Bayesian Nested Partially-Latent Models for Dependent Binary Data

Estimating Disease Etiology

Zhenke Wu

Postdoctoral Fellow
Department of Biostatistics

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

09 November 2015

R Package: https://github.com/zhenkewu/baker

Question: What's Causing Her Lung Infection?

Measurements From a Random Case

Measurements using different specimens

		BCX	PFCX	LACX	NPCX	ISCX2	PFPCR	LAPCR	NPPCR	ISPCR
Bacterium	HINF	0				0			1	1
	MCAT	0				1			1	1
	PNEU	0			1	1			1	1
	SASP	0				0			0	0
	SAUR	0				0			0	0
	BORD								0	0
	C_PNEU								0	0
	M_PNEU								0	1
	PCP								0	0
Virus	ADENOVIRUS								0	0
	CMV								0	0
	COR_229								0	0
	COR_43								0	0
	COR_63								0	0
	COR_HKU								0	0
	FLU_C								0	0
	HBOV								0	1
	HMPV_A_B								0	0
	INFLUENZA_A								0	0
	INFLUENZA_B								0	0
	PARA1								0	0
	PARA2								0	0
	PARA3								0	0
	PARA4								0	0
	PV_EV								0	0
	RHINO								0	0
	RSV_A_B								0	0

Motivating Application

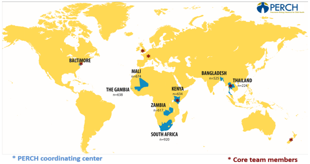
Pneumonia Etiology Research for Child Health (PERCH)

Background:

- $>$ 1 million deaths per year among children under 5
- $>$ 30 possible pathogen causes

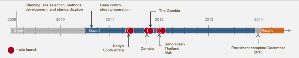
Goal:

- To determine the etiology and risk factors for pneumonia



Design:

- 7-country, case-control study
- Multiple modern diagnostic tools
- ~5,000 cases and ~5,000 controls



Common Questions on Individual and Population Health

1. a. What is the person's health state given health measurements?
   b. What is the population distribution of health states?
   (Wu et al., 2015a,b,c)
2. How to make robust inference?

Picture source: http://www.diabetesdaily.com/voices/2014/07/why-one-size-fits-all-doesnt-work-in-diabetes

Problem and Data Features

Latent health state:

- Estimating population distribution + individual diagnosis

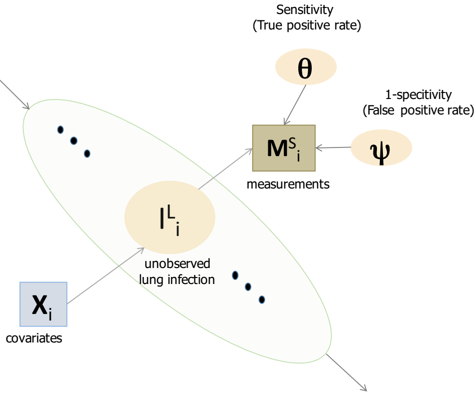
Data Features:

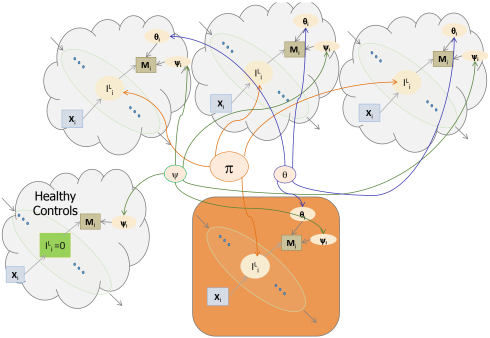
1. Gold-standard measure: few or none
2. Latent state: many categories
3. Measurements: many, with distinct error rates, missingness
4. **Blessing**: **control data**

No effective and principled methods to estimate the etiologic distribution ("pie") using such data.

