Regression Analysis of Dependent Binary Data for Estimating Disease Etiology from Case-Control Studies
(“Small-Area Estimation” of Disease Etiology)

Zhenke Wu

Assistant Professor of Biostatistics
Research Assistant Professor of Michigan Institute for Data Science (MIDAS)
University of Michigan, Ann Arbor

Twitter handle: @*ZhenkeWu*

R package “baker”: `https://github.com/zhenkewu/baker`

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; “latent” ☹

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; “latent” ☹
- Scientists measure peripheral sites, e.g., nose, for presence or absence of > 30 pathogens (hence multivariate binary data)

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; "latent" ☹
- Scientists measure peripheral sites, e.g., nose, for presence or absence of > 30 pathogens (hence multivariate binary data)
- However, tests lack sensitivity or specificity ("Bronze-Standard", BrS) ☹

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; "latent" ☹
- Scientists measure peripheral sites, e.g., nose, for presence or absence of > 30 pathogens (hence multivariate binary data)
- However, tests lack sensitivity or specificity ("Bronze-Standard", BrS) ☹
- Absent lung infection, healthy **controls** provide requisite information about specificity and covariations

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; "latent" ☹
- Scientists measure peripheral sites, e.g., nose, for presence or absence of > 30 pathogens (hence multivariate binary data)
- However, tests lack sensitivity or specificity ("Bronze-Standard", BrS) ☹
- Absent lung infection, healthy **controls** provide requisite information about specificity and covariations
- **Statistical problem**: estimate cause-specific case fractions, or **"population etiologic fractions" (PEFs)**; Think "Pie chart"

Background

In large-scale **disease etiology** studies:

- some diseases may have multiple potential causes (e.g., childhood pneumonia: lung infections due to distinct pathogen causes)
- Infeasible to directly observe; "latent" ☹
- Scientists measure peripheral sites, e.g., nose, for presence or absence of >30 pathogens (hence multivariate binary data)
- However, tests lack sensitivity or specificity ("Bronze-Standard", BrS) ☹
- Absent lung infection, healthy **controls** provide requisite information about specificity and covariations
- **Statistical problem**: estimate cause-specific case fractions, or **"population etiologic fractions" (PEFs)**; Think "Pie chart"
- **Motivation for this talk**: PEFs may vary by season, a child's age, HIV status, disease severity

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$
- $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$: binary measurements; Indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$.

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$
- $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$: binary measurements; Indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$.
- Y_i: case (1) or a control (0).

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$
- $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$: binary measurements; Indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$.
- Y_i: case (1) or a control (0).
- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$: covariates that may influence case i's etiologic fractions

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$
- $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$: binary measurements; Indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$.
- Y_i: case (1) or a control (0).
- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$: covariates that may influence case i's etiologic fractions
- $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$: possibly different from $\boldsymbol{X}_i$; may influence control distribution $[\boldsymbol{M}_i \mid \boldsymbol{W}_i, Y_i = 0]$. For example, healthy controls do not have disease severity information (which can be included in $\boldsymbol{X}_i$). Cases also have $\boldsymbol{W}_i$.

Data (with Covariates)

- $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$
- $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$: binary measurements; Indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$.
- Y_i: case (1) or a control (0).
- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$: covariates that may influence case i's etiologic fractions
- $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$: possibly different from $\boldsymbol{X}_i$; may influence control distribution $[\boldsymbol{M}_i \mid \boldsymbol{W}_i, Y_i = 0]$. For example, healthy controls do not have disease severity information (which can be included in $\boldsymbol{X}_i$). Cases also have $\boldsymbol{W}_i$.
- Continuous covariates: the first p_1 and q_1 elements of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$, respectively.

Motivating Application: PERCH Study

Data : 494 cases and 944 controls from one site

Motivating Application: PERCH Study

Data : 494 cases and 944 controls from one site

Goal a. : Estimate PEFs at all covariate values, and assign cause-specific probabilities for each case

Motivating Application: PERCH Study

Data : 494 cases and 944 controls from one site

Goal a. : Estimate PEFs at all covariate values, and assign cause-specific probabilities for each case

Goal b. : Quantify overall cause-specific disease burdens in a population, i.e., overall PEFs $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ as an empirical average of the stratum-specific PEFs (by $\boldsymbol{X}$); Of policy interest (vaccine/antibiotics development and manufacture)

Motivating Application: PERCH Study

Data : 494 cases and 944 controls from one site

Goal a. : Estimate PEFs at all covariate values, and assign cause-specific probabilities for each case

Goal b. : Quantify overall cause-specific disease burdens in a population, i.e., overall PEFs $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ as an empirical average of the stratum-specific PEFs (by $\boldsymbol{X}$); Of policy interest (vaccine/antibiotics development and manufacture)

Model :

- $J = 7$: noisy presence/absence of 2 bacteria and 5 viruses in the nose
- Causes: seven single-pathogen causes plus an "Not Specified" (NoS) cause; So $L = J + 1$
- $\boldsymbol{X}_i$: enrollment date, age ($<$ or $>$ 1 year), disease severity for cases (severe or very severe), HIV status ($+$/-)
- $\boldsymbol{W}_i$: $\boldsymbol{X}_i$ minus "disease severity".

PERCH Data: Sparsely-Populated Strata☹

Table: The observed count (frequency) of cases and controls by age, disease severity and HIV status (1: yes; 0: no). The marginal fractions among cases and controls for each covariate are shown at the bottom. Regression results will be shown for the first two strata.

age ≥ 1	very severe (VS)	HIV positive	# cases (%)	# controls (%)
	(case-only)		total: 524 (100)	total: 964 (100)
0	0	0	208 (39.7)	545 (56.5)
1	0	0	72 (13.7)	278 (28.8)
0	1	0	116 (22.1)	0
1	1	0	33 (6.3)	0
0	0	1	37 (7.1)	85 (8.8)
1	0	1	24 (4.5)	51 (5.3)
0	1	1	25 (4.8)	0
1	1	1	3 (0.6)	0
case: 25.2%	34.5%	17.0%		
control: 34.3%	-	14.1%		

Current Methods (per covariate stratum)

1. Nested partially-latent class models, npLCM (Wu et al., 2017; Wu et al., 2016)

Current Methods (per covariate stratum)

1. Nested partially-latent class models, npLCM (Wu et al., 2017; Wu et al., 2016)
2. A finite mixture model for multivariate binary data, where the control component is observed, and there are L unobserved case mixture components (e.g., each representing a cause of lung infection)

Current Methods (per covariate stratum)

1. Nested partially-latent class models, npLCM (Wu et al., 2017; Wu et al., 2016)
2. A finite mixture model for multivariate binary data, where the control component is observed, and there are L unobserved case mixture components (e.g., each representing a cause of lung infection)
3. The goal is to estimate the mixing distribution/weights for the L case components (we called them PEFs, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top$)

Current Methods (per covariate stratum)

1. Nested partially-latent class models, npLCM (Wu et al., 2017; Wu et al., 2016)
2. A finite mixture model for multivariate binary data, where the control component is observed, and there are L unobserved case mixture components (e.g., each representing a cause of lung infection)
3. The goal is to estimate the mixing distribution/weights for the L case components (we called them PEFs, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top$)
4. Fitted in a Bayesian framework; can also estimate the posterior probability of the cause of disease for an **individual** case given her measurements.

Current Methods (per covariate stratum)

1. Nested partially-latent class models, npLCM (Wu et al., 2017; Wu et al., 2016)
2. A finite mixture model for multivariate binary data, where the control component is observed, and there are L unobserved case mixture components (e.g., each representing a cause of lung infection)
3. The goal is to estimate the mixing distribution/weights for the L case components (we called them PEFs, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top$)
4. Fitted in a Bayesian framework; can also estimate the posterior probability of the cause of disease for an **individual** case given her measurements.
5. Related to restricted latent class models (RLCM, Xu, 2017, AOS); Major differences: multiple sets of responses probabilities ("nested"); use control data ("partially-latent")

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM (more later) to the case-control data in each covariate stratum.

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM (more later) to the case-control data in each covariate stratum.

  Like pLCM, the npLCM is partially-identified in each stratum, necessitating multiple sets of *independent* informative priors across multiple strata.

  Two primary issues:

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM (more later) to the case-control data in each covariate stratum.

  Like pLCM, the npLCM is partially-identified in each stratum, necessitating multiple sets of *independent* informative priors across multiple strata.

  Two primary issues:

Gap 1a Unstable PEF estimates due to sparsely-populated strata.

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM (more later) to the case-control data in each covariate stratum.

  Like pLCM, the npLCM is partially-identified in each stratum, necessitating multiple sets of *independent* informative priors across multiple strata.

  Two primary issues:

Gap 1a Unstable PEF estimates due to sparsely-populated strata.

Gap 1b Informative TPR priors are often elicited for a case population and rarely for each stratum; Reusing independent prior distributions of the TPRs across all the strata will lead to overly-optimistic posterior uncertainty in $\boldsymbol{\pi}^*$, hampering policy decisions.

