Regression Analysis for Probabilistic Cause-of-Disease Assignment Using Case-Control Diagnostic Tests

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`

Motivating Application

Pneumonia Etiology Research for Child Health (PERCH) (PERCH Study Group, Lancet 2019, In Press)

Background:

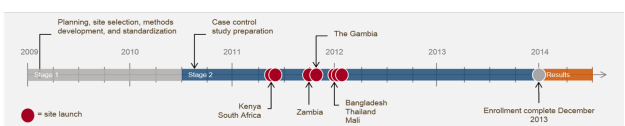
- $>$ 30 possible infectious causes
- Difficult to directly observe

Goal:

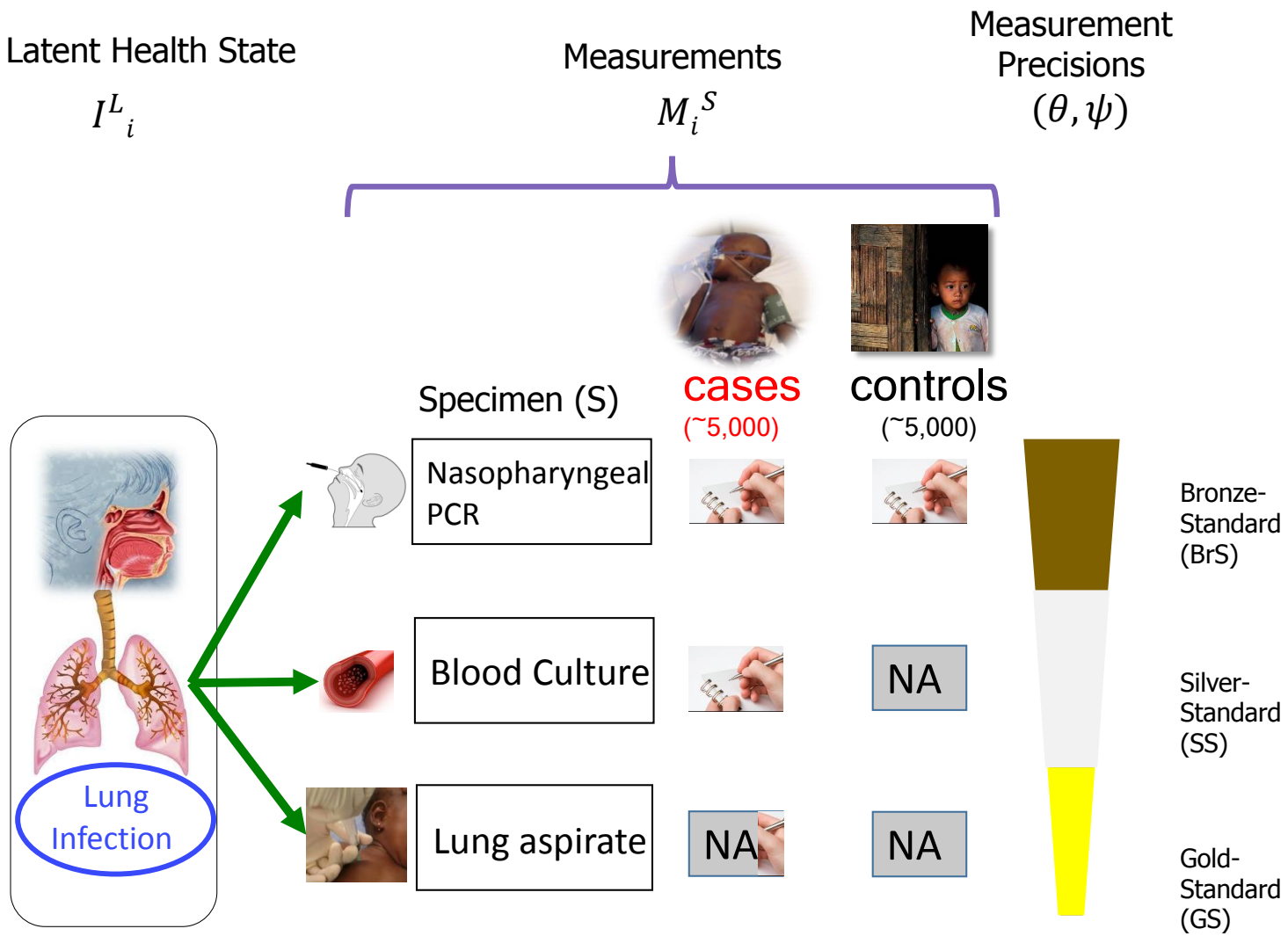
- Population disease etiology estimation
- Individual diagnosis

Study details:

- \$40-mil, Gates-funded 7-country study; Sites at Sub-Saharan Africa and South Asia
- Diverse measures; variable precisions
- $\sim$5,000 cases and $\sim$5,000 controls



Measurements of Different Quality



---

*NP: nasopharyngeal; PCR: polymerase chain reaction; LA: lung aspirate

Data From A Random Case

Measurement defined by (specimen+technology)

For example, "BCX" for **B**lood **C**ulture, "NPPCR" for **N**asal-**P**haryngeal **P**olymerase **C**hain **R**eaction

	BCX	PFCX	LACX	NPCX	ISCX2	PFPCR	LAPCR	NPPCR	ISPCR
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
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

Bacteria

Viruses

1: detected
0: absence
blank: missing

Problem and Data Features

Summary

Problem:

1. To infer individual latent health state
2. To estimate population distribution of latent health states (CSCFs)

Features:

- **case data**:
  1. Few or no gold-standard measure
  2. A large number of categories of latent health states
  3. Multiple sources of measurements of differential quality
- **extra control data** to integrate

No method has effectively estimated the etiologic distribution ("pie") using such data.

Previous Statistical Methods for Etiology Research

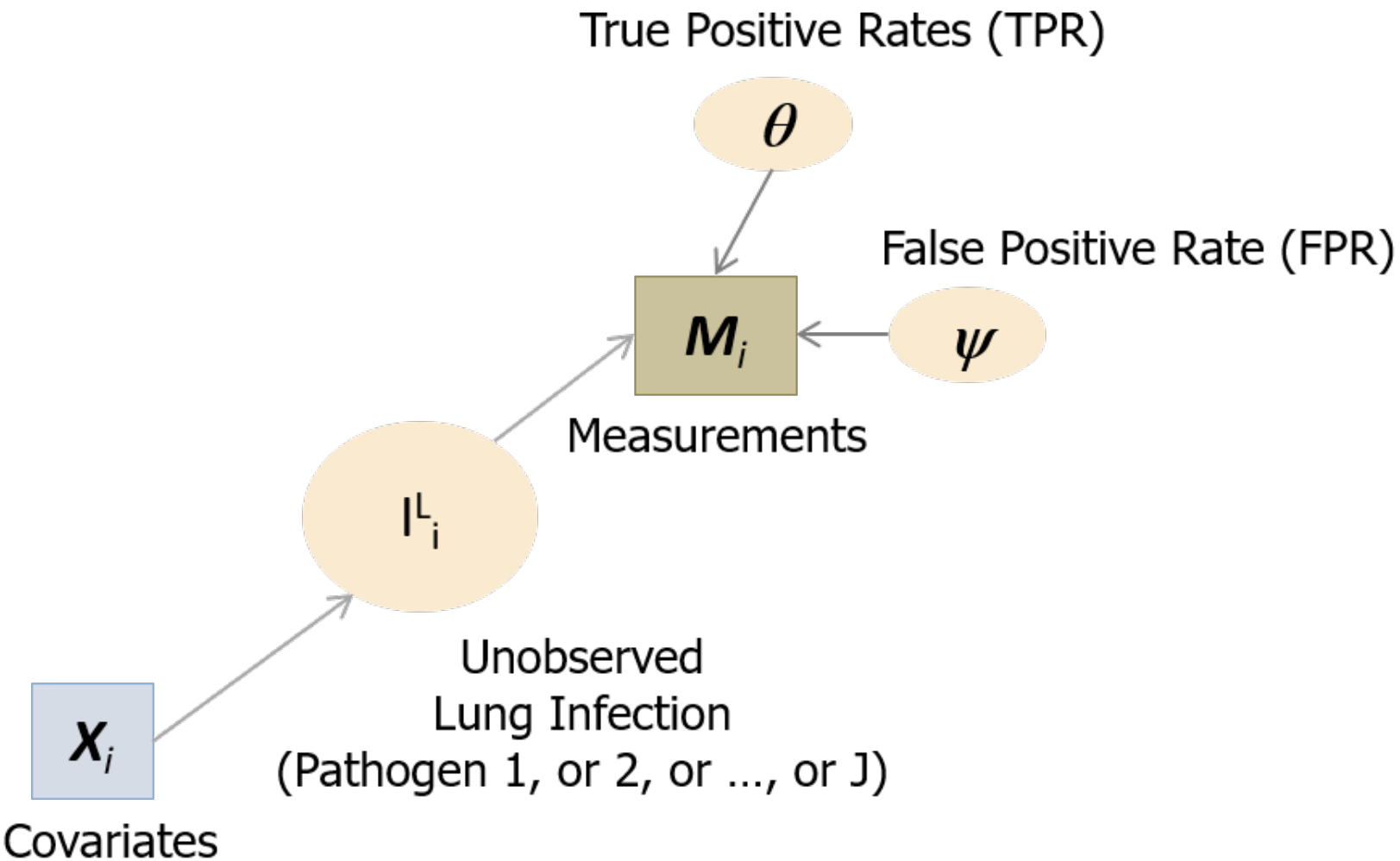
A Selected Review

- **Case-only, needs lots of GS data**: verbal autopsy methods for areas without medical death certification; Kernel smoothing for estimating sparse probability contingency table $Pr[\boldsymbol{M}^{\text{BrS}} \mid I]$ (King and Lu, 2008, *Stat. Sci.*)
- **Case-only, BrS data**: Bayesian nonparametric clustering (Hoff, 2004, *Biometrics*); Subset clustering (Friedman and Meulman, 2004, *JRSS-B*). Both no pre-defined cluster labels.
- **Case-control, only allows BrS data, assumes perfect test sensitivities**: Attributable fraction method (Bruzzi et al., 1985, AJE) based on logistic regression