Our Approach: Direct Modeling

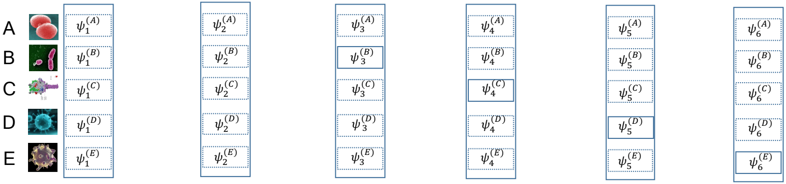
Connect Latent States and Measurements for Individual *i*





Latent Class Models (LCM)

Review



- IDEA: marginal correlations are caused by confounding of unobserved cluster indicators (I_i)
- Assumption 1: Within-Class Homogeneity

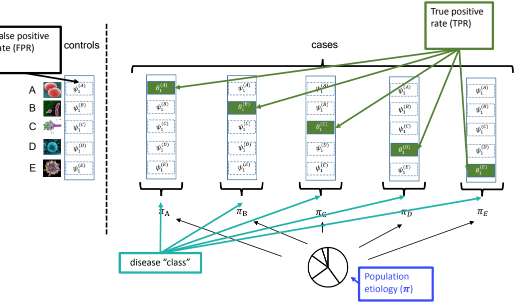
$$P[M_{ij} = 1 \mid I_i = k] = \psi_k^{(j)}, k = 1, ..., K$$

- Assumption 2: Local Independence (LI)

$$P[M_{i1} = m_1, ..., M_{iJ} = m_J \mid I_i = k] = \prod_{j=1}^{J} Pr[M_{ij} = m_j \mid I_i = k], \ \forall (m_1, ..., m_J)' = \boldsymbol{m}$$

Partially-Latent Class Models (pLCM; Wu et al. 2015a)

Model Structure



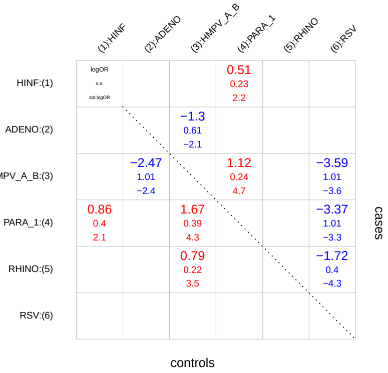
- *Partially-observed class*: Controls have no lung infection;
- *Non-interference*:

$$P(M_{[-j]} \mid Y = 0) = P(M_{[-j]} \mid I^L = j, Y = 1);$$

- *Local independence (LI)*: independence among measurements given class (I_i^L).

Next: relax both non-interference and LI assumptions.

Modeling Local Dependence (LD)



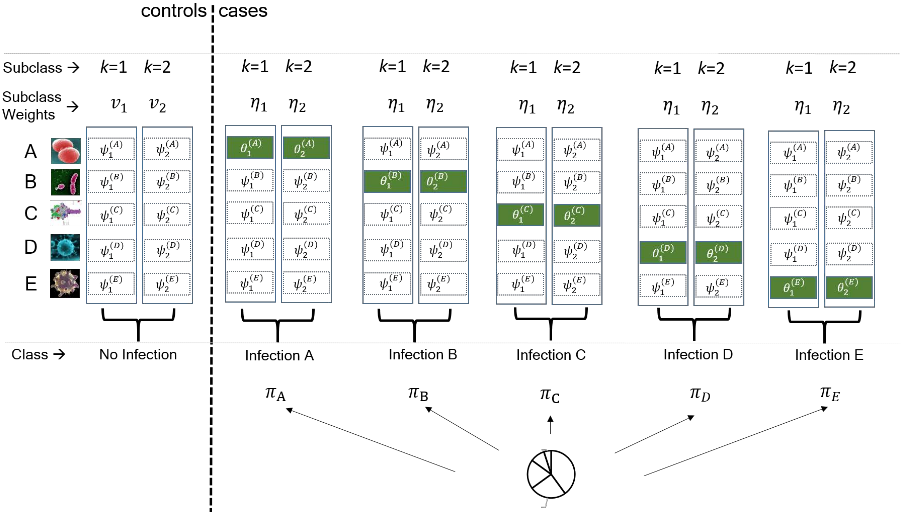
- Direct evidence from control data; symmetry (see Figure); pathogen interactions
- Impact on inference (Pepe and Janes, 2007; Albert et al., 2001)
- Modeling cross-classified probability contingency tables

$$P(M_{i1} = m_1, \ldots, M_{iJ} = m_J)$$

  - Log-linear parametrization
  - Generalized linear mixed-effect models (GLMM)
  - Mixed-membership models
  - Other non-negative decompositions