This talk☺

More focus on model formulation; Inference done by ‘baker‘

Extend the npLCM to perform regression analysis in case-control disease etiology studies that

This talk☺

More focus on model formulation; Inference done by 'baker'

Extend the npLCM to perform regression analysis in case-control disease etiology studies that

(a) incorporates controls to estimate the PEFs ($\boldsymbol{\pi}$),

This talk☺

More focus on model formulation; Inference done by ‘baker‘

Extend the npLCM to perform regression analysis in case-control disease etiology studies that

(a) incorporates controls to estimate the PEFs ($\boldsymbol{\pi}$),

(b) specifies parsimonious functional dependence of $\boldsymbol{\pi}$ upon covariates such as additivity, and

This talk☺

More focus on model formulation; Inference done by ‘baker‘

Extend the npLCM to perform regression analysis in case-control disease etiology studies that

(a) incorporates controls to estimate the PEFs ($\boldsymbol{\pi}$),

(b) specifies parsimonious functional dependence of $\boldsymbol{\pi}$ upon covariates such as additivity, and

(c) correctly assesses the posterior uncertainty of the PEF functions and the overall PEFs $\boldsymbol{\pi}^*$ by applying the TPR priors just once.

Okay, what is npLCM?

Quick Technical Review: Nested Partially Latent Class Models (npLCM)

For simplicity, we assume "single-pathogen causes"

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

a. PEFs/cause-specific case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top =$

$$\{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1};$$

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

a. PEFs/cause-specific case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top =$

$$\{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1};$$

b. $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$;

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

a. PEFs/cause-specific case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top =$

$$\{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1};$$

b. $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$;

c. $\boldsymbol{P}_0 = \{\boldsymbol{P}_0(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = 0, Y = 0)\}$: the same probability table as above but for controls.

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

a. PEFs/cause-specific case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top =$

$$\{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1};$$

b. $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$;

c. $\boldsymbol{P}_0 = \{\boldsymbol{P}_0(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = 0, Y = 0)\}$: the same probability table as above but for controls.

Cases' disease classes are **unobserved**, so the distribution of their measurements is a weighted finite-mixture model: $\boldsymbol{P}_1 = \sum_{\ell=1}^{L} \pi_\ell \boldsymbol{P}_{1\ell}$

npLCM Framework (no Covariates)

Three components of an npLCM likelihood function:

a. PEFs/cause-specific case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top =$

$$\{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1};$$

b. $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$;

c. $\boldsymbol{P}_0 = \{\boldsymbol{P}_0(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = 0, Y = 0)\}$: the same probability table as above but for controls.

Cases' disease classes are **unobserved**, so the distribution of their measurements is a weighted finite-mixture model: $\boldsymbol{P}_1 = \sum_{\ell=1}^{L} \pi_\ell \boldsymbol{P}_{1\ell}$

The likelihood:

$$L = L_1 \cdot L_0 = \left\{ \prod_{i: Y_i = 1} \sum_{\ell=1}^{L} \pi_\ell \cdot \boldsymbol{P}_{1\ell}(\boldsymbol{M}_i; \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \right\} \times \prod_{i': Y_{i'} = 0} \boldsymbol{P}_0(\boldsymbol{M}_{i'}; \boldsymbol{\Psi}, \boldsymbol{\nu})$$

Special Case: pLCM (Wu et al., 2016)

Setting $\eta_1 = 1$ and $\nu_1 = 1$

Control model for multivariate binary data $\{\boldsymbol{M}_i : where\ Y_i = 0\}$:

1. $\boldsymbol{P}_0(\boldsymbol{m}) = \prod_{j=1}^{J} \{\psi_j\}^{m_j} \{1 - \psi_j\}^{1-m_j} = \Pi(\boldsymbol{m}; \boldsymbol{\psi})$

   1a. $\Pi(\boldsymbol{m}; \boldsymbol{s}) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}} \{1 - s_j\}^{1-m_{ij}}$ is the probability mass function for a product Bernoulli distribution given the success probabilities $\boldsymbol{s} = (s_1, \ldots, s_J)^\top$, $0 \leq s_j \leq 1$

   1b. Parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_J)^\top$ represent the positive rates absent disease, referred to as “false positive rates” (FPRs).

Local Independence: $M_{ij} \perp M_{ij'} \mid I = 0$

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made

   2a. on the *causative* pathogen ℓ, $\mathbb{P}(M_\ell \mid I = \ell, Y = 1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1 - \theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ are "true positive rates" (TPRs), larger than FPRs.

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made
   - 2a. on the *causative* pathogen ℓ,
     $\mathbb{P}(M_\ell \mid I = \ell, Y = 1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1 - \theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ are “true positive rates” (TPRs), larger than FPRs.
   - 2b. on the *non-causative* pathogens
     $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \psi_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \psi_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$.

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made
   - 2a. on the *causative* pathogen ℓ, $\mathbb{P}(M_\ell \mid I = \ell, Y = 1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1 - \theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ are "true positive rates" (TPRs), larger than FPRs.
   - 2b. on the *non-causative* pathogens $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\psi}_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$.
   - 2c. Under the single-pathogen-cause assumption, pLCM uses J TPRs $\boldsymbol{\theta}$ for $L = J$ causes and J FPRs $\boldsymbol{\psi}$.

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made
   - 2a. on the *causative* pathogen ℓ, $\mathbb{P}(M_\ell \mid I=\ell, Y=1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1-\theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ are "true positive rates" (TPRs), larger than FPRs.
   - 2b. on the *non-causative* pathogens $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \psi_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \psi_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$.
   - 2c. Under the single-pathogen-cause assumption, pLCM uses J TPRs $\boldsymbol{\theta}$ for $L = J$ causes and J FPRs ψ.

2a-2b: Local Independence (LI): $M_{ij} \perp M_{ij'} \mid I = \ell \neq 0$

Special Case: pLCM (Wu et al., 2016)

Model for the multivariate binary data in case class $\ell \neq 0$

2. $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ is a product of the probabilities of measurements made
   - 2a. on the *causative* pathogen ℓ, $\mathbb{P}(M_\ell \mid I = \ell, Y = 1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1-\theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ are "true positive rates" (TPRs), larger than FPRs.
   - 2b. on the *non-causative* pathogens $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \psi_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \psi_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$.
   - 2c. Under the single-pathogen-cause assumption, pLCM uses J TPRs $\boldsymbol{\theta}$ for $L = J$ causes and J FPRs ψ.

2a-2b: Local Independence (LI): $M_{ij} \perp M_{ij'} \mid I = \ell \neq 0$

2a-2b. Non-interference: disease-causing pathogen(s) are more frequently detected among cases than controls ($\theta_\ell > \psi_\ell$) and the non-causative pathogens are observed with the same rates among cases as in controls

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$
  - test cross-reactions (prevented in PERCH assays)
  - lab technicians effect
  - heterogeneity in subjects' immunity level

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$
    - test cross-reactions (prevented in PERCH assays)
    - lab technicians effect
    - heterogeneity in subjects' immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1},...,M_{iJ})'$
    - test cross-reactions (prevented in PERCH assays)
    - lab technicians effect
    - heterogeneity in subjects' immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)
- Modeling Deviation from LI Modeling a cross-classified probability contingency table

$$\mathbb{P}[M_{i1}=m_1,...,M_{iJ}=m_J],\ \forall \boldsymbol{m}=(m_1,...,m_J)'$$

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$
    - test cross-reactions (prevented in PERCH assays)
    - lab technicians effect
    - heterogeneity in subjects' immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)
- Modeling Deviation from LI Modeling a cross-classified probability contingency table

$$\mathbb{P}[M_{i1} = m_1, ..., M_{iJ} = m_J],\ \forall \boldsymbol{m} = (m_1, ..., m_J)'$$

- Log-linear parametrization

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$
  - test cross-reactions (prevented in PERCH assays)
  - lab technicians effect
  - heterogeneity in subjects' immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)
- Modeling Deviation from LI Modeling a cross-classified probability contingency table

$$\mathbb{P}[M_{i1} = m_1, ..., M_{iJ} = m_J],\ \forall \boldsymbol{m} = (m_1, ..., m_J)'$$

  - Log-linear parametrization
  - Generalized linear mixed-effect models (GLMM)

“nested” pLCM

Relax the LI and Non-interference Assumption

- Direct evidence: control measurements $(M_{i1}, ..., M_{iJ})'$
    - test cross-reactions (prevented in PERCH assays)
    - lab technicians effect
    - heterogeneity in subjects’ immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)
- Modeling Deviation from LI Modeling a cross-classified probability contingency table

$$\mathbb{P}[M_{i1} = m_1, ..., M_{iJ} = m_J],\ \forall \boldsymbol{m} = (m_1, ..., m_J)'$$

- Log-linear parametrization
- Generalized linear mixed-effect models (GLMM)
- Simplex factor model; similar to mixed-membership model (Cf. Bhattacharya and Dunson, 2012, *JASA*)