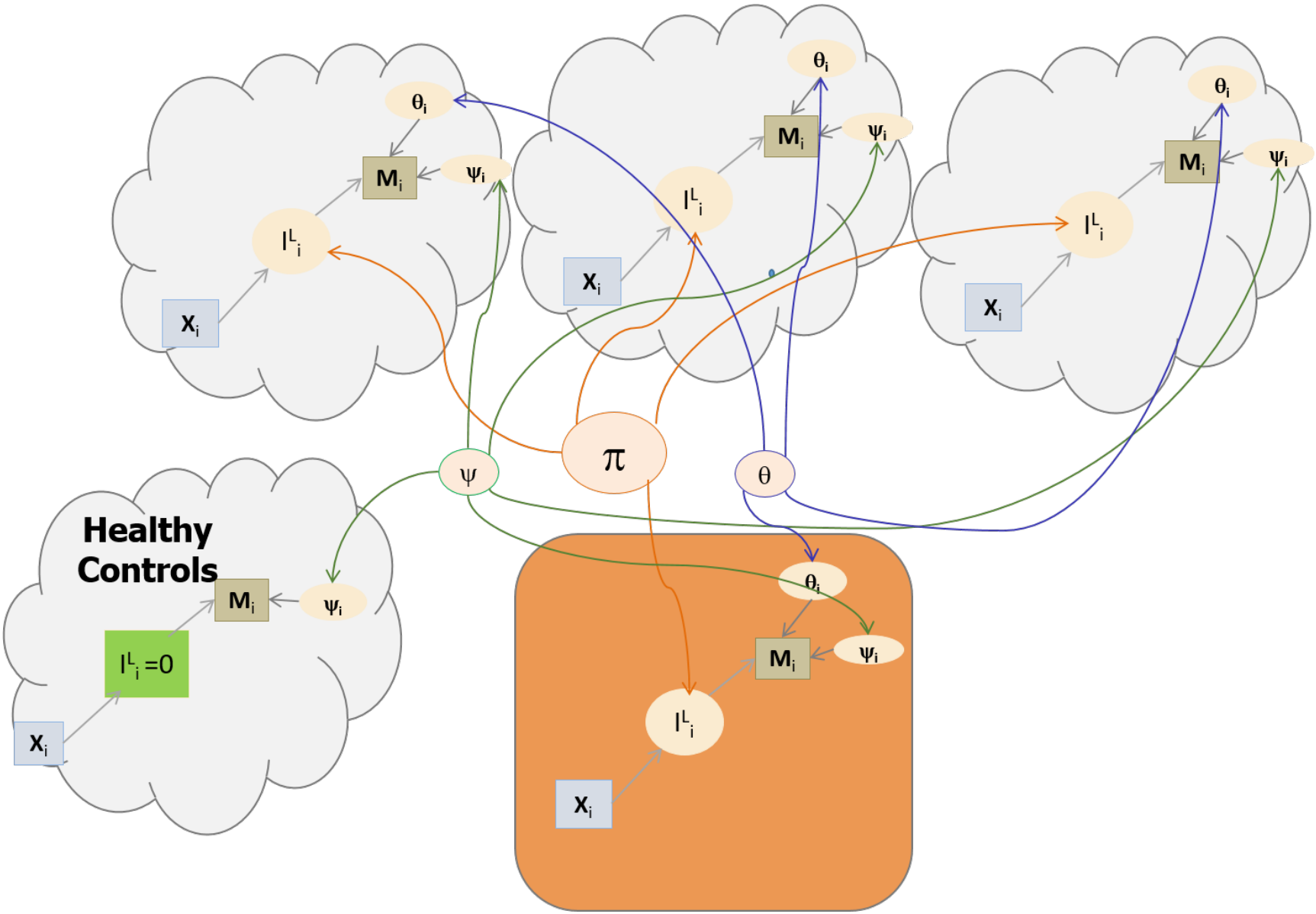
$$\text{logit } Pr[Y_i = 1 \mid \boldsymbol{M}_i^{\text{BrS}}, \boldsymbol{X}_i] = \sum_{j=1}^{J} \beta_j M_{ij}^{\text{BrS}} + \boldsymbol{X}_i' \gamma$$

Case Measurement Model

Joint Distribution of (Health State, Measurements)



Hierarchical Bayes Model for Etiology Research



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

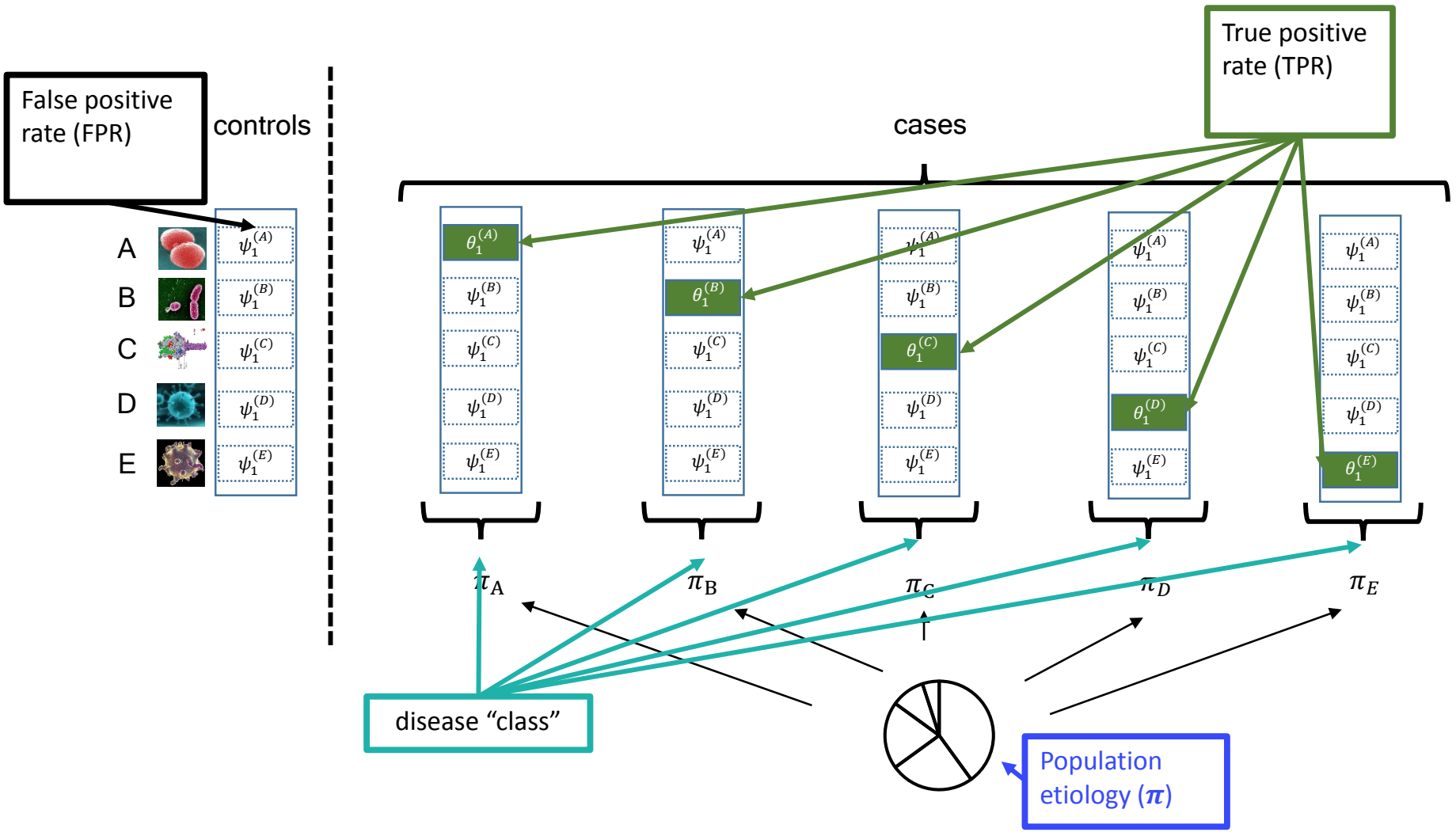
Notation

- $Y_i = \begin{cases} 0, \text{control} \\ 1, \text{case} \end{cases}$
- $I_i^L = \begin{cases} 0, \text{control} \\ 1, \text{pathogen 1} \\ ... \\ L, \text{pathogen } L \end{cases}$
- $\boldsymbol{M}_i^S = (M_{i1}^S, ..., M_{iJ_S}^S)'$ - Measurement vector
  - Specimen S on individual i
  - 1 for presence of pathogen from the test; 0 for absence

Partially-Latent Class Models (pLCM; Wu et al., 2016)

Model Structure (Bronze-Standard Data Only)