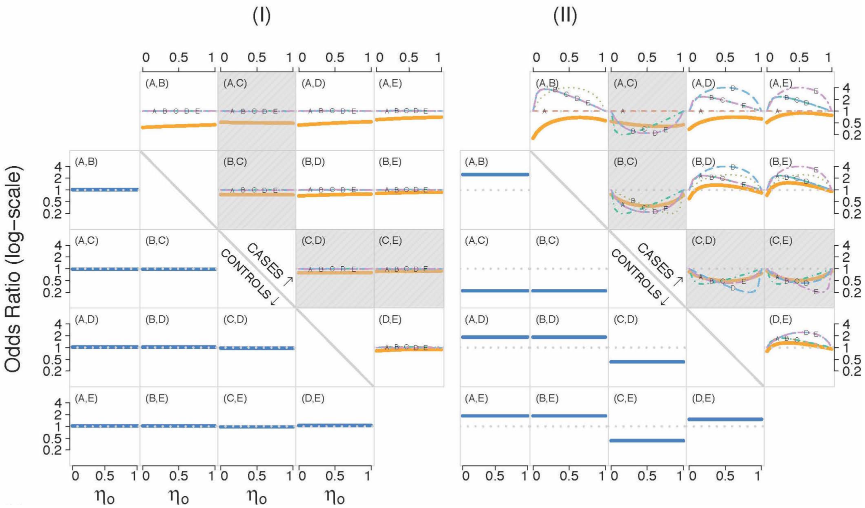
Nested pLCM

Example: 5 Pathogens, 2 Subclasses



Example: Dependence Structure; 2 Subclasses

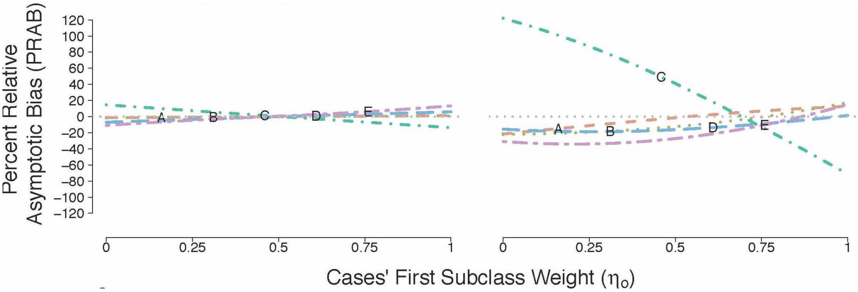
Left: weak LD *Right*: strong LD



Simulation: Relative Asymptotic Bias

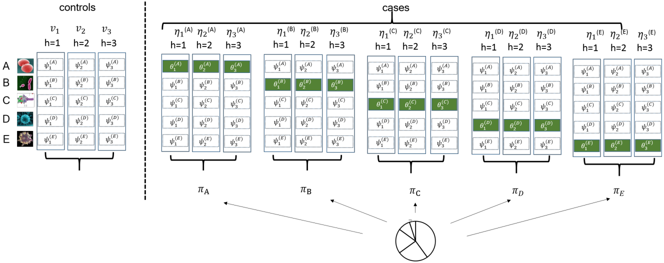
Bias if Estimated by Working LI Model (pLCM)

Left: weak LD *Right*: strong LD



Estimation in Finite Samples: How Many Subclasses?

Example: 3 Subclasses



A model selection problem:

- Extra subclasses: rich correlation structure;
- Few subclasses: parsimonious approximation in finite samples.

Proposed solution:

Model averaging by stick-breaking prior: to encourage few but allow more if data have rich dependence

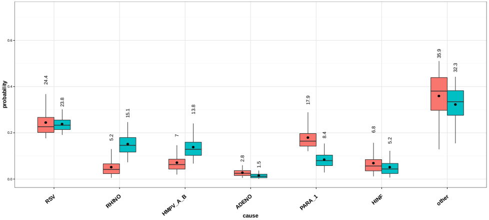
Finite-Sample Simulations: Smaller MSE by npLCM

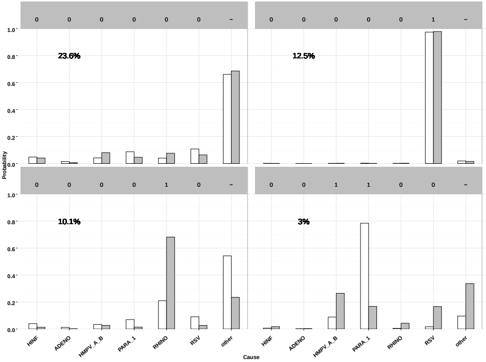
Scenario **II**: Strong LD; $N_{case} = N_{control} = 500$

	Truth: Cases' First Subclass Weight (η_o)				
	0	0.25	0.5	0.75	1
Class	100×Ratio of MSE(Standard Error)				
A	82(4)	25(1)	47(2)	115(6)	221(12)
B	516(11)	177(5)	80(3)	62(4)	140(8)
C	2379(77)	711(26)	131(7)	268(13)	357(8)
D	397(14)	152(6)	94(5)	79(4)	60(4)
E	357(13)	151(6)	102(5)	95(6)	82(5)

Table: ratio of mean squared errors (MSE) for pLCM vs npLCM. All numbers are averaged across 1,000 replications.