“nested” pLCM

Relax the LI and Non-interference Assumption

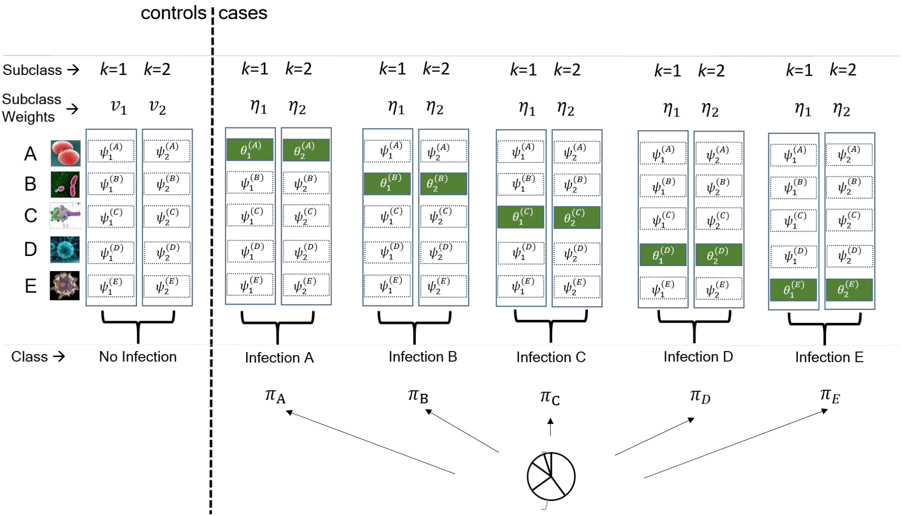
- Direct evidence: control measurements $(M_{i1},\ldots,M_{iJ})'$
    - test cross-reactions (prevented in PERCH assays)
    - lab technicians effect
    - heterogeneity in subjects’ immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert et al., 2001, *Biometrics*)
- Modeling Deviation from LI Modeling a cross-classified probability contingency table

$$\mathbb{P}[M_{i1}=m_1,\ldots,M_{iJ}=m_J],\ \forall \boldsymbol{m}=(m_1,\ldots,m_J)'$$

- Log-linear parametrization
- Generalized linear mixed-effect models (GLMM)
- Simplex factor model; similar to mixed-membership model (Cf. Bhattacharya and Dunson, 2012, *JASA*)
- PARAFAC decomposition (Cf. Dunson and Xing, 2009, *JASA*)

Nested Partially-Latent Class Models (npLCM; Wu and Zeger, 2016)

Example: 5 Pathogens, 2 Subclasses; BrS Data Only



controls | cases

Subclass → k=1 k=2 (repeated for each class)

Subclass Weights → ν_1 ν_2 (controls); η_1 η_2 (cases)

A, B, C, D, E

No Infection: $\psi_1^{(A)}$ $\psi_2^{(A)}$, $\psi_1^{(B)}$ $\psi_2^{(B)}$, $\psi_1^{(C)}$ $\psi_2^{(C)}$, $\psi_1^{(D)}$ $\psi_2^{(D)}$, $\psi_1^{(E)}$ $\psi_2^{(E)}$

Infection A: $\theta_1^{(A)}$ $\theta_2^{(A)}$, $\psi_1^{(B)}$ $\psi_2^{(B)}$, $\psi_1^{(C)}$ $\psi_2^{(C)}$, $\psi_1^{(D)}$ $\psi_2^{(D)}$, $\psi_1^{(E)}$ $\psi_2^{(E)}$

Infection B: $\psi_1^{(A)}$ $\psi_2^{(A)}$, $\theta_1^{(B)}$ $\theta_2^{(B)}$, $\psi_1^{(C)}$ $\psi_2^{(C)}$, $\psi_1^{(D)}$ $\psi_2^{(D)}$, $\psi_1^{(E)}$ $\psi_2^{(E)}$

Infection C: $\psi_1^{(A)}$ $\psi_2^{(A)}$, $\psi_1^{(B)}$ $\psi_2^{(B)}$, $\theta_1^{(C)}$ $\theta_2^{(C)}$, $\psi_1^{(D)}$ $\psi_2^{(D)}$, $\psi_1^{(E)}$ $\psi_2^{(E)}$

Infection D: $\psi_1^{(A)}$ $\psi_2^{(A)}$, $\psi_1^{(B)}$ $\psi_2^{(B)}$, $\psi_1^{(C)}$ $\psi_2^{(C)}$, $\theta_1^{(D)}$ $\theta_2^{(D)}$, $\psi_1^{(E)}$ $\psi_2^{(E)}$

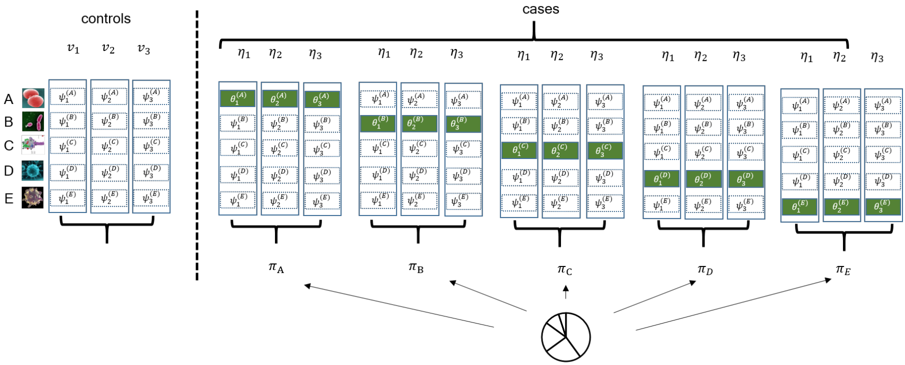
Infection E: $\psi_1^{(A)}$ $\psi_2^{(A)}$, $\psi_1^{(B)}$ $\psi_2^{(B)}$, $\psi_1^{(C)}$ $\psi_2^{(C)}$, $\psi_1^{(D)}$ $\psi_2^{(D)}$, $\theta_1^{(E)}$ $\theta_2^{(E)}$

Class → No Infection, Infection A, Infection B, Infection C, Infection D, Infection E

π_A π_B π_C π_D π_E

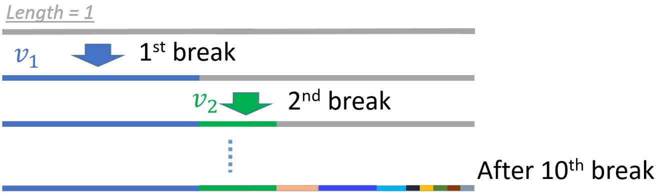
Nested Partially-Latent Class Models (npLCM; Wu and Zeger, 2016)

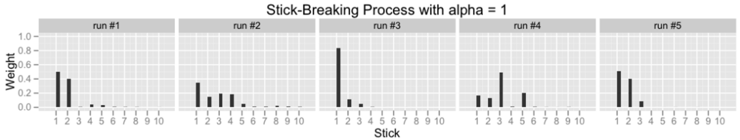
Example: 5 Pathogens, 3 Subclasses; BrS Data Only


controls
cases
ν_1 ν_2 ν_3
A
B
C
D
E
$\psi_1^{(A)}$ $\psi_2^{(A)}$ $\psi_3^{(A)}$
$\psi_1^{(B)}$ $\psi_2^{(B)}$ $\psi_3^{(B)}$
$\psi_1^{(C)}$ $\psi_2^{(C)}$ $\psi_3^{(C)}$
$\psi_1^{(D)}$ $\psi_2^{(D)}$ $\psi_3^{(D)}$
$\psi_1^{(E)}$ $\psi_2^{(E)}$ $\psi_3^{(E)}$
η_1 η_2 η_3
$\theta_1^{(A)}$ $\theta_2^{(A)}$ $\theta_3^{(A)}$
$\theta_1^{(B)}$ $\theta_2^{(B)}$ $\theta_3^{(B)}$
$\theta_1^{(C)}$ $\theta_2^{(C)}$ $\theta_3^{(C)}$
$\theta_1^{(D)}$ $\theta_2^{(D)}$ $\theta_3^{(D)}$
$\theta_1^{(E)}$ $\theta_2^{(E)}$ $\theta_3^{(E)}$
π_A
π_B
π_C
π_D
π_E

Encourage Few Subclasses: Stick-Breaking Prior

$V_j \sim \text{Beta}(1, \alpha)$; Example: $K = 10$, $\alpha = 1$





- On average, the first several segments receive most weights

npLCM: Likelihood and Prior

BrS Data Only

- Likelihood

$$
\begin{aligned}
\boldsymbol{P}_0(\boldsymbol{M}_i=\boldsymbol{m}) &= \sum_{k=1}^{K}\nu_k\prod_{j=1}^{J}\left\{\psi_k^{(j)}\right\}^{m_j}\left\{1-\psi_k^{(j)}\right\}^{1-m_j},\\
\boldsymbol{P}_1(\boldsymbol{M}_i=\boldsymbol{m}) &= \sum_{j=1}^{J}\pi_j\sum_{k=1}^{K}\left[\eta_k\left\{\theta_k^{(j)}\right\}^{m_j}\left\{1-\theta_k^{(j)}\right\}^{1-m_j}\prod_{\ell\neq j}\left\{\psi_k^{(j)}\right\}^{m_\ell}\left\{1-\psi_k^{(j)}\right\}^{1-m_\ell}\right],
\end{aligned}
$$

- Prior:

$$
\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(.5,\ldots,.5),\\
\psi_k^{(j)} &\sim \text{Beta}(1,1),\ \ \theta_k\sim\text{Beta}(c_{1kj},c_{2kj}), j=1,...,J; k=1,...,\infty,\\
Z_{i'}\mid I_{i'}^{L}=j &\sim \sum_{k=1}^{\infty}U_k\prod_{\ell<k}\left[1-U_\ell\right]\delta_k,\qquad U_k\sim\text{Beta}(1,\alpha_0),\ \text{for all cases},\\
Z_i &\sim \sum_{k=1}^{\infty}V_k\prod_{\ell<k}[1-V_\ell]\delta_k,\qquad V_k\sim\text{Beta}(1,\alpha_0),\ \text{for all controls},\\
\alpha_0 &\sim \text{Gamma}(0.25,0.25),
\end{aligned}
$$

Now, how to incorporate covariates, to which quantities?

