Finite mixture clustering of risk behaviors for an infectious disease

Joseph Kang, Ph.D.

Mathematical Statistician

The U.S. Census Bureau

General aims

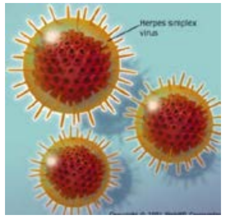
- **Overarching aim:**

  Estimation of Herpes (an infectious disease) prevalence among latent classes of sexual partners using NHANES complex survey data

- **Statistical methods:**

1) Latent class analysis of partners' gender & frequencies
2) Expected estimating equation (EEE) approach for missing latent class
3) Propensity weights for comparing classes

Herpes is an important infection



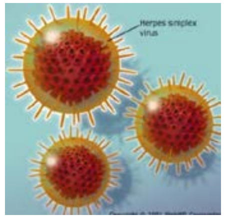
- **Genital herpes (more commonly known as “herpes”)**

Genital herpes is common in the United States. More than one out of every six people aged 14 to 49 years have genital herpes. (https://www.cdc.gov/std/herpes/stdfact-herpes.htm)

- **Herpes simplex viruses (HSV)**

1) HSVs are categorized into two types: herpes type 1 (HSV-1, or oral herpes) and herpes type 2 (HSV-2, or genital herpes)
2) In HSV-2, the infected person may have sores around the genitals or rectum
3) Most of the time, HSVs cause no symptoms, but some infected people have "outbreaks" of blisters and ulcers

Health problems with HSV-2



- **There is no cure for herpes**
  Once infected, people remain infected for life. However, there are medications that can prevent or shorten outbreaks (https://www.cdc.gov/std/herpes/)

- **HSV-2 is related to psychological issues**
  Feelings of shame, embarrassment, anxiety, or depression are the most common psychological issues related to HSV-2 (Merin et al, 2011)

- **Herpes is related to HIV**
  Having genital herpes can increase the risk of being infected with HIV, the virus that causes AIDS (https://www.nih.gov/)

HSV-2 is associated with the number of partners

- The risk of having HSV-2 increases with respect to the number of partners

- CDC researchers epidemiologically defined six categories for the number of partners: 0, 1, 2-4, 5-9, 10-49, 50+ (Xu et al 2006, JAMA)

- **An issue with the complex patterns of combinations:** all possible combinations are $6^4 = 1296$!

Past year		Life time	
Male Partner	Female Partner	Male Partner	Female Partner
0	0	0	0
0	0	0	1
0	0	1	1
…	…	…	…
50+	50+	50+	50+

Statistical challenges

- **Partners' gender & frequencies are high-dimensional**

  Latent class analysis (LCA) can identify commonly occurring behavioral clusters

- **Missing latent class variable in NHANES complex survey data**

  Estimating equations can accommodate survey design features

NAHNES data sets

- Data

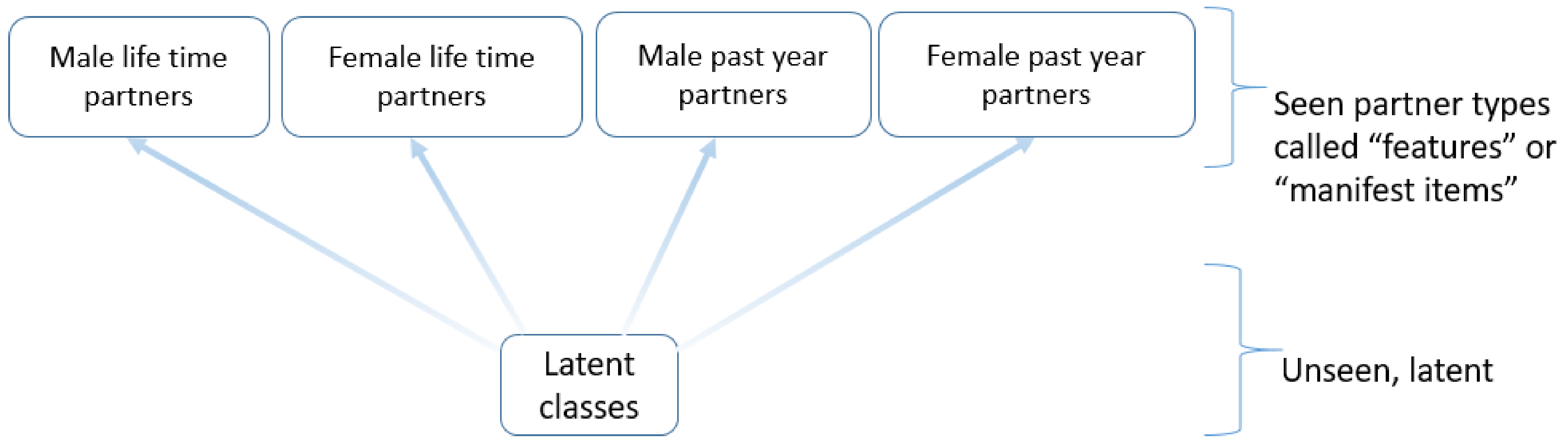
Our analysis sample was from the National Health and Nutrition Examination Surveys (NHANES) from 2001–2014; N=2,204