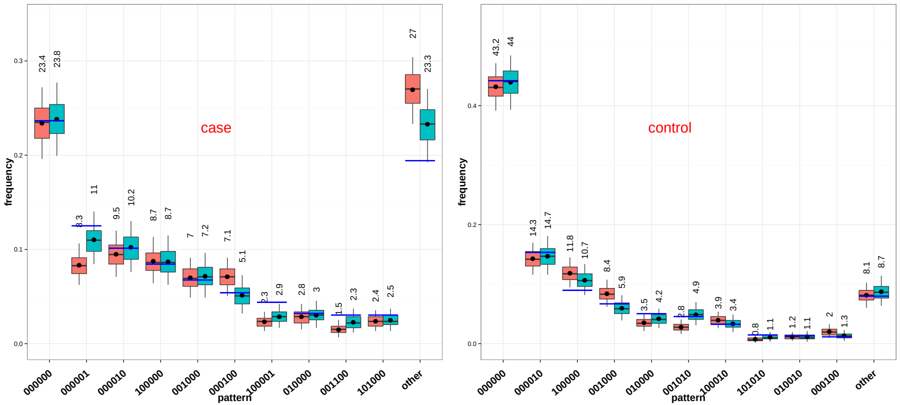
Analysis of PERCH Data





Model Checking: Frequent Binary Patterns

Left: pLCM; *Right*: npLCM

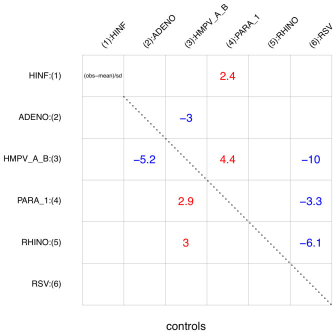


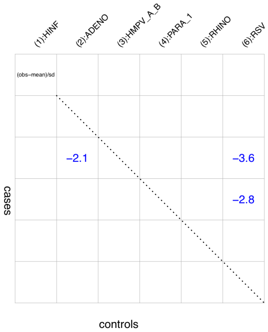
Main Points Once Again

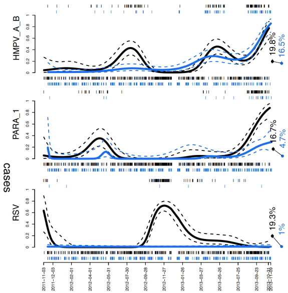
- Input: multivariate binary data in case-control studies
- Output: two histograms: 1) the fraction of cases caused by each pathogen; 2) the probability of a particular case caused by each pathogen; both given measurements.
- Proposed a larger model family (nested pLCM) to
  1) Borrow covariation and measurement precision from controls;
  2) Account for residual measurement correlations, or local dependence (LD);
  3) Parsimoniously approximate LD by sparse Bayesian fitting
- Compared to pLCM, the extended model family can
  1) Reduce bias
  2) Retain efficiency
  3) Have near-nominal coverage

Regression Analysis

Left: pLCM (bad fit) *Middle*: npLCM (improved fit) *Right*: Seasonality







Thanks!

Collaborators
Scott Zeger
Maria Deloria-Knoll
Laura Hammitt
Katherine O'Brien

Funding
Patient-Centered Outcome Research Institute [PCORI ME-1408-20318]
Bill & Melinda Gates Foundation [48968]
Hopkins Individualized Health (*in*Health) Initiative

Related Papers (More at: `zhenkewu.com`)

1. **Wu Z**, Deloria-Knoll M, Hammitt LL, and Zeger SL, for the PERCH Core Team (2015a). Partially Latent Class Models (pLCM) for Case-Control Studies of Childhood Pneumonia Etiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. doi: 10.1111/rssc.12101.
2. **Wu Z**, Zeger SL (2015b). Nested Partially-Latent Class Models for Estimating Disease Etiology from Case-Control Data. *Johns Hopkins Biostatistics Working Papers No. 276.*
3. **Wu Z**, Zeger SL (2015c). Regression Analysis for Estimating Disease Etiology from Case-Control Data. *Johns Hopkins Biostatistics Working Papers.*