Regression Extension for P_0 and P_1: letting π_ℓ, ν_k, η_k depend on covariates

Roadmap

Let three sets of parameters in an npLCM (pg.17) depend on the observed covariates

1x. Etiology regression function among cases, $\{\pi_\ell(\boldsymbol{x}), \ell \neq 0\}$, which is of primary scientific interest

2x. Conditional probability of measurements $\boldsymbol{m}$ given covariates $\boldsymbol{w}$ in controls: $\boldsymbol{P}_0(\boldsymbol{m}; \boldsymbol{w}) = [\boldsymbol{M} = \boldsymbol{m} \mid \boldsymbol{W} = \boldsymbol{w}, I = 0]$,

3x. 2x above, but in the case class ℓ: $\boldsymbol{P}_{1\ell}(\boldsymbol{m}; \boldsymbol{w}) = [\boldsymbol{M} = \boldsymbol{m} \mid \boldsymbol{W} = \boldsymbol{w}, I = \ell]$, $\ell = 1, \ldots, L$

note Keep the specifications for the TPRs and FPRs ($\boldsymbol{\Theta}$, $\boldsymbol{\Psi}$) as in the original npLCM.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.
3. Over-parameterized multinomial logistic regression: $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\}/\sum_{\ell'=1}^{L}\exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, ..., L$, where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.
3. Over-parameterized multinomial logistic regression: $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\}/\sum_{\ell'=1}^{L}\exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, ..., L$, where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$.
4. Without specifying a baseline category, we treat all the disease classes symmetrically which simplifies prior specification.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.
3. Over-parameterized multinomial logistic regression: $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\}/\sum_{\ell'=1}^{L} \exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, \ldots, L$, where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$.
4. Without specifying a baseline category, we treat all the disease classes symmetrically which simplifies prior specification.
5. Additive models for $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi$

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.
3. Over-parameterized multinomial logistic regression: $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\}/\sum_{\ell'=1}^{L}\exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, ..., L$, where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$.
4. Without specifying a baseline category, we treat all the disease classes symmetrically which simplifies prior specification.
5. Additive models for $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \gamma_\ell^\pi$

5a. Use B-spline basis expansion to approximate $f_{\ell j}^\pi(\cdot)$ and use P-spline for estimating smooth functions.

Etiology Regression $\boldsymbol{\pi}_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference.

1. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ.
2. Occurs with probability $\pi_{i\ell}$ that depends upon covariates.
3. Over-parameterized multinomial logistic regression: $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\}/\sum_{\ell'=1}^{L} \exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, ..., L$, where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$.
4. Without specifying a baseline category, we treat all the disease classes symmetrically which simplifies prior specification.
5. Additive models for $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \gamma_\ell^\pi$

5a. Use B-spline basis expansion to approximate $f_{\ell j}^\pi(\cdot)$ and use P-spline for estimating smooth functions.

5b. $\widetilde{\boldsymbol{x}}$ is the subvector of the predictors $\boldsymbol{x}$; $\boldsymbol{\Gamma}_\ell^\pi = (\boldsymbol{\beta}_{\ell j}^\pi, \boldsymbol{\gamma}_\ell^\pi)$.

$\boldsymbol{P}_0$: Multivariate binary regression for controls

Desirable properties

Model Specification:

- Model space large enough for complex conditional dependence of $\boldsymbol{M}$ given covariates $\boldsymbol{W}$
- Upward compatibility, or reproducibility (invariant parameter interpretation with increasing dimensions or complex patterns of missing responses)

Estimation:

- Consistency and Efficiency
- Adaptivity: regularization to adapt to the difficulty of the problem, e.g., model residual dependence $[\boldsymbol{M} \mid \boldsymbol{W}, I = 0]$ only if necessary; model the effect of covariates only if necessary

Let $\boldsymbol{P}_0$ depend on $\boldsymbol{W}_i$

Regression model for controls

- The pmf for controls' measurements: $Pr(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{W}_i, I_i = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}_i)\Pi(\boldsymbol{m}; \Psi_k)$, $\Psi_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})'$

Let P_0 depend on $\boldsymbol{W}_i$

Regression model for controls

- The pmf for controls' measurements:
  $Pr(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{W}_i, I_i = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}_i)\Pi(\boldsymbol{m}; \Psi_k)$,
  $\Psi_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})'$
  - The vector $(\nu_1(\boldsymbol{W}_i), \ldots, \nu_K(\boldsymbol{W}_i))$ lies in a $(K-1)$-simplex

Let P_0 depend on $\boldsymbol{W}_i$

Regression model for controls

- The pmf for controls' measurements:
  $Pr(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{W}_i, I_i = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}_i)\Pi(\boldsymbol{m}; \Psi_k)$,
  $\Psi_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})'$
  - The vector $(\nu_1(\boldsymbol{W}_i), \ldots, \nu_K(\boldsymbol{W}_i))$ lies in a $(K-1)$-simplex
  - $\Pi(\boldsymbol{m}; s) = \prod_{j=1}^{J}\{s_j\}^{m_{ij}}(1-s_j)^{1-m_{ij}}$

Let P_0 depend on $\boldsymbol{W}_i$

Regression model for controls

- The pmf for controls' measurements: $Pr(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{W}_i, I_i = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}_i)\Pi(\boldsymbol{m}; \Psi_k)$, $\Psi_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})'$
    - The vector $(\nu_1(\boldsymbol{W}_i), \ldots, \nu_K(\boldsymbol{W}_i))$ lies in a $(K-1)$-simplex
    - $\Pi(\boldsymbol{m}; s) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}}(1-s_j)^{1-m_{ij}}$
- An equivalent generative process:

Let P_0 depend on W_i

Regression model for controls

- The pmf for controls' measurements:
  $Pr(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{W}_i, I_i = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}_i)\Pi(\boldsymbol{m}; \Psi_k)$,
  $\Psi_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})'$
  - The vector $(\nu_1(\boldsymbol{W}_i), \ldots, \nu_K(\boldsymbol{W}_i))$ lies in a $(K-1)$-simplex
  - $\Pi(\boldsymbol{m}; s) = \prod_{j=1}^{J}\{s_j\}^{m_{ij}}(1-s_j)^{1-m_{ij}}$
- An equivalent generative process:

$$
\begin{aligned}
\text{sample subclass indicator}: &\quad Z_i \mid \boldsymbol{W}_i \sim \text{Categorical}_K(\boldsymbol{\nu}(\boldsymbol{W}_i))\\
\text{generate measurements}: &\quad M_{ij} \mid Z_i = k \sim \text{Bernoulli}(\psi_k^{(j)}),\\
&\quad \text{independently for } j = 1, ..., J.
\end{aligned}
$$

Let $\boldsymbol{P}_0$ depend on $\boldsymbol{W}_i$

Regression model for controls Stick-breaking parametrization of weight functions $\nu_k(\boldsymbol{W}_i) = P(Z_i = k \mid \boldsymbol{W}_i)$ by

$$\underbrace{h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\nu})}_{stick\ \ k} = \begin{cases} g(\alpha_{ik}^{\nu}) \prod_{s<k} \left\{1 - g(\alpha_{is}^{\nu})\right\}, & \text{if } k < K, \\ \prod_{s<k} \left\{1 - g(\alpha_{is}^{\nu})\right\}, & \text{if } k = K, \end{cases}$$

Let P_0 depend on $\boldsymbol{W}_i$

Regression model for controls Stick-breaking parametrization of weight functions $\nu_k(\boldsymbol{W}_i) = P(Z_i = k \mid \boldsymbol{W}_i)$ by

$$\underbrace{h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^\nu)}_{stick\ \ k} = \begin{cases} g(\alpha_{ik}^\nu) \prod_{s<k} \{1 - g(\alpha_{is}^\nu)\}, & \text{if } k < K, \\ \prod_{s<k} \{1 - g(\alpha_{is}^\nu)\}, & \text{if } k = K, \end{cases}$$

We specify α_{ik}^ν via additive models at g^{-1} scale $(g(\cdot) = 1/(1 + \exp\{-(\cdot)\}))$:

$$\alpha_{ik}^\nu = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(\boldsymbol{W}_{ij}; \beta_{kj}^\nu) + \widetilde{\boldsymbol{W}}_i^\top \gamma_k^\nu,\ \ k = 1, \ldots, K-1.$$