- Main results foreseen with LCA

1) Two latent classes were found: class1 (9.8%) vs. class 2 (91.1%)
2) The HSV-2 rate was significantly higher in class 1 than class 2 (20.6% vs. 13%, P-value=0.02)
3) What is LCA?

LCA: an unsupervised clustering method (machine learning) and a finite mixture model (statistics)

The name "latent" indicates that there are unseen clusters that exist to explain manifested values



The LCA model is quite simple… to some

- Notation
  U: partner variables (features, manifest items); Z: latent class membership;

- LCA as a mixture of C probability models:
  $$P(U = u) = P(U = u|Z = 1)P(Z = 1)+\ldots+P(U = u|Z = C)P(Z = C)$$

- Consider $P(U = u|Z = 1)$:
  $$\begin{aligned} & P(U_1 = u_1, \ldots, U_4 = u_4|Z = 1) \\ = & P(U_1 = u_1|Z = 1) \times \cdots \times P(U_4 = u_4|Z = 1) \times P(Z = 1) \\ = & \, constant \times \cdots \times constant \times constant \end{aligned}$$

- The constant parameters are estimated with an EM-type algorithm

- The log-likelihood is weighted with survey weights (Patterson et al., JASA, 2002)

A typical LCA algorithm

- The goal is to maximize a weighted log-likelihood

$$wgt \times logP(U) = wgt \times \sum_c I(Z=c) \times log\{P(U|Z=c)P(Z=c)\}$$

➢ E-step

Weighted log-likelihood is expected with the conditional probability of Z given U:

$$\delta_c = P(Z=c|U) = \frac{P(U|Z=c)P(Z=c)}{\sum_{c'} P(U|Z=c')P(Z=c')}$$

➢ M-step

Solve the equation below for $\rho = P(U_m = a|L=c)$

$$\sum wgt \times \delta_c \times (I(U_m = a) - \rho) = 0$$

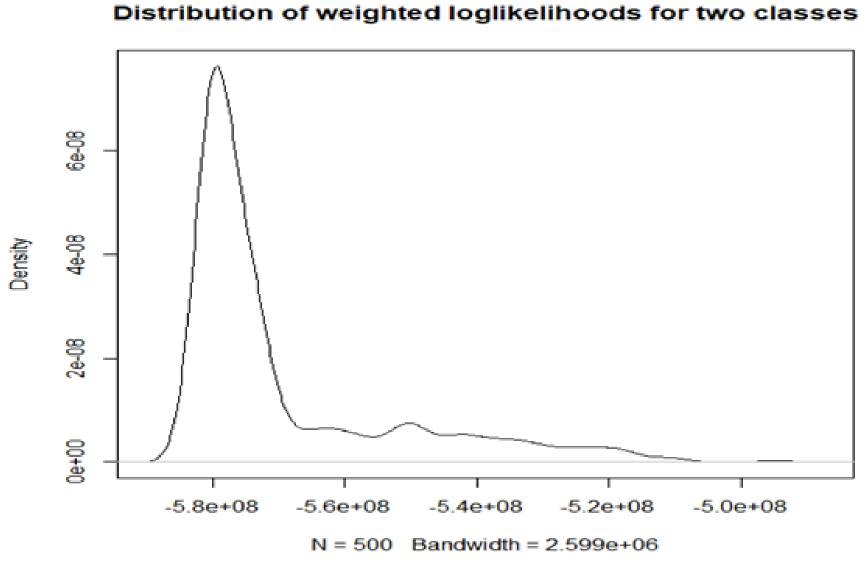
$$\sum wgt \times expected\ class \times (differnce\ of\ data\ with\ parameter\ ) = 0$$

LCA fitting

1) AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria), d_AIC (design-based AIC) all supported the two class solution

