al. (2015). kml and kml3d: R packages to cluster longitudinal data. Journal of Statistical Software, 65(4):1–34.
- Genolini, C. and Falissard, B. (2010). Kml: k-means for longitudinal data. Computational Statistics, 25(2):317–328.

Methodology

Algorithm

- An algorithm of the **expectation-maximization** (EM) class
- Initially, each observation is assigned to a cluster
- Then, the optimal clustering is reached by alternating two phases:
  1)expectation phase: centers of the different clusters are computed
  2)maximization phase: assigns each observation to its nearest cluster
- The alternation of the two phases is repeated until no further changes occur in the clusters

Choosing the K

- Uses **Calinski-Harabasz** criterion
- Also consider **interpretability**
- Chose **4 clusters**
- Comparability with previous LCA research

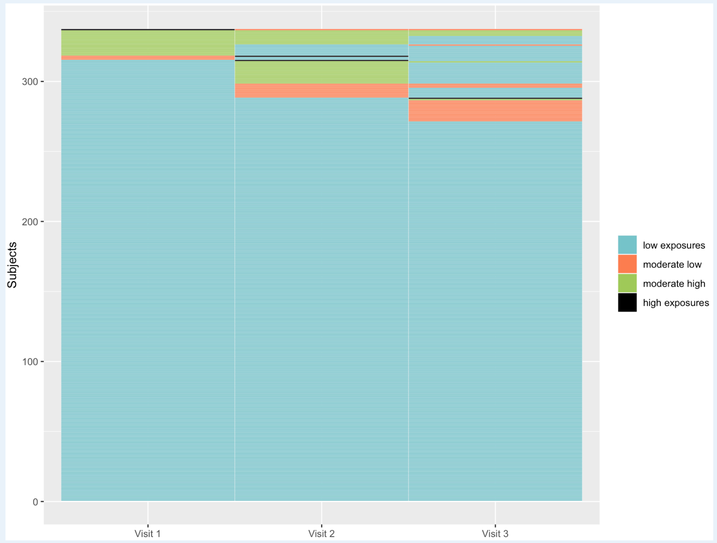
Approaches

- **Longitudinal k-means** with R package kml3d
  - repeatedly measured variables are seen as trajectories
  - allows to examine the evolution
- **Regular k-means** at each time point with R package stats

Regression Models

- Examine the **association** between cluster membership and urinary oxidative stress biomarkers
- With or without cluster membership as a variable

Figure 1: a heat map of cluster memberships over time



Results

Longitudinal k-means

Table 1: Number of subjects in each cluster

Low	Moderate low	Moderate high	High	Total
333	2	1	1	337

Table 2: Regression models results

		AIC	Res. Deviance	P-value
OHdG	w/out cluster	185.17	28.27	**0.01**
	w/ cluster	183.08	27.34	
ISO	w/out cluster	614.53	166.68	0.35
	w/ cluster	617.25	164.43	

Independent k-means

Table 3: Number of subjects in each cluster

	Low	Moderate low	Moderate high	High	Total
Visit 1	315	3	18	1	337
Visit 2	299	14	22	2	337
Visit 3	310	20	6	1	337

Table 4: Regression models results

OHdG		AIC	Res. Dev.	p-value
Visit 1	w/out cluster	582.23	106.35	0.98
	w/ cluster	588.07	106.30	
Visit 2	w/out cluster	341.48	51.73	0.51
	w/ cluster	345.17	51.32	
Visit 3	w/out cluster	387.74	61.13	**0.01**
	w/ cluster	382.60	58.78	

ISO		AIC	Res. Dev.	p-value
Visit 1	w/out cluster	897.70	275.05	0.74
	w/ cluster	902.47	274.03	
Visit 2	w/out cluster	850.21	291.88	0.20
	w/ cluster	852.21	287.27	
Visit 3	w/out cluster	783.99	246.72	0.69
	w/ cluster	788.52	245.45	

Table 5: Demographic characteristics within clusters of visit 1

Visit 1	Covariate	Overall	Low	Moderate low	Moderate high	High
Race	White	202(59.9%)	187(59.4%)	1(33.3%)	14(77.8%)	0(0.0%)
	Black	52(15.4%)	48(15.2%)	2(66.7%)	2(11.1%)	0(0.0%)
	Other	83(24.6%)	80(25.4%)	0(0.0%)	2(11.1%)	1(100.0%)
Education	<= HS	47(13.9%)	43(13.7%)	2(66.7%)	1(5.6%)	1(100.0%)
	Technical school	52(15.4%)	51(16.2%)	0(0.0%)	1(5.6%)	0(0.0%)
	Some college	99(29.4%)	93(29.5%)	1(33.3%)	5(27.8%)	0(0.0%)
	>= college	139(41.2%)	128(40.6%)	0(0.0%)	11(61.1%)	0(0.0%)
Age	<25	41(12.2%)	39(12.4%)	1(33.3%)	0(0.0%)	1(100.0%)
	25-29	68(20.2%)	63(20.0%)	1(33.3%)	5(26.3%)	0(0.0%)
	30-34	133(39.5%)	126(40.0%)	1(33.3%)	6(31.6%)	0(0.0%)
	35+	95(28.2%)	87(27.6%)	0(0.0%)	8(42.1%)	0(0.0%)
BMI	<25	184 (54.6%)	177(56.2%)	1(33.3%)	15(78.9%)	0(0.0%)
	25-30	92 (27.3%)	86(27.3%)	1(33.3%)	2(11.1%)	1(100.0%)
	>30	61 (18.1%)	52(16.5%)	1(33.3%)	1(5.6%)	0(0.0%)