▸ partial identifiability ▸ statistical information



Assumptions

pLCM

- Non-interference assumptions for BrS data:

$$P(\boldsymbol{M}^{\text{BrS}}_{[-(j,j')]} \mid I^L = j, Y = 1) = P(\boldsymbol{M}^{\text{BrS}}_{[-(j,j')]} \mid I^L = j', Y = 1), \quad j, j' = 1, ..., J.$$

$$P(\boldsymbol{M}^{\text{BrS}}_{[-j]} \mid Y = 0) = P(\boldsymbol{M}^{\text{BrS}}_{[-j]} \mid I^L = j, Y = 1), \quad j = 1, ..., J$$

- Independence of measurements given class label (I^L_i)

Likelihood

pLCM

- Bronze-standard

$$
\begin{aligned}
P_i^{0,\mathrm{BrS}} &= \prod_{j=1}^{J}\left(\psi_j^{\mathrm{BrS}}\right)^{m_j}\left(1-\psi_j^{\mathrm{BrS}}\right)^{1-m_j}, \\
P_{i'}^{1,\mathrm{BrS}} &= \sum_{j=1}^{J}\pi_j\cdot\left(\theta_j^{\mathrm{BrS}}\right)^{m_j}\left(1-\theta_j^{\mathrm{BrS}}\right)^{1-m_j}\prod_{l\neq j}\left(\psi_l^{\mathrm{BrS}}\right)^{m_l}\left(1-\psi_l^{\mathrm{BrS}}\right)^{1-m_l}, \\
&\qquad\qquad \boldsymbol{m}=\boldsymbol{m}_{i'}^{\mathrm{BrS}}
\end{aligned}
$$

- Silver-standard

$$
P_{i'}^{1,\mathrm{SS}}=\Pr(\boldsymbol{M}_{i'}^{\mathrm{SS}}=\boldsymbol{m}|\boldsymbol{\pi},\boldsymbol{\theta}^{\mathrm{SS}})=\sum_{j=1}^{J'}\pi_j\cdot\left(\theta_j^{\mathrm{SS}}\right)^{m_j}\left(1-\theta_j^{\mathrm{SS}}\right)^{1-m_j}\mathbf{1}_{\left\{\sum_{l=1}^{J'}m_l\leq 1\right\}},\ \boldsymbol{m}=\boldsymbol{m}_{i'}^{\mathrm{SS}}
$$

- Gold-standard

$$
P_{i'}^{1,\mathrm{GS}}=\Pr\left(M_{i'}^{\mathrm{GS}}=\boldsymbol{m}|\boldsymbol{\pi}\right) = \prod_{j=1}^{J}\pi_j^{\mathbf{1}\left\{m_j=1\right\}}\mathbf{1}_{\{\sum_j m_j=1\}},\ \boldsymbol{m}=\boldsymbol{m}_{i'}^{\mathrm{GS}}\ .
$$

Partial Identifiablility

Necessity of Informative Priors on True Positive Rate

- pLCM implies: ◂ Model structure

$$\Pr\left[M_{ij}^{\text{BrS}} = 1\right] = \pi_j \theta_j^{\text{BrS}} + (1-\pi_j)\psi_j^{\text{BrS}}$$

- Formal argument: singular vectors and values of Jacobian matrix of model parametrization
- Bayesian framework sidesteps partial identifiability problem
  - Use TPR prior elicited from laboratory scientists (Cf. Wu et al., 2015, *JRSS-C*)
  - **No Bayesian free lunch**: posterior of unidentified parameters not shrinking to point mass as sample size grows
  - Identified set of parameter values; Valuable in epidemiology, econometrics, sociology (Cf. Greenland, 2005, *JRSS-A*; Gustafson, 2009, *JASA*; Gustafson, 2005, *Stat. Sci.*; Manski, 2010, *PNAS*)

Priors

pLCM

- Informative
  - $\theta_j^{\text{BrS}} \sim \text{Beta}(c_{1j}, c_{2j})$ - true positive rates for BrS data
  - $\theta_j^{\text{SS}} \sim \text{Beta}(d_{1j}, d_{2j})$ - true positive rates for SS data
- Non-informative
  - $\boldsymbol{\pi} \sim \text{Dirichlet}(0.5, ..., 0.5)$ - population etiology
  - $\psi_j^{\text{BrS}} \sim \text{Beta}(1, 1)$ - false positive rates for BrS data

Joint prior for $\boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\psi}^{\text{BrS}}, \boldsymbol{\theta}^{\text{BrS}}, \boldsymbol{\theta}^{SS})'$, *a priori* independent:

$$[\boldsymbol{\gamma}] = [\boldsymbol{\pi}][\boldsymbol{\psi}^{\text{BrS}}][\boldsymbol{\theta}^{\text{BrS}}][\boldsymbol{\theta}^{SS}]$$

Posterior Computing

- Gibbs sampler: construct correlated samples to approximate the shape of joint posterior distribution of the unknowns
- Unknowns:
  - π-population etiology distribution
  - $(\psi^{\text{BrS}}, \theta^{\text{BrS}})'$- TPRs and FPRs for BrS measurements
  - θ^{SS} - TPRs for SS measurements
  - **I_i^L-latent health state; for case i**
- Individual diagnosis: For a case with new measurements $\boldsymbol{m}_*$, approximate by

$$\Pr(I_i^L = j \mid \boldsymbol{m}_*, \mathcal{D}) = \int \Pr(I_i^L = j \mid \boldsymbol{m}_*, \gamma) \Pr(\gamma \mid \boldsymbol{m}_*, \mathcal{D}) \mathrm{d}\gamma,$$

$$j = 1, \ldots J$$

Information for Correct Individual Diagnosis

- Log relative probability of $I_i^L = j$ versus $I_i^L = \ell$ given others is

$$R_{j\ell} = \log\left(\frac{\pi_j}{\pi_\ell}\right) + \log\left\{\left(\frac{\theta_j^{\text{BrS}}}{\psi_j^{\text{BrS}}}\right)^{m_{*j}}\left(\frac{1-\theta_j^{\text{BrS}}}{1-\psi_j^{\text{BrS}}}\right)^{1-m_{*j}}\right\} + \log\left\{\left(\frac{\psi_\ell^{\text{BrS}}}{\theta_\ell^{\text{BrS}}}\right)^{m_{*\ell}}\left(\frac{1-\psi_\ell^{\text{BrS}}}{1-\theta_\ell^{\text{BrS}}}\right)^{1-m_{*\ell}}\right\}$$

- Suppose $I_i^L = j$. Averaging over $\boldsymbol{m}_*$:

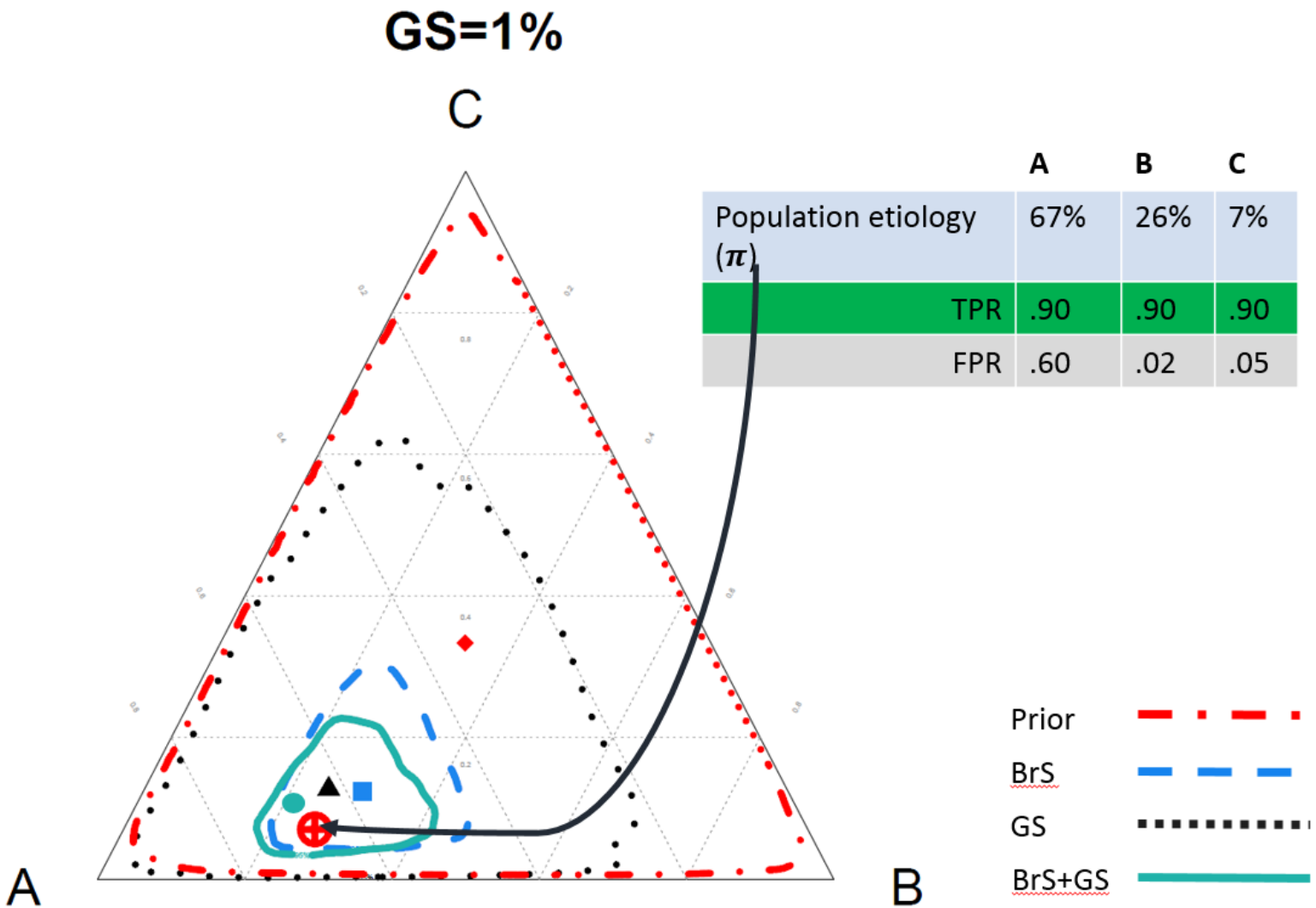
$$E[R_{j\ell}] = \log\left(\pi_j/\pi_\ell\right) + \underbrace{I(\theta_j^{\text{BrS}}\,;\psi_j^{\text{BrS}}) + I(\psi_\ell^{\text{BrS}}\,;\theta_\ell^{\text{BrS}})}_{\textit{large \& positive if the arguments are discrepant}}$$

◂ Model structure

- $I(v_1; v_2)$: expected amount of information in $m_{*j} \sim \text{Bern}(v_1)$ for discriminating against $m_{*j} \sim \text{Bern}(v_2)$.