Class	AIC($\times 10^5$)	BIC($\times 10^5$)	d_AIC($\times 10^{12}$)
2	**1.52**	**1.52**	**3.01**
3	2.13	2.14	5.74
4	3.21	3.21	8.46
5	4.44	4.44	11.57
6	6.15	6.15	18.63

2) 500 random starting values were used to evaluate the distribution of weighted maximum likelihood estimates (the global estimate from weighted log-likelihoods was used)



Two latent classes

Class 1 (8.9%)						
Partners	0	1	2-4	5-9	10-49	50<=
M-YR	9.6	40.4	32.1	7	7.9	2.9
F-YR	78.7	8.1	12.4	0.6	0.2	0
M-LT	0	4.3	20.7	16.7	42.5	15.9
F-LT	35.2	14	24.9	7.8	15.9	2.3

Class 2 (91.1%)						
Partners	0	1	2-4	5-9	10-49	50<=
M-YR	100	0	0	0	0	0
F-YR	7	72.9	15.2	3.4	1.3	0.2
M-LT	94.1	3.9	1.6	0.3	0.1	0
F-LT	0	11.8	21.6	24.3	35.6	6.7

❖ Glossary

- M-YR: Male past year partners
- F-YR : Female past year partners
- M-LT : Male lifetime partners
- F-LT : Female lifetime partners

❖ Foreseen results

- HSV-2: Class 1's 20.6% vs Class 2's 13% (15.4%, adjusted)
- Next slides will explain class characteristics

Class 1 characteristics

Class 1 (8.9%)						
Partners	0	1	2-4	5-9	10-49	50<=
M-YR	9.6	40.4	32.1	7	7.9	2.9
F-YR	78.7	8.1	12.4	0.6	0.2	0
M-LT	0	4.3	20.7	16.7	42.5	15.9
F-LT	35.2	14	24.9	7.8	15.9	2.3

- Class 1: Mostly **male** partners

➢ 1≤M-YR (90.4%=40.4%+…+2.9%)

➢ 1≤M-LT (100%=4.3%+…+15.9%)

➢ M-YR>F-YR for 1≤partners (90.4% vs. 21.3%=8.1%+…+0.2%)

➢ M-LT>F-LT for 5≤partners (75.1% vs. 26%)

Class 2 characteristics

Class 2 (91.1%)						
Partners	0	1	2-4	5-9	10-49	50<=
M-YR	100	0	0	0	0	0
F-YR	7	72.9	15.2	3.4	1.3	0.2
M-LT	94.1	3.9	1.6	0.3	0.1	0
F-LT	0	11.8	21.6	24.3	35.6	6.7

- Class 2: Mostly **female** partners for both the previous year and lifetime

➢ Single F-YR (72.9%)

➢ Multiple F-LT (89.2% =21.6%+…+6.7%)

Comparison of two classes with propensity weights

Variable	Class 1 (%)	Class 2 (%)	StdDiff
Age group:<=24	13.7±2.6	13.7±1.8	0
Age group:25-29	15.6±1.9	15.4±1.7	0.005
Age group:30-39	42.7±4.4	33.2±1.6	0.196
Age group:40-49	28±3.5	37.6±1.8	0.206
Race:Black	9.6±1.2	10.2±1.5	0.022
Race:Mex	7.3±1.5	9.8±1.1	0.092
Race:Other	6±1.5	4.4±1.2	0.07
Race:OtherHis	8.2±0.9	5.4±1.7	0.113
Race:White	69±2.9	70.2±2.7	0.026
Poverty: Yes	12.6±3.5	11.3±1.4	0.041
Education years<12	29.4±3.9	44±2.3	0.308
Marriage: Married	7.6±2.7	57.4±2.1	1.255
Marriage: Unmarried	64.8±4.1	23.6±2.1	0.913
Marriage: Partner	19.5±3.1	8.9±1.1	0.309
Marriage: Separate	8.1±2.1	10.1±1.8	0.073
Cocain usage: No	68.1±5.7	70.5±2	0.053
Age at first sex<18	26.4±6.1	19.4±1.3	0.168
Circumcised: No	13.7±1.7	18.8±2.2	0.14