Let $\boldsymbol{P}_0$ depend on $\boldsymbol{W}_i$

Regression model for controls Stick-breaking parametrization of weight functions $\nu_k(\boldsymbol{W}_i) = P(Z_i = k \mid \boldsymbol{W}_i)$ by

$$\underbrace{h_k(\boldsymbol{W}_i;\boldsymbol{\Gamma}_k^{\nu})}_{stick\ \ k} = \begin{cases} g(\alpha_{ik}^{\nu})\prod_{s<k}\{1-g(\alpha_{is}^{\nu})\}, & \text{if } k<K, \\ \prod_{s<k}\{1-g(\alpha_{is}^{\nu})\}, & \text{if } k=K, \end{cases}$$

We specify α_{ik}^{ν} via additive models at g^{-1} scale $(g(\cdot) = 1/(1+\exp\{-(\cdot)\}))$:

$$\alpha_{ik}^{\nu} = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(\boldsymbol{W}_{ij};\boldsymbol{\beta}_{kj}^{\nu}) + \widetilde{\boldsymbol{W}}_i^{\top}\gamma_k^{\nu},\ \ k = 1,\ldots,K-1.$$

Expand the smooth functions by B-spline bases with coefficients $\boldsymbol{\beta}_{kj}^{\nu}$; $\widetilde{\boldsymbol{w}}$ is a subvector of covariates $\boldsymbol{w}$

Adaptivity Considerations☺

Proposed Model

- Prevent overfitting when the regression is easy, and improve interpretability
- We *a priori* place substantial probabilities on models with the following two features:
  a) Few subclasses with effective weights (in the sense that $\nu_k(\cdot)$ is bounded away from 0 and 1): a novel additive half-Cauchy prior for μ_{k0}.
  b) Smooth weight regression curves $\nu_k(\cdot)$: by Bayesian Penalized-Splines (P-Splines) combined with mixture priors on spline coefficients to sensitively distinguish constant $\alpha_k^\nu(\cdot)$ from flexible smooth curves

On Consideration a) Selective Stopping, or “Uniform Shrinkage over Simplex” for $\nu_k(\boldsymbol{W})$

Proposed Model

On Consideration a) Selective Stopping, or "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu^{*}_{k0}$ where $u_{kj}, 1 \leq j \leq k$ are pre-specified trangular array of positive values. a large k has a large intercept: increasing burden only when justified

On Consideration a) Selective Stopping, or "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu^*_{k0}$ where $u_{kj}, 1 \leq j \leq k$ are pre-specified trangular array of positive values. a large k has a large intercept: increasing burden only when justified
- $u_{kj} = 1, j = 1, \ldots, k$; other choices: $u_{kj} = \mathbb{I}\{k = j\}$ or $u_{kj} = 1/k$ may be useful in other settings.

On Consideration a) Selective Stopping, or "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu^*_{k0}$ where $u_{kj}, 1 \leq j \leq k$ are pre-specified trangular array of positive values. a large k has a large intercept: increasing burden only when justified
- $u_{kj} = 1, j = 1, \ldots, k$; other choices: $u_{kj} = \mathbb{I}\{k = j\}$ or $u_{kj} = 1/k$ may be useful in other settings.
- We specify the prior distributions for μ^*_{k0} to be heavy-tailed:

$$\mu^*_{k0} \sim Cauchy^{+}(0, s_k),\ k = 1, \ldots, K,$$

On Consideration a) Selective Stopping, or "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu_{k0}^{*}$ where $u_{kj}, 1 \leq j \leq k$ are pre-specified trangular array of positive values. a large k has a large intercept: increasing burden only when justified
- $u_{kj} = 1, j = 1, \ldots, k$; other choices: $u_{kj} = \mathbb{I}\{k = j\}$ or $u_{kj} = 1/k$ may be useful in other settings.
- We specify the prior distributions for μ_{k0}^{*} to be heavy-tailed:

$$\mu_{k0}^{*} \sim \mathit{Cauchy}^{+}(0, s_k),\ k = 1, \ldots, K,$$

- A large s_k produce large μ_{k0}^{*} and help stop stick-breaking at class k, while small values let the stick-breaking continue to Step $k+1$.

On Consideration a) Selective Stopping, or "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

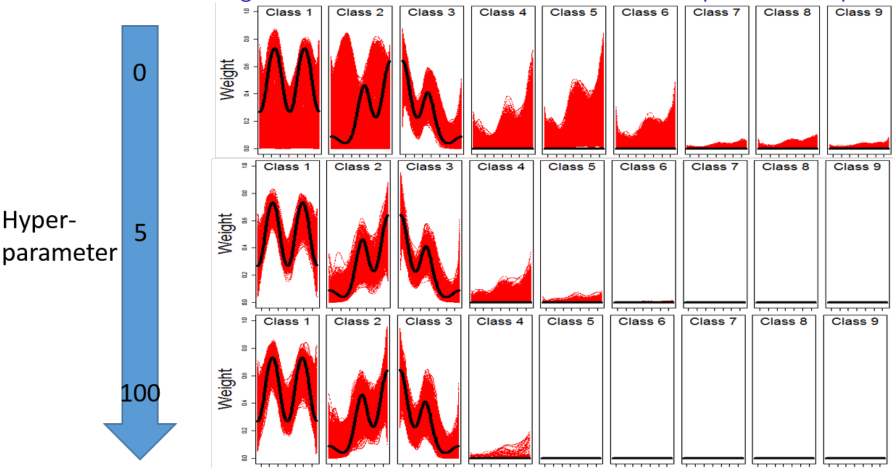
- We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu^*_{k0}$ where $u_{kj}, 1 \leq j \leq k$ are pre-specified trangular array of positive values. a large k has a large intercept: increasing burden only when justified
- $u_{kj} = 1, j = 1, \ldots, k$; other choices: $u_{kj} = \mathbb{I}\{k = j\}$ or $u_{kj} = 1/k$ may be useful in other settings.
- We specify the prior distributions for μ^*_{k0} to be heavy-tailed:

$$\mu^*_{k0} \sim Cauchy^+(0, s_k),\ k = 1, \ldots, K,$$

- A large s_k produce large μ^*_{k0} and help stop stick-breaking at class k, while small values let the stick-breaking continue to Step $k + 1$.
- Encourages using a small number of effective classes ($< K$) to approximate the observed 2^J probability contingency table in finite samples

Inference of $\nu_k(x)$ at three hyperparameter values s_k

Simulation: with a single continuous covariate; "—": truth, "—": posterior samples



Hyper-
parameter

X-axis: covariate values

Y-axis: weight; 0 to 1.

Let P_1 depend on X and W

Subclass Weight Regression: For Cases

Let $\boldsymbol{P}_1$ depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$$

Let P_1 depend on X and W

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$

Let $\boldsymbol{P}_1$ depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise

Let P_1 depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise
- Subclass weight regression $\eta_k(\boldsymbol{W})$ is also specified via stick-breaking: $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\eta})$, $k = 1, \ldots, K-1$

Let $\boldsymbol{P}_1$ depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ: $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise
- Subclass weight regression $\eta_k(\boldsymbol{W})$ is also specified via stick-breaking: $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\eta})$, $k = 1, \ldots, K-1$
- α_{ik}^{η}: GAMs

Let P_1 depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ: $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise
- Subclass weight regression $\eta_k(\boldsymbol{W})$ is also specified via stick-breaking: $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\eta})$, $k = 1, \ldots, K-1$
- α_{ik}^{η}: GAMs
- $\alpha_{ik}^{\eta} = \alpha_k^{\eta}(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\boldsymbol{\eta}}) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(W_{ij}; \boldsymbol{\beta}_{kj}^{\eta}) + \widetilde{\boldsymbol{W}}_i^{\top} \boldsymbol{\gamma}_k^{\eta}$, where $\boldsymbol{\Gamma}_k^{\eta} = \{\mu_{k0}, \{\boldsymbol{\beta}_{kj}^{\eta}\}, \boldsymbol{\gamma}_k^{\eta}\}$ are the regression parameters.

Let $\boldsymbol{P}_1$ depend on $\boldsymbol{X}$ and $\boldsymbol{W}$

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

$$Pr(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_{i\ell} \sum_{k=1}^{K} \eta_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$$

- $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ are positive rates for J measurements in subclass k of disease class ℓ: $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise
- Subclass weight regression $\eta_k(\boldsymbol{W})$ is also specified via stick-breaking: $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\eta})$, $k = 1, \ldots, K-1$
- α_{ik}^{η}: GAMs
- $\alpha_{ik}^{\eta} = \alpha_k^{\eta}(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\boldsymbol{\eta}}) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(W_{ij}; \boldsymbol{\beta}_{kj}^{\eta}) + \widetilde{\boldsymbol{W}}_i^{\top} \boldsymbol{\gamma}_k^{\eta}$, where $\boldsymbol{\Gamma}_k^{\eta} = \{\mu_{k0}, \{\boldsymbol{\beta}_{kj}^{\eta}\}, \boldsymbol{\gamma}_k^{\eta}\}$ are the regression parameters.
- we use μ_{k0} from the controls (why?)

npLCM Regression Framework

The npLCM regression framework is then obtained as:

npLCM Regression Framework

The npLCM regression framework is then obtained as:

- Control likelihood with covariates: $L_0^{\mathrm{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$.

npLCM Regression Framework

The npLCM regression framework is then obtained as:

- Control likelihood with covariates:
  $L_0^{\text{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$.
- Cases likelihood with covariates:

$$L_1^{\text{reg}} = \prod_{i:Y_i=1} \left\{ \sum_{\ell=1}^{L} \left[\underbrace{\pi_\ell(\boldsymbol{X}_i; \boldsymbol{\Gamma}_\ell^\pi)}_{PEF\ \ell} \sum_{k=1}^{K} \{\eta_{ik} \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})\} \right] \right\} \quad (2)$$

- $\nu_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^\nu)$: The S$\underline{????}$-B$\underline{????}$ parameterization
- $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^\eta)$

npLCM Regression Framework

The npLCM regression framework is then obtained as:

- Control likelihood with covariates:
  $L_0^{\text{reg}} = \prod_{i:Y_i=0}\sum_{k=1}^{K}\nu_{ik}\Pi(\boldsymbol{M}_i;\boldsymbol{\Psi}_k)$.
- Cases likelihood with covariates:

$$L_1^{\text{reg}} = \prod_{i:Y_i=1}\left\{\sum_{\ell=1}^{L}\left[\underbrace{\pi_\ell(\boldsymbol{X}_i;\boldsymbol{\Gamma}_\ell^{\pi})}_{PEF\ \ell}\sum_{k=1}^{K}\{\eta_{ik}\cdot\Pi(\boldsymbol{M}_i;\boldsymbol{p}_{k\ell})\}\right]\right\} \tag{2}$$

- $\nu_{ik} = h_k(\boldsymbol{W}_i;\boldsymbol{\Gamma}_k^{\nu})$: The S$\underline{????}$-B$\underline{????}$ parameterization
- $\eta_{ik} = h_k(\boldsymbol{W}_i;\boldsymbol{\Gamma}_k^{\eta})$

The joint likelihood for the regression model can be written as:
$L^{\text{reg}} = L_1^{\text{reg}} \times L_0^{\text{reg}}$.

Prior Specifications

Unknown parameters:

- etiology regression coefficients ($\{\mathbf{\Gamma}_\ell^\pi\}$),
- subclass mixing weight parameters for cases ($\{\mathbf{\Gamma}_k^\eta\}$) and controls ($\{\mathbf{\Gamma}_k^\nu\}$),
- true and false positive rates ($\mathbf{\Theta} = \{\theta_k^{(j)}\}, \mathbf{\Psi} = \{\psi_k^{(j)}\}$).

Prior Specifications

Unknown parameters:

- etiology regression coefficients ($\{\mathbf{\Gamma}^{\pi}_{\ell}\}$),
- subclass mixing weight parameters for cases ($\{\mathbf{\Gamma}^{\eta}_{k}\}$) and controls ($\{\mathbf{\Gamma}^{\nu}_{k}\}$),
- true and false positive rates ($\mathbf{\Theta} = \{\theta^{(j)}_k\}, \mathbf{\Psi} = \{\psi^{(j)}_k\}$).

To avoid potential overfitting, we *a priori* introduce:

- (a) few non-trivial subclasses via novel additive half-Cauchy prior for the intercepts $\{\mu_{k0}\}$
- (b) for continuous variable: smooth regression curves $\pi_{\ell}(\cdot)$, $\nu_k(\cdot)$ and $\eta_k(\cdot)$ by Bayesian Penalized-splines (Lang, 2004) combined with shrinkage priors on spline coefficients (Ni et.al, 2015) (to encourage towards constant values)

Posterior Inference

Use Markov chain Monte Carlo (MCMC) algorithm to approximate joint posterior distribution

Posterior Inference

Use Markov chain Monte Carlo (MCMC) algorithm to approximate joint posterior distribution

- Posterior inference is flexible and can be obtained from any functions of model parameters and individual latent variables

Posterior Inference

Use Markov chain Monte Carlo (MCMC) algorithm to approximate joint posterior distribution

- Posterior inference is flexible and can be obtained from any functions of model parameters and individual latent variables

Fit npLCMs (w/ or w/out covariates using R package baker (`https://github.com/zhenkewu/baker`)

Posterior Inference

Use Markov chain Monte Carlo (MCMC) algorithm to approximate joint posterior distribution

- Posterior inference is flexible and can be obtained from any functions of model parameters and individual latent variables

Fit npLCMs (w/ or w/out covariates using R package baker (`https://github.com/zhenkewu/baker`)

- calls Bayesian model fitting software JAGS 4.2.0 (Plummer et al., 2003) from within R
- provides functions to visualize the posterior distributions of the unknowns
- also performs posterior predictive model checking

Simulation Results

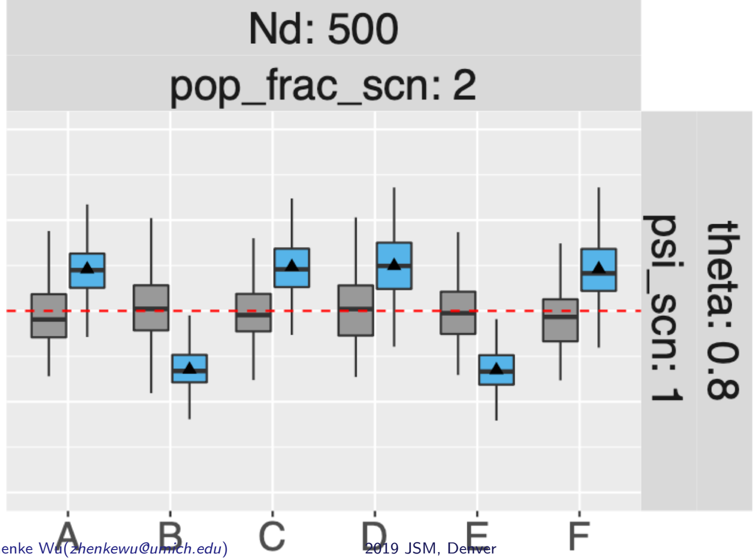
Simulation Results

- Simulation I: demonstrate flexible statistical inferences about the PEF functions $\{\pi_\ell(\cdot)\}$ (not shown in this talk; check paper)

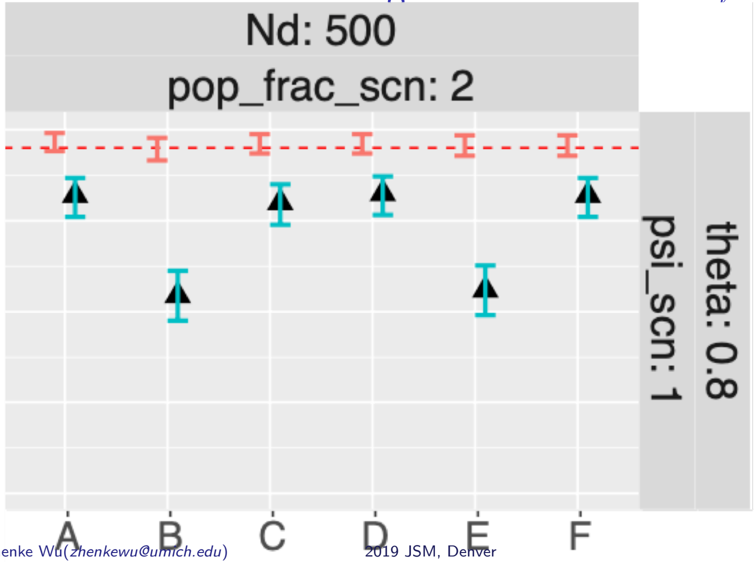
Simulation Results

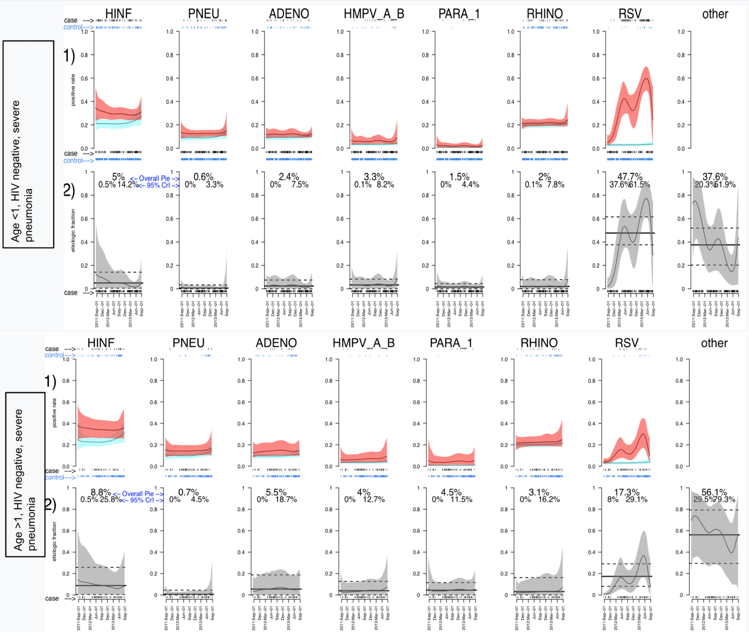
- Simulation I: demonstrate flexible statistical inferences about the PEF functions $\{\pi_\ell(\cdot)\}$ (not shown in this talk; check paper)
- Simulation II: determining overall PEF π_ℓ^* (empirical average) to quantify disease burdens in a population (potential policy interest)