Inference with BrS+GS Data

Simulation: 3 Pathogens; 500 Cases/Controls; 5 Cases with GS Measure



	A	B	C
Population etiology (π)	67%	26%	7%
TPR	.90	.90	.90
FPR	.60	.02	.05

"nested" pLCM

Relax the LI and Non-interference Assumption

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

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence against LI: 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 against LI: 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 against LI: 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

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

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence against LI: 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

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

  - Log-linear parameterization

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence against LI: 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

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

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

"nested" pLCM

Relax the LI and Non-interference Assumption

- Direct evidence against LI: 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

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

  - Log-linear parameterization
  - 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

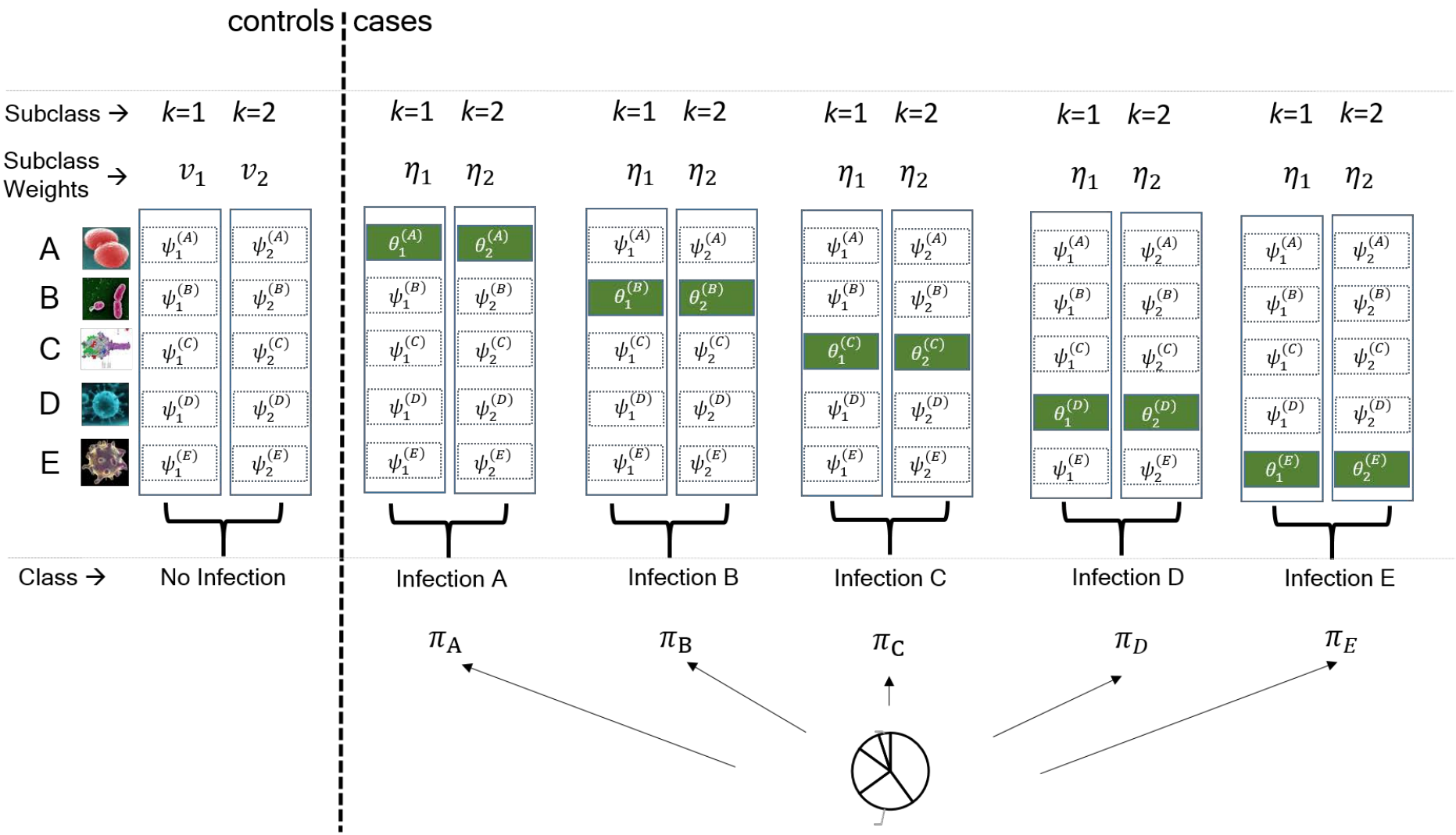
- Direct evidence against LI: control measurements $(M_{i1}, ..., M_{iJ})'$
  - test cross-reactions (prevented in PERCH assays)
  - lab technicians effect
  - heterogeneity in subjects' immunity level
- Deviations from independence impacts inference (Cf. Pepe and Janes, 2007, *Biostatistics*; Albert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

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

  - Log-linear parameterization
  - 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*)

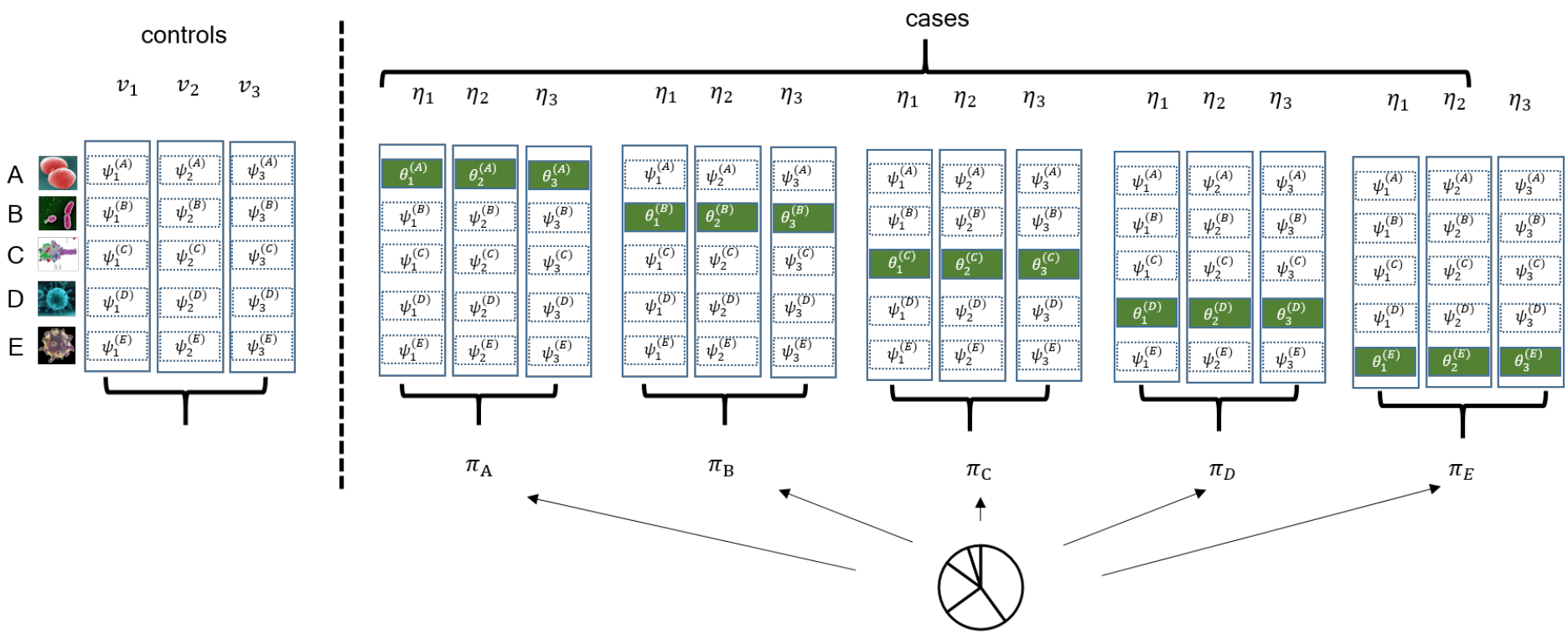
Nested Partially-Latent Class Models (npLCM; Wu and Zeger, 2016)