Class 1	Class 2 (propensity-weighted)	StdDiff
13.7±2.6	14.0±2.9	0.008
15.6±1.9	16.1±1.9	0.012
42.7±4.4	42.4±4.9	0.005
28±3.5	27.6±4	0.01
9.6±1.2	10.3±1.5	0.022
7.3±1.5	6.6±1.4	0.025
6±1.5	5.2±1.2	0.031
8.2±0.9	8.9±1.3	0.023
69±2.9	69±3.1	0
12.6±3.5	13.7±4.7	0.034
29.4±3.9	29.7±3.7	0.008
7.6±2.7	7.7±2.7	0.002
64.8±4.1	64.6±4.4	0.003
19.5±3.1	19.6±3.2	0.002
8.1±2.1	8.1±2	0.001
68.1±5.7	67.6±5.6	0.011
26.4±6.1	26.4±6.5	0.001
13.7±1.7	13.8±1.6	0.003

StdDiff (Austin, 2009, SIM):

- Standardized Difference

$$\frac{P_1 - P_2}{\sqrt{\frac{P_1(1-P_1) + P_2(1-P_2)}{2}}}$$

- An StdDiff of 0.1 denotes meaningful imbalance
- Highlighted are >=0.1

Latent classes were un-confounded, balanced!!

Next slides explain about observational studies

What is an observational study?

- A goal of an observational study: to attain balance among the comparison groups

- Consider a simple example below:

	Treatment	Control
Male	40%	60%
Mortality	30%	20%

- Aim to get the treatment effect on mortality controlling the gender confounder

Compare apples to apples

- The gender proportion needs to be balanced

	Treatment	Control
Male	40%	60%

	Treatment	Control
Male	50%	50%

How?

The propensity score: the probability of being exposed to a cause

- The treatment condition (cause) is highly correlated with the gender variable
- That is, the propensity score $P(Z = 1|Gender)$ is 75% for male and 17% for female

	Treatment (Z=1)	Control (Z=2)
Male	30%	10%
Female	10%	50%

Magic with the propensity score

- The inverse of the propensity score $P(Z = 1|X)$ makes the treatment condition independent from the gender variable!

	Treatment (Z=1)	Control (Z=2)
Male	40%$=30\%*\frac{40}{30}$	40%$=10\%*\frac{40}{10}$
Female	60%$=10\%*\frac{60}{10}$	60%$=50\%*\frac{60}{50}$

- The odds ratio of this 2x2 table is 1 (propensity of 0.5)

Another look on the propensity score

- We would like to have a weight that makes the following equality:

$$wgt \times P(X|Z=1) = P(X).$$

By the Bayes theorem, it becomes:

$$wgt \times \frac{P(Z=1|X)P(X)}{P(Z=1)} = P(X),$$

which is

$$wgt = \frac{P(Z=1)}{P(Z=1|X)}.$$

Next slides will explain wgt in association with the potential outcome (Rubin, 2005)

Weighted prevalence for the majority class

- We weight the majority class to make it look like the minority class

- The weighted estimator for HSV2 prevalence rateis

$$\frac{\sum_i \delta_c^* \times w_c \times y}{\sum_i \delta_c^* \times w_c} = \frac{\sum_i Membership \times weights \times HSV2}{\sum_i Membership \times weights},$$

where

- Membership probability: δ_c^*=$P(Z = c|U, X, Y)$
- U is features, and X is confounding factors related to Z
- w_1 is the original NHANES weight for the majority class
- $w_2 = w_1 \times \frac{P(Z=1|X)}{P(Z=0|X)}$ for the minority class

How to get the estimator? Use EEE!

- Let P_c denotes HSV2 prevalence rate for class c.
- Then our estimate $\frac{\sum_i \delta_c^* \times w_c \times y}{\sum_i \delta_c^* \times w_c}$ for class c is the solution to

$$\sum_{i \in S} \delta_c^* \times w_c \times (y - P_c) \ = 0,$$

which is the weighted and expected estimating equation of

$$\sum_{i \in S} I(\boldsymbol{Z} = \boldsymbol{c}) \times w_c \times (y - P_c) \ = 0,$$

and the expectation was done w.r.t a membership probability: $\delta_c^* = P(Z = c | U = u, X = x, Y = y)$.

What is $\boldsymbol{\delta_c^*}$?

- δ_c^* is defined as

$$\delta_c^* = P(Z=c|U,X,Y) = \frac{P(U,X,Y,Z=c)}{P(U,X,Y)} = \frac{P(U,X,Y|Z=c)P(Z=c)}{\sum_z P(U,X,Y|Z=c)P(Z=c)}$$

- $P(U,X,Y|Z=c)$ is modeled as $P(U|Z=c,X,Y)P(Y|Z=c,X_i)P(Z=c|X)$.