Simulation II: Regression Model Reduces the Percent Relative Bias in Recovering the Overall PEFs π_ℓ^*



Simulation II: Regression Model Produces More Valid 95% CrIs in Recovering the Overall PEFs π^*_ℓ




Age <1, HIV negative, severe pneumonia
1)
2)
HINF
PNEU
ADENO
HMPV_A_B
PARA_1
RHINO
RSV
other
case -->
control-->
positive rate
etiologic fraction
5%
0.5% 14.2%
<- Overall Pie ->
<- 95% CrI ->
0.6%
0% 3.3%
2.4%
0% 7.5%
3.3%
0.1% 8.2%
1.5%
0% 4.4%
2%
0.1% 7.8%
47.7%
37.6% 61.5%
37.6%
20.3% 51.9%
Age >1, HIV negative, severe pneumonia
8.8%
0.5% 25.8%
0.7%
0% 4.5%
5.5%
0% 18.7%
4%
0% 12.7%
4.5%
0% 11.5%
3.1%
0% 16.2%
17.3%
8% 29.1%
56.1%
29.5% 79.3%
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2011:Sep-01
Dec-01
2012:Mar-01
Jun-01
Sep-01
Dec-01
2013:Mar-01
Jun-01
Sep-01

Main Points Once Again

Context: Modern large-scale etiology studies generate complex measurements of unobserved causes of disease, and have raised the analytic needs of estimating cause-specific case fractions, or "Population Etiologic Fractions" (PEFs)

Main Points Once Again

Context: Modern large-scale etiology studies generate complex measurements of unobserved causes of disease, and have raised the analytic needs of estimating cause-specific case fractions, or "Population Etiologic Fractions" (PEFs)
Gap: Despite recent methodological advances, the need of describing the relationship between covariates and PEFs, remains unmet

Main Points Once Again

Context: Modern large-scale etiology studies generate complex measurements of unobserved causes of disease, and have raised the analytic needs of estimating cause-specific case fractions, or "Population Etiologic Fractions" (PEFs)

Gap: Despite recent methodological advances, the need of describing the relationship between covariates and PEFs, remains unmet

Contribution: A general etiology regression framework building on npLCM that is broadly applicable to case-control studies

A general framework for a class of statistical problems that can be formulated as estimating covariate-dependent class-mixing weights.

Main Points Once Again

Three features of our approach:

Main Points Once Again

Three features of our approach:

- 1) allows analysts to specify a model that links important covariates to PEFs ☺

Main Points Once Again

Three features of our approach:

- 1) allows analysts to specify a model that links important covariates to PEFs ☺
- 2) produces covariate-dependent reference distribution for controls, which is critical for assigning cause-specific probabilities to a given case ☺
    - because we can compare control measurements to case measurements with similar covariate values

Main Points Once Again

Three features of our approach:

- 1) allows analysts to specify a model that links important covariates to PEFs ☺
- 2) produces covariate-dependent reference distribution for controls, which is critical for assigning cause-specific probabilities to a given case ☺
    - because we can compare control measurements to case measurements with similar covariate values
- 3) TPR priors are only used once; avoids overly-optimistic etiology uncertainty estimates. ☺

Thank You!

Student
Irena Chen

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

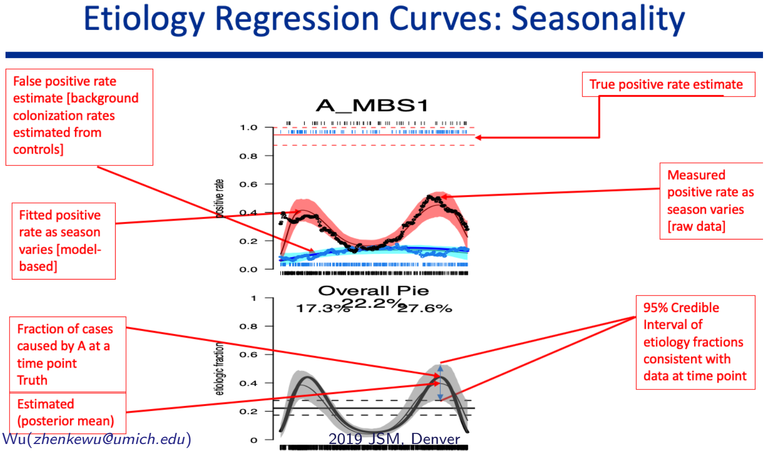
Funding
Patient-Centered Outcome Research Institute [PCORI ME-1408-20318]
Bill & Melinda Gates Foundation [48968]
Michigan Precision Health Investigator Award
National Cancer Institute (P30CA046592, U01CA229437)

Some References (More at: `zhenkewu.com`)

1. **Wu Z** and Chen I (2019+).
   Regression Analysis of Dependent Binary Data: Estimating Disease Etiology from Case-Control Studies.
   Submitted. https://arxiv.org/abs/1906.08436
2. PERCH Study Group (2019+).
   Causes of severe pneumonia re- quiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi- country case-control study.
   The Lancet. https://doi.org/10.1016/S0140-6736(19)30721-4
3. **Wu Z**, Deloria-Knoll M and Zeger SL (2019+).
   A Bayesian Approach to Restricted Latent Class Mod- els for Scientifically-Structured Clustering of Multivariate Binary Outcomes.
   Submitted. https://doi.org/10.1101/400192
4. **Wu Z**, Deloria-Knoll M and Zeger SL (2017).
   Nested Partially-Latent Class Models for Estimating Disease Etiology from Case-Control Data.
   Biostatistics. 18 (2): 200-213.
5. **Wu Z**, Deloria-Knoll M, Hammitt LL, and Zeger SL, for the PERCH Core Team (2015).
   Partially Latent Class Models (pLCM) for Case-Control Studies of Childhood Pneumonia Etiology.
   Journal of the Royal Statistical Society: Series C (Applied Statistics). 65:97-114.

Simulation I Results

- $N_d = 500$ cases and $N_u = 500$ controls for each of two levels of S (discrete covariate); Uniformly sample the subjects' enrollment dates over a period of 300 days.



Etiology Regression Curves: Seasonality

False positive rate estimate [background colonization rates estimated from controls]

True positive rate estimate

Measured positive rate as season varies [raw data]

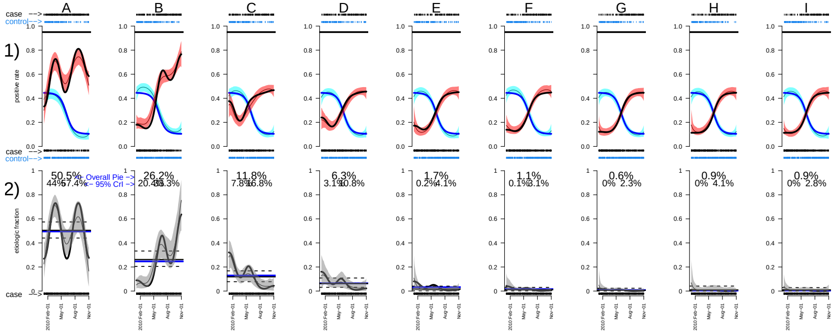
Fitted positive rate as season varies [model-based]

Fraction of cases caused by A at a time point Truth

Estimated (posterior mean)

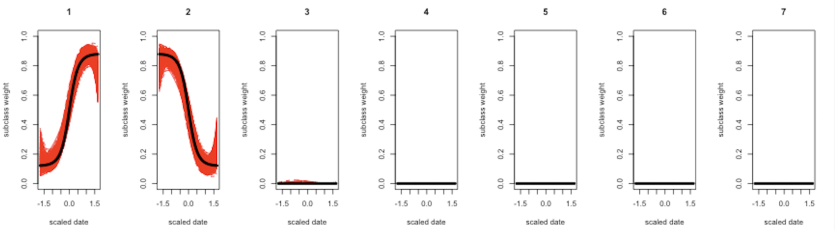
95% Credible Interval of etiology fractions consistent with data at time point

Simulation I: Recovery of Truth $\pi_\ell^0(t, S = s)$

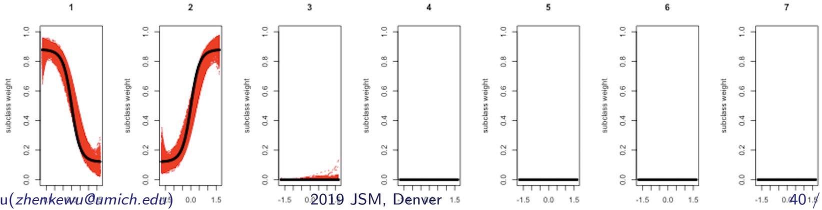


Simulation I: Recovery of $\nu_k(t)$ and $\eta_k(t)$

True $K^0 = 2$; Model fitted using a working number $K = 7$



(a) case



Appendix: Simulation II Setup

- npLCM regression analysis with $K^* = 3$, $R = 200$ replication data sets simulated under 48 different scenarios
- $L = J = 3, 6, 9$ causes, under single-pathogen-cause assumption, BrS measurements made on N_d cases and N_u controls for each level of X where $N_d = N_u = 250$ or 500.
- $\phi_\ell(X) = \beta_{0\ell} + \beta_{1\ell}\,\mathbb{I}\{X = 2\}$ take two sets of values to reflect PEF variability across X: i) $\boldsymbol{\beta}_0^{\text{i}} = (0, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_1^{\text{i}} = (-1.5, 0, -1.5, -1.5, 0, -1.5)$; ii) $\boldsymbol{\beta}_0^{\text{ii}} = (1, 0, 1, 1, 0, 1)$ and $\boldsymbol{\beta}_1^{\text{ii}} = (-1.5, 1, -1.5, -1.5, 1, -1.5)$
- TPRs $\theta_k^{(j)} = 0.95$ or 0.8 and FPRs $(\psi_1^{(j)}, \psi_2^{(j)}) \in \{(0.5, 0.05), (0.5, 0.15)\}$, for $j = 1, \ldots, J$.
- $\nu_k(W) = \eta_k(W) = logit^{-1}\left(\gamma_{k0} + \gamma_{k1}\,\mathbb{I}\{W = 2\}\right)$ where $(\gamma_{10}, \gamma_{11}) = (-0.5, 1.5)$ and $(\gamma_{20}, \gamma_{21}) = (1, -1.5)$.