Example: 5 Pathogens, 2 Subclasses; BrS Data Only



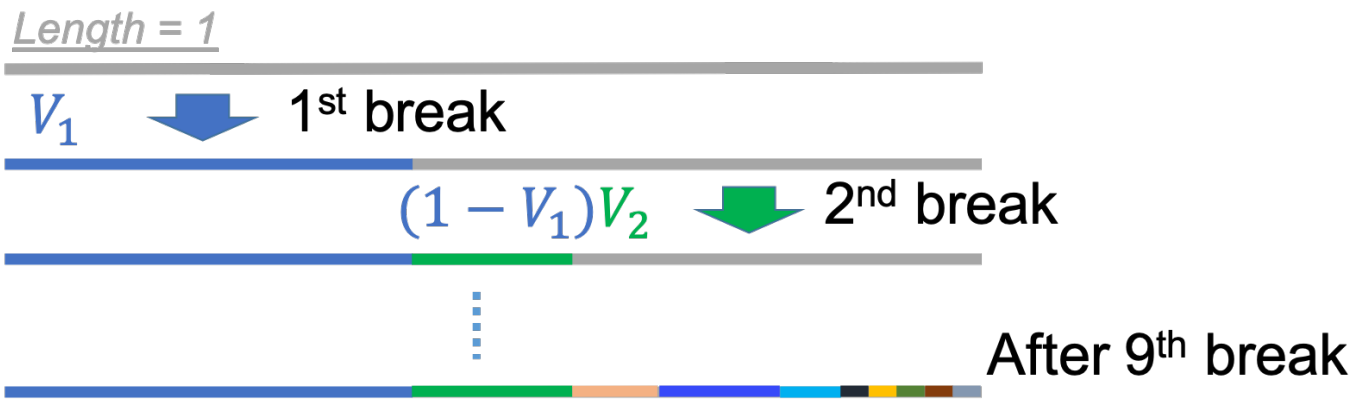
Nested Partially-Latent Class Models (npLCM; Wu and Zeger, 2016)

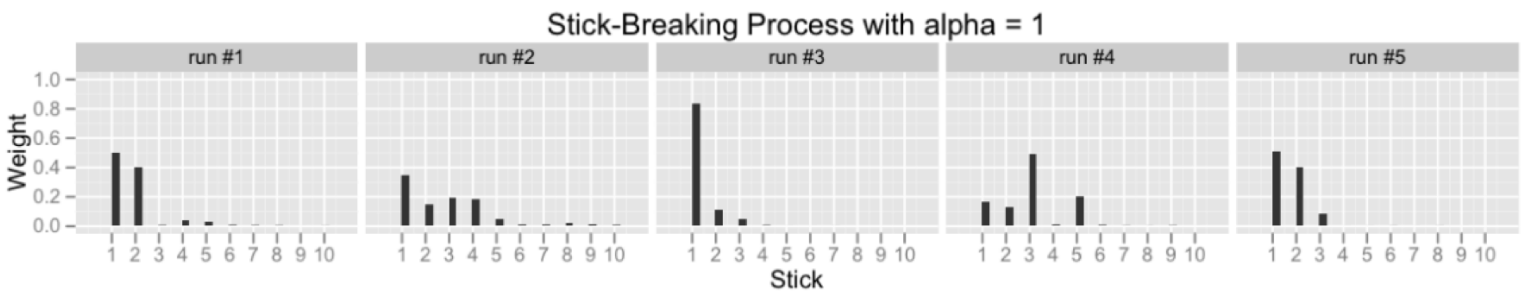
Example: 5 Pathogens, 3 Subclasses; BrS Data Only



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

$$\boldsymbol{P}_0(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{k=1}^{K} \nu_k \prod_{j=1}^{J} \left\{\psi_k^{(j)}\right\}^{m_j} \left\{1-\psi_k^{(j)}\right\}^{1-m_j},$$

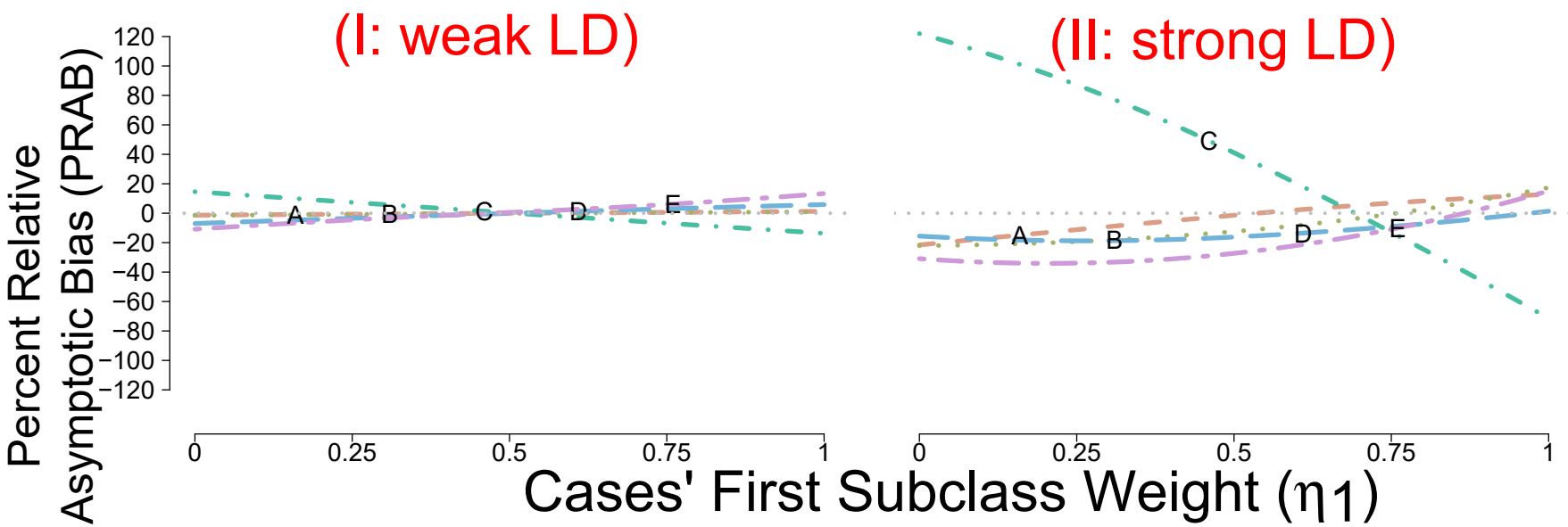
$$\boldsymbol{P}_1(\boldsymbol{M}_i = \boldsymbol{m}) = \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K} \left[\eta_k \left\{\theta_k^{(j)}\right\}^{m_j} \left\{1-\theta_k^{(j)}\right\}^{1-m_j} \prod_{\ell\neq j} \left\{\psi_k^{(j)}\right\}^{m_\ell} \left\{1-\psi_k^{(j)}\right\}^{1-m_\ell}\right],$$

- Prior:

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

Estimation Bias if Ignoring Local Dependence (LD)

Simulation: LD Truth (npLCM) Estimated by Working LI Models (pLCM)



Nested Partially Latent Class Models (npLCM)

For simplicity, we assume "single-pathogen causes", or a single relevant feature per cluster, or more visually, "one row of green boxes per disease class"

npLCM Framework (no Covariates)

Three components of a likelihood function:

npLCM Framework (no Covariates)

Three components of a likelihood function:

npLCM Framework (no Covariates)

Three components of a likelihood function:

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

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

npLCM Framework (no Covariates)

Three components of a likelihood function:

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

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

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 a likelihood function:

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

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

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 a likelihood function:

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

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

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 a likelihood function:

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

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

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 $\{\boldsymbol{M}_i : \textit{where } Y_i = 0\}$:

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
     $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\psi}_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$.

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
     $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\psi}_{[-\ell]})$, 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 $\boldsymbol{\psi}$.

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
     $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\psi}_{[-\ell]})$, 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 $\boldsymbol{\psi}$.

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

Regression Analysis in nested PLCM

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

- **Data**: case-control diagnostic tests, multivariate binary observations
- **Scientific problem**: estimate cause-specific case fractions (CSCF); Think "Pie chart" for cases

Regression Analysis in nested PLCM

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

- **Data**: case-control diagnostic tests, multivariate binary observations
- **Scientific problem**: estimate cause-specific case fractions (CSCF); Think "Pie chart" for cases
- **Statistical problem**: Using nested PLCM to estimate the mixing distribution among the cases

Regression Analysis in nested PLCM

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

- **Data**: case-control diagnostic tests, multivariate binary observations
- **Scientific problem**: estimate cause-specific case fractions (CSCF); Think "Pie chart" for cases
- **Statistical problem**: Using nested PLCM to estimate the mixing distribution among the cases
- **Motivation for regression analyses**: CSCFs 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
- $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$: shared by cases and controls; 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 $\boldsymbol{X}_i$).

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
- $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$: shared by cases and controls; 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 $\boldsymbol{X}_i$).
- Continuous covariates: the first p_1 and q_1 elements of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$, respectively.

Motivating Application Again: PERCH Study

Data : 494 cases and 944 controls from one site

Motivating Application Again: PERCH Study

Data : 494 cases and 944 controls from one site

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

Motivating Application Again: PERCH Study

Data : 494 cases and 944 controls from one site

Goal a. : Estimate CSCFs 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 CSCFs $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ as an empirical average of the stratum-specific CSCFs (by $\boldsymbol{X}$); Of policy interest (vaccine/antibiotics development and manufacture)

Motivating Application Again: PERCH Study

Data : 494 cases and 944 controls from one site

Goal a. : Estimate CSCFs 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 CSCFs $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ as an empirical average of the stratum-specific CSCFs (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) (case-only)	HIV positive	# cases (%) total: 524 (100)	# controls (%) 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)	-
1	1	0	33 (6.3)	-
0	0	1	37 (7.1)	85 (8.8)
1	0	1	24 (4.5)	51 (5.3)
0	1	1	25 (4.8)	-
1	1	1	3 (0.6)	-
case: 25.2%	34.5%	17.0%		
control: 34.3%	-	14.1%		

Current Methods Fall Short☹

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

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM 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 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 CSCF estimates due to sparsely-populated strata.