- $P(U|Z=c,X,Y)$ is reduced to $P(U|Z=c)$ and to $\prod_{m=1}^{M} P(U_m|Z=c)$ by the local independence assumption

- Estimation procedure:

1) ρ of $P(U|Z=c;\rho)$ is estimated
2) α of $P(Z=c|X;\alpha)$ is estimated using $\hat{\rho}$
3) β of $P(Y|Z=c,X_i;\beta)$ is estimated using $\hat{\alpha}$ and $\hat{\rho}$ with X including a function of $P(Z=c|X)$

Propensity weights?

- w_1 is the originial sample weights for class 1, but w_2 must meet the below condition as in Ridgeway et al. (2015)

$$w_2 \times P(X|S=1,Z=2) = P(X|Z=1)$$

$$\leftrightarrow \qquad w_2 \times \frac{P(S=1|X,Z=2)\times P(Z=2|X)\times P(X)}{P(S=1,Z=2)} = \frac{P(Z=1|X)\times P(X)}{P(Z=1)}$$

$$\leftrightarrow \qquad w_2 = cosntant \times \frac{1}{P(S=1|Z=2,X)} \times \frac{P(Z=1|X)}{P(Z=2|X)}$$

$$\rightarrow \qquad w_2 = cosntant \times w_1 \times \frac{P(Z=1|X)}{P(Z=2|X)}$$

The three-step estimation

- Unlike Kang and Schafer (2010), we use stepwise estimation:

  Step 1) Build an LCA model

  Step 2) Fit a propensity model with the estimated LCA parameters from step 1

  Step 3) Estimate mean potential outcomes using estimated LCA and propensity
  parameters from previous steps

- Jackknife estimation or Taylor-linearization for the variance calculation

Propensity results

Variable	Class 1 (%)	Class 2 (%)	Balanced?	Class 1	Class 2 (propensity-weighted)	Balanced?
Age group:<=24	13.7±2.6	13.7±1.8	Yes	13.7±2.6	14.0±2.9	Yes
Age group:25-29	15.6±1.9	15.4±1.7	Yes	15.6±1.9	16.1±1.9	Yes
Age group:30-39	42.7±4.4	33.2±1.6	No	42.7±4.4	42.4±4.9	Yes
Age group:40-49	28±3.5	37.6±1.8	No	28±3.5	27.6±4	Yes
Race:Black	9.6±1.2	10.2±1.5	Yes	9.6±1.2	10.3±1.5	Yes
Race:Mex	7.3±1.5	9.8±1.1	Yes	7.3±1.5	6.6±1.4	Yes
Race:Other	6±1.5	4.4±1.2	Yes	6±1.5	5.2±1.2	Yes
Race:OtherHis	8.2±0.9	5.4±1.7	No	8.2±0.9	8.9±1.3	Yes
Race:White	69±2.9	70.2±2.7	Yes	69±2.9	69±3.1	Yes
Poverty: Yes	12.6±3.5	11.3±1.4	Yes	12.6±3.5	13.7±4.7	Yes
Education years<12	29.4±3.9	44±2.3	No	29.4±3.9	29.7±3.7	Yes
Marriage: Married	7.6±2.7	57.4±2.1	No	7.6±2.7	7.7±2.7	Yes
Marriage: Unmarried	64.8±4.1	23.6±2.1	No	64.8±4.1	64.6±4.4	Yes
Marriage: Partner	19.5±3.1	8.9±1.1	No	19.5±3.1	19.6±3.2	Yes
Marriage: Separate	8.1±2.1	10.1±1.8	Yes	8.1±2.1	8.1±2	Yes
Cocain usage: No	68.1±5.7	70.5±2	Yes	68.1±5.7	67.6±5.6	Yes
Age at first sex<18	26.4±6.1	19.4±1.3	No	26.4±6.1	26.4±6.5	Yes
Circumcised: No	13.7±1.7	18.8±2.2	No	13.7±1.7	13.8±1.6	Yes

Variable	Class 1	Class 2	P-value	Class 1	Class 2	P-value
HSV2	20.6±3.2	13.0±1.6	0.021	20.9±3.3	15.4±1.9	0.239

HSV-2:
- HSV-2 rate increased for class 2 with the propensity adjustment
- SEs and P-values were estimated by the Jack-knife resampling method

Summary

- The point estimates were assessed by the expected estimation equation frame work
- The estimating functions were expected with respect to the LCA posterior membership probability
- Variance was computed using the Jackknife method (Patterson, 2002, JASA) for simplicity purposes