Appendix

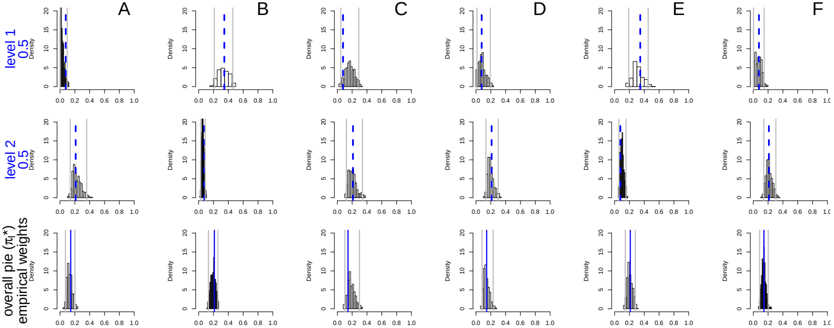


Figure: Posterior distributions of the stratum-specific (Row 1 and 2) and the overall (Bottom Row) PEFs based on a simulation with a two-level discrete covariate and $L = J = 6$ causes. The vertical gray lines indicate the 2.5% and 97.5% posterior quantiles, respectively; The truths are indicated by vertical blue dashed lines. *Row 1-2*) PEFs by stratum (level = 1,2) and cause (A-F); *Bottom*) π_ℓ^*: overall population etiologic fraction for cause A-F (empirical average of the two PEFs above).

Appendix

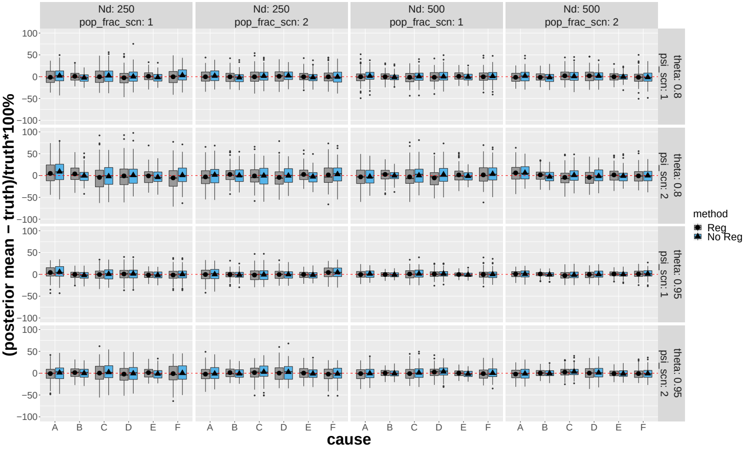
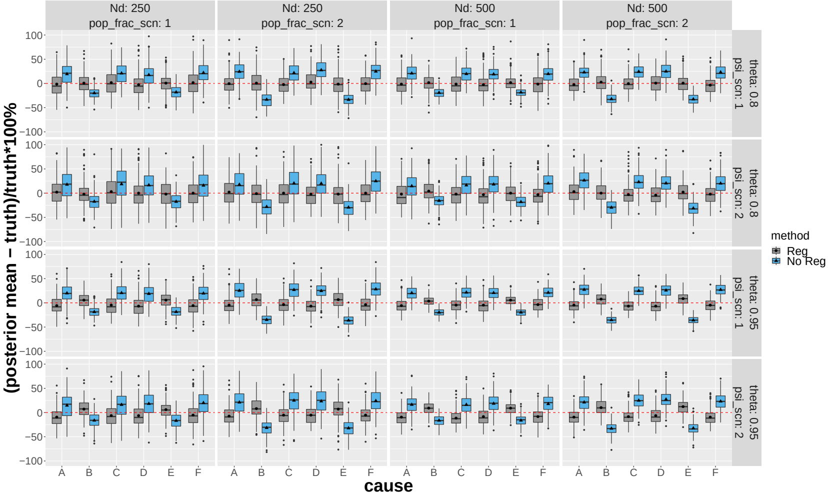
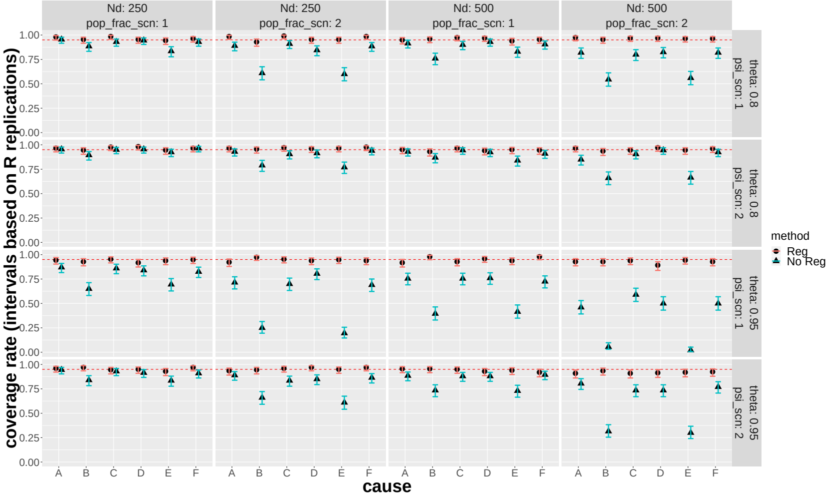


Figure: NPLCM analyses with or without regression perform similarly in terms of percent relative bias (top) and empirical coverage rates (bottom) over $R = 100$ replications in simulations where the case and control subclass weights *do not* vary by covariates. Each panel corresponds to one of 16 combinations of true parameter values and sample sizes.

Simulation II: Regression Model Reduces the Percent Relative Bias in Recovering the Overall PEFs π_ℓ^*



Simulation II: Regression Model Produces More Valid 95% CrIs in Recovering the Overall PEFs π^*_ℓ



Appendix

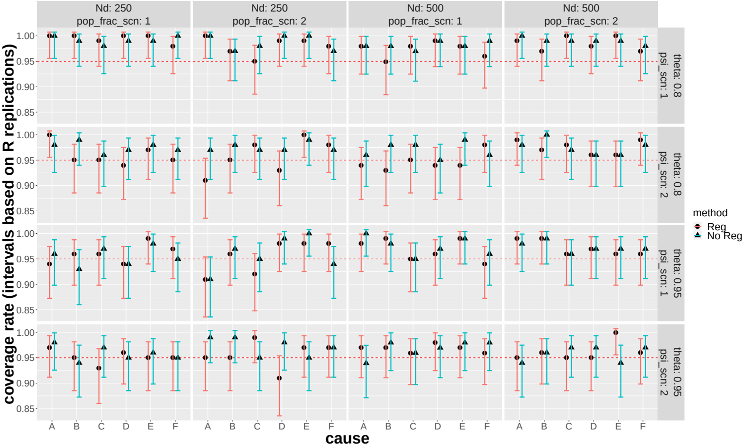
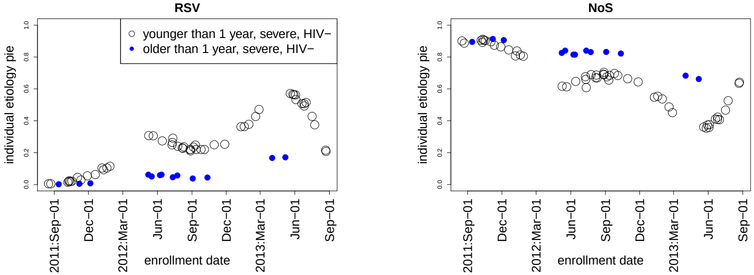


Figure: NPLCM analyses with or without regression perform similarly in terms of percent relative bias (top) and empirical coverage rates (bottom) over $R = 100$ replications in simulations where the case and control subclass weights *do not* vary by covariates. Each panel corresponds to one of 16 combinations of true parameter values and sample sizes.

Appendix



(a) Cause: RSV

(b) Cause: NoS

Figure: Individual etiology fraction estimates for RSV (left) and NoS (right) differ by age and season among HIV negative and severe pneumonia cases for whom the seven pathogens were *all tested negative* in the nasopharyngeal specimens.