Current Methods Fall Short☹

- *Fully-stratified analysis*: fit an npLCM 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 CSCF 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.

The Rest of Talk☺

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

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

The Rest of 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 CSCFs ($\boldsymbol{\pi}$),

The Rest of 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 CSCFs (π),

(b) specifies parsimonious functional dependence of π upon covariates such as additivity, and

The Rest of 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 CSCFs (π),

(b) specifies parsimonious functional dependence of π upon covariates such as additivity, and

(c) correctly assesses the posterior uncertainty of the CSCF functions and the overall CSCFs π^* by applying the TPR priors just once.

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 $\boldsymbol{\pi_\ell(X)}$

$\pi_\ell(\boldsymbol{X})$ 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 $\boldsymbol{\pi_\ell(X)}$

$\pi_\ell(\boldsymbol{X})$ 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 $\pi_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ 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 $\pi_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ 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 $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi$

Etiology Regression $\pi_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ 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 $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi$

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

Etiology Regression $\pi_\ell(\boldsymbol{X})$

$\pi_\ell(\boldsymbol{X})$ 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 $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi$

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 $\boldsymbol{x}$; $\boldsymbol{\Gamma}_\ell^\pi = (\boldsymbol{\beta}_{\ell j}^\pi, \boldsymbol{\gamma}_\ell^\pi)$.

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:

- 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 P_0 depend on W_i

Regression model for controls

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

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

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
  - $\Pi(\boldsymbol{m}; s) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}} (1 - s_j)^{1 - m_{ij}}$

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:

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:

sample subclass indicator : $Z_i \mid \boldsymbol{W}_i \sim \text{Categorical}_K(\boldsymbol{\nu}(\boldsymbol{W}_i))$

generate measurements : $M_{ij} \mid Z_i = k \sim \text{Bernoulli}(\psi_k^{(j)})$,
independently for $j = 1, ..., J$.

Let P_0 depend on 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

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

Let P_0 depend on 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

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

$g(\cdot) = 1/(1 + \exp\{-(\cdot)\})$

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

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

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

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

Let P_0 depend on 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

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

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

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

Expand the smooth functions by B-spline bases with coefficients β_{kj}^ν; $\widetilde{\boldsymbol{w}}$ 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) "Uniform Shrinkage over Simplex" for $\nu_k(\boldsymbol{W})$

Proposed Model

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

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} \mu_{j0}^*$, $\mu_{j0}^* > 0$. A large μ_{k0} for a large k.

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

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} \mu_{j0}^*$, $\mu_{j0}^* > 0$. A large μ_{k0} for a large k.
- μ_{k0} increases with k: making the stick-breaking *a priori* more likely to stop for a large k

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

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} \mu_{j0}^*$, $\mu_{j0}^* > 0$. A large μ_{k0} for a large k.
- μ_{k0} increases with k: making the stick-breaking *a priori* more likely to stop for a large k
- We specify the prior distributions for μ_{j0}^* to be heavy-tailed:

$$\mu_{j0}^* \sim \mathit{Cauchy}^+(0, s_j),\ j = 1, \ldots, K,$$

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

Proposed Model

- We let $\mu_{k0} = \sum_{j=1}^{k} \mu_{j0}^*$, $\mu_{j0}^* > 0$. A large μ_{k0} for a large k.
- μ_{k0} increases with k: making the stick-breaking *a priori* more likely to stop for a large k
- We specify the prior distributions for μ_{j0}^* to be heavy-tailed:

$$\mu_{j0}^* \sim Cauchy^+(0, s_j),\ j = 1, \ldots, K,$$

- A large s_k produces a large μ_{k0}^* and helps stop the stick-breaking at class k.

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

Proposed Model

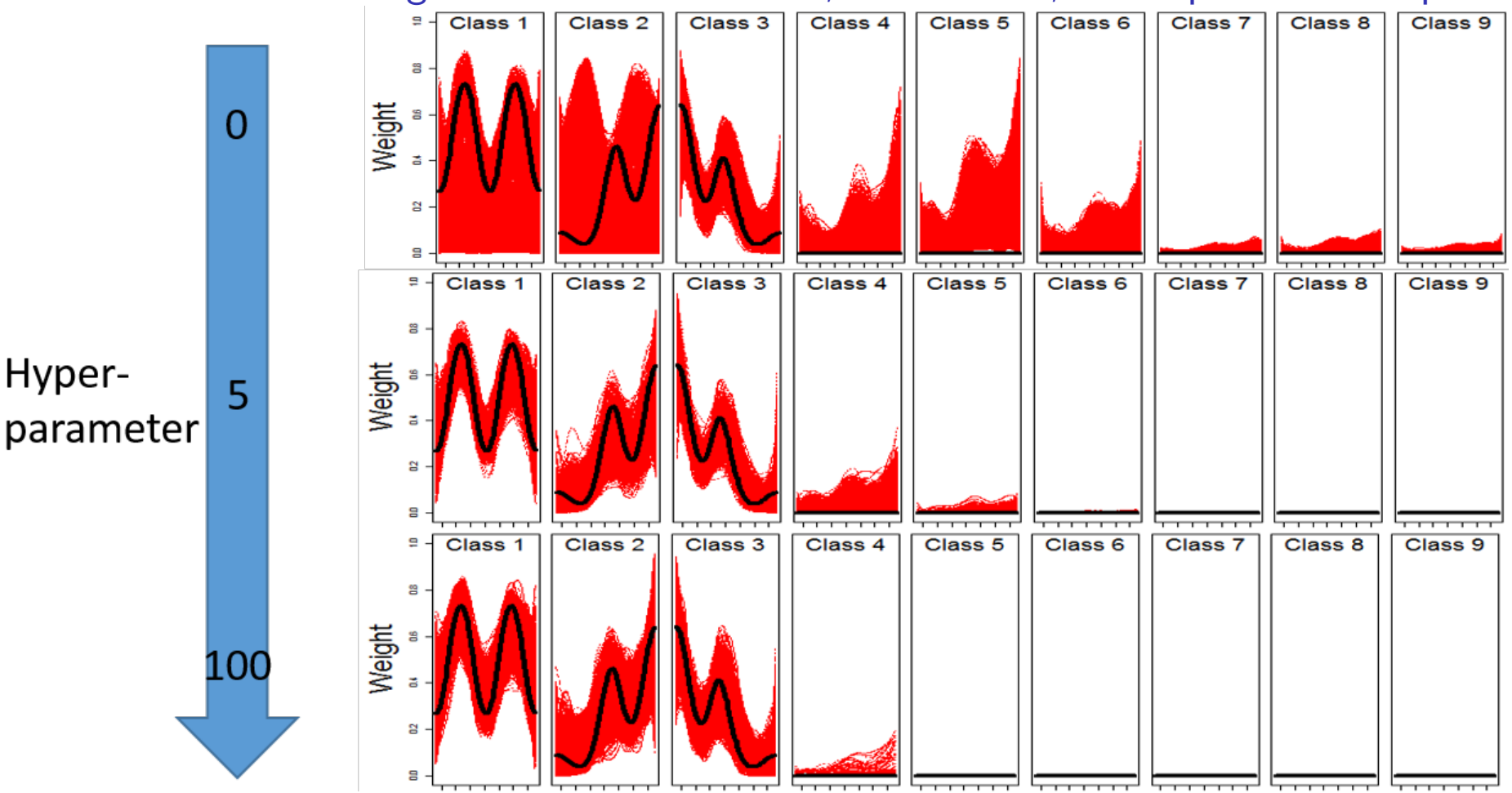
- We let $\mu_{k0} = \sum_{j=1}^{k} \mu_{j0}^*$, $\mu_{j0}^* > 0$. A large μ_{k0} for a large k.
- μ_{k0} increases with k: making the stick-breaking *a priori* more likely to stop for a large k
- We specify the prior distributions for μ_{j0}^* to be heavy-tailed:

$$\mu_{j0}^* \sim Cauchy^+(0, s_j),\ j = 1, \ldots, K,$$

- A large s_k produces a large μ_{k0}^* and helps stop the stick-breaking at class k.
- 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_j

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



X-axis: covariate values

Y-axis: weight; 0 to 1.

Let P_1 depend on X and W

Subclass Weight Regression: For Cases

Let P_1 depend on X and W

Subclass Weight Regression: For Cases

The pmf for cases' measurements:

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

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 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:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- Equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise

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:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- 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 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:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- 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 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:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- 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 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:
  $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$
- 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:
  $L_0^{\text{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$.

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:

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

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

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

- $\nu_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^\nu)$: The S????-B???? 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 ($\{\mathbf{\Gamma}_\ell^\pi\}$),
- subclass mixing weight parameters for cases ($\{\mathbf{\Gamma}_k^\eta\}$) and controls ($\{\mathbf{\Gamma}_k^\nu\}$),
- true and false positive rates ($\mathbf{\Theta} = \{\theta_k^{(j)}\}, \mathbf{\Psi} = \{\psi_k^{(j)}\}$).

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

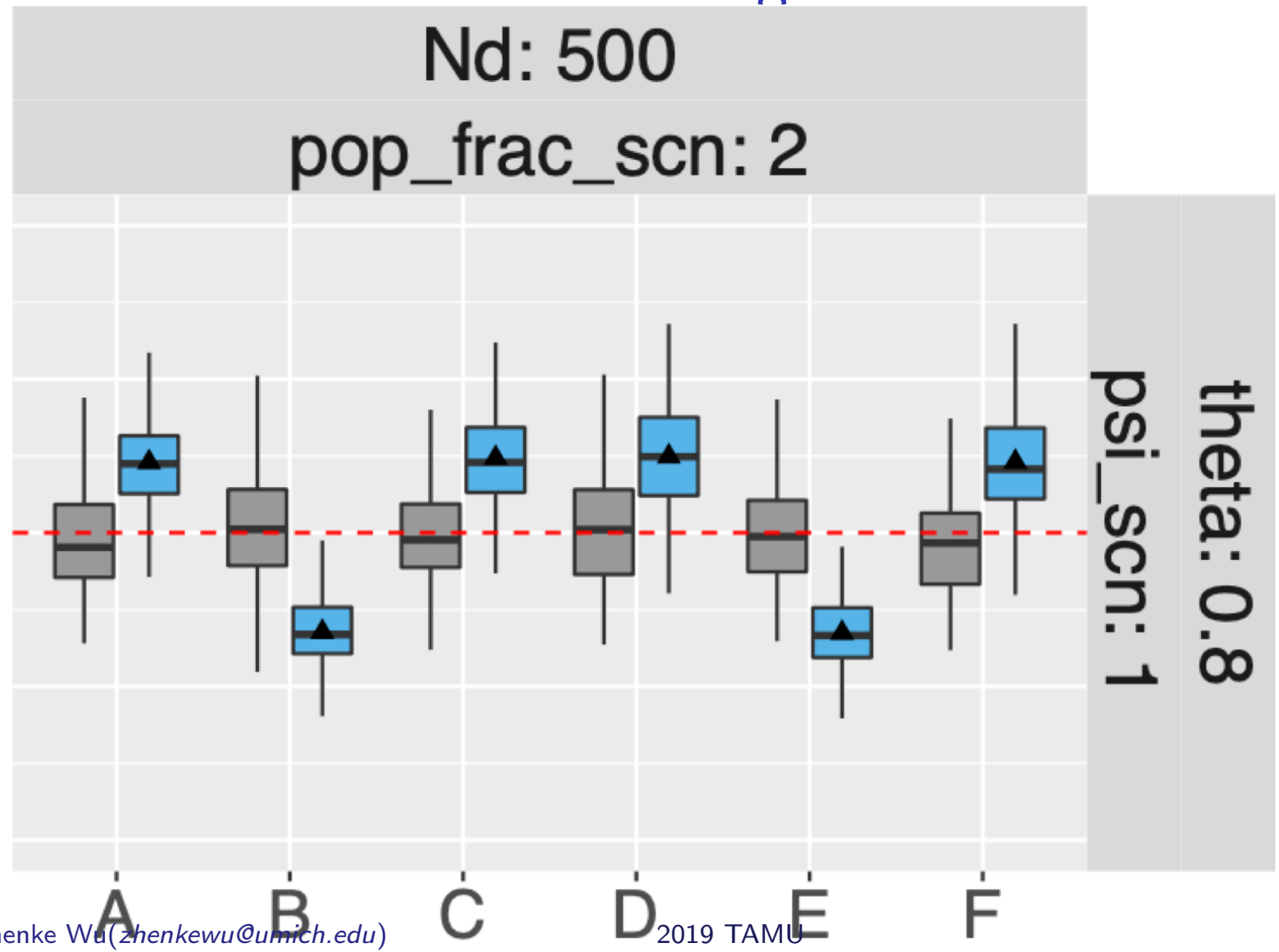
Simulation Results

- Simulation I: flexible and valid statistical inferences about the CSCF functions $\{\pi_\ell(\cdot)\}$ (not shown here)

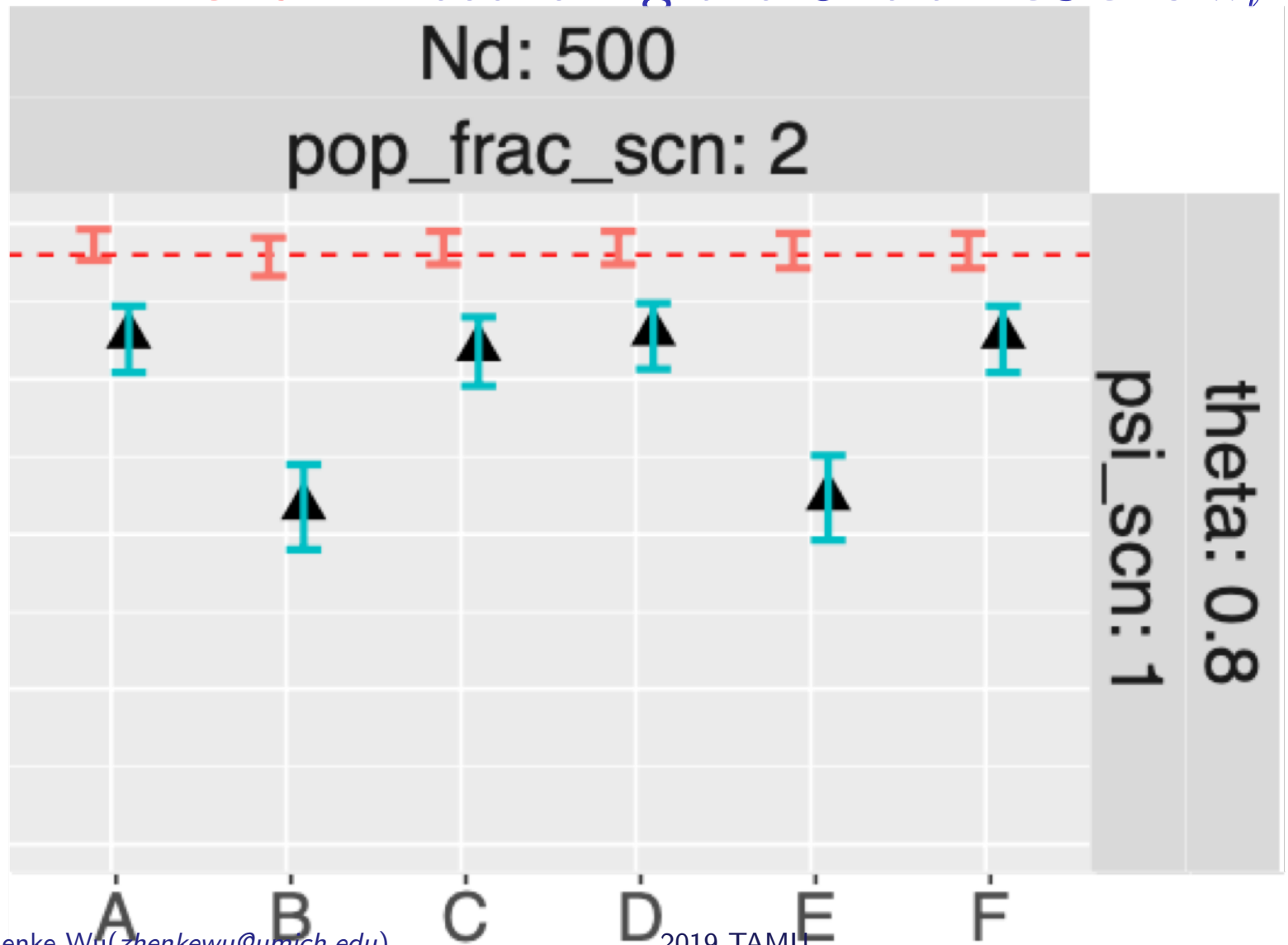
Simulation Results

- Simulation I: flexible and valid statistical inferences about the CSCF functions $\{\pi_\ell(\cdot)\}$ (not shown here)
- Simulation II: valid inferences about the overall CSCF π^*_ℓ (empirical average) to quantify disease burdens in a population (of policy interest)

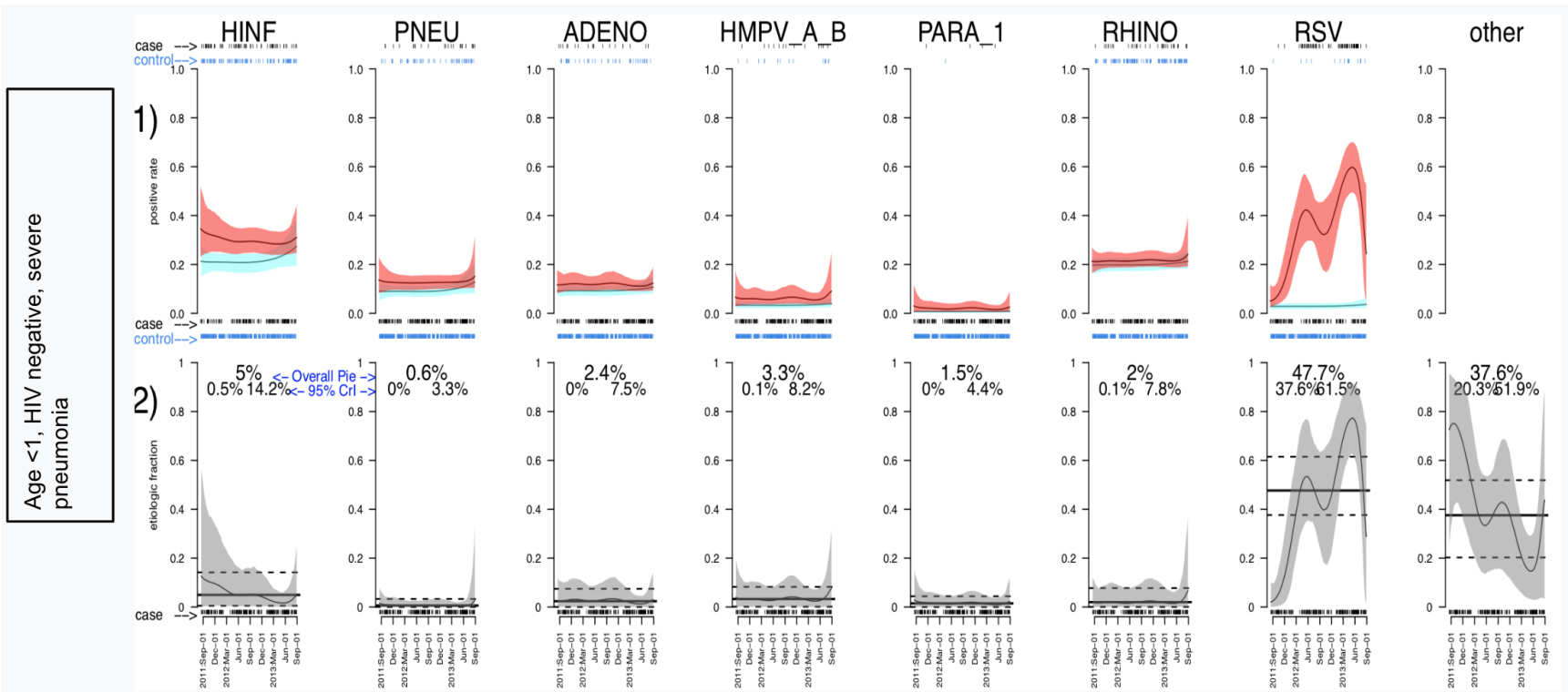
Simulation II: Regression Model Reduces the Percent Relative Bias in Recovering the Overall CSCFs π_ℓ^*



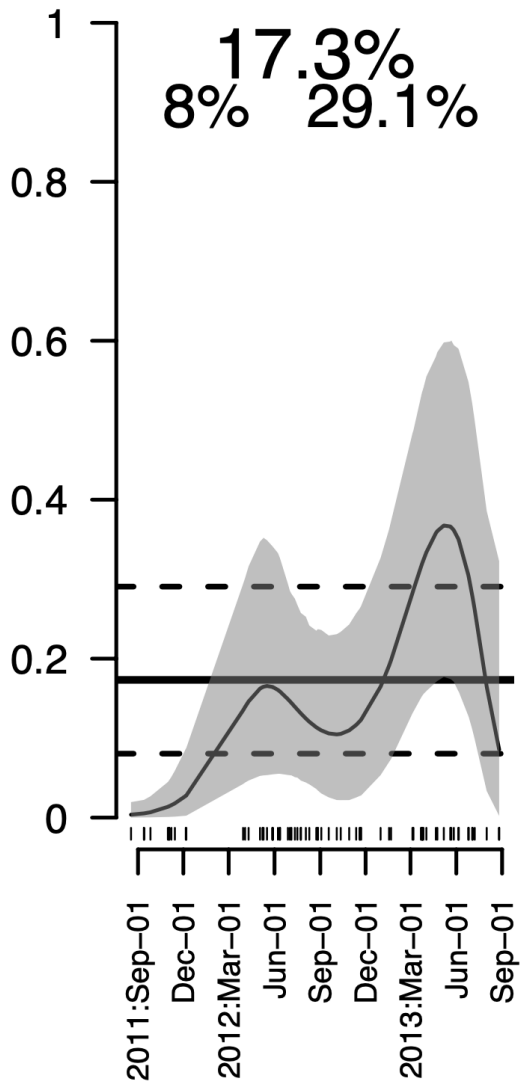
Simulation II: Regression Model Produces More Valid 95% CrIs in Recovering the Overall CSCFs π_ℓ^*



Regression analysis of PERCH data from one site: Age<1, Severe Pneumonia, HIV negative



Seasonal Trend for π_{RSV}: Age<1, Severe Pneumonia, HIV negative



Summary of the Regression Approach

Summary of the Regression Approach

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

Summary of the Regression Approach

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

Summary of the Regression Approach

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

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 (CSCFs)

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 (CSCFs)
Gap: Despite recent methodological advances, the need of describing the relationship between covariates and CSCFs, 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 (CSCFs)

Gap: Despite recent methodological advances, the need of describing the relationship between covariates and CSCFs, 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.

Discussions

- Related to restricted latent class models (RLCM, Xu, 2017, AOS; Wu 2019);
- "Restricted" means the response probability for a measurement depends on the latent state in a monotonic way (e.g., we have TPR greater than FPR in the pneumonia example)
- Established sufficient and necessary conditions for theoretical identifibility (based on likelihood only).
- Also related to boolean matrix decomposition (Rukat 2017, ICML) and double feature allocation (Ni and Mueller, 2019, JASA)
- Other applications in autoimmune disease subsetting (Wu et al, 2019, Biostatistics) and electronic health records (Ni and Mueller, 2019) and verbal autopsy (King and Lu, 2008 Stat Sci; McCormick et al., 2016, JASA)

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

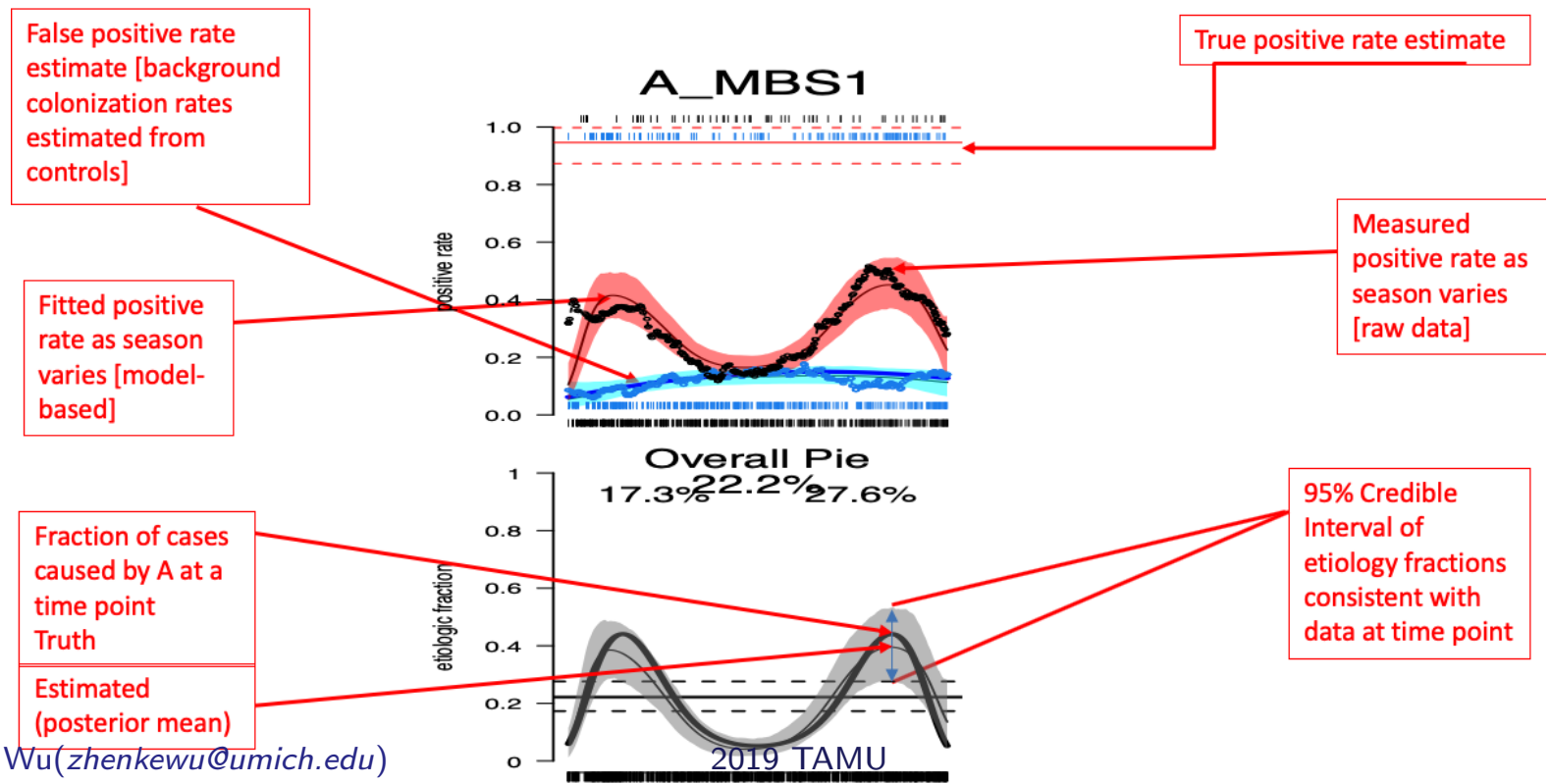
4. **Wu Z**, Deloria-Knoll M and Zeger SL (2017).
Nested Partially-Latent Class Models for Estimating Disease Etiology from Case-Control Data.
Biostatistics. 18 (2): 200-213.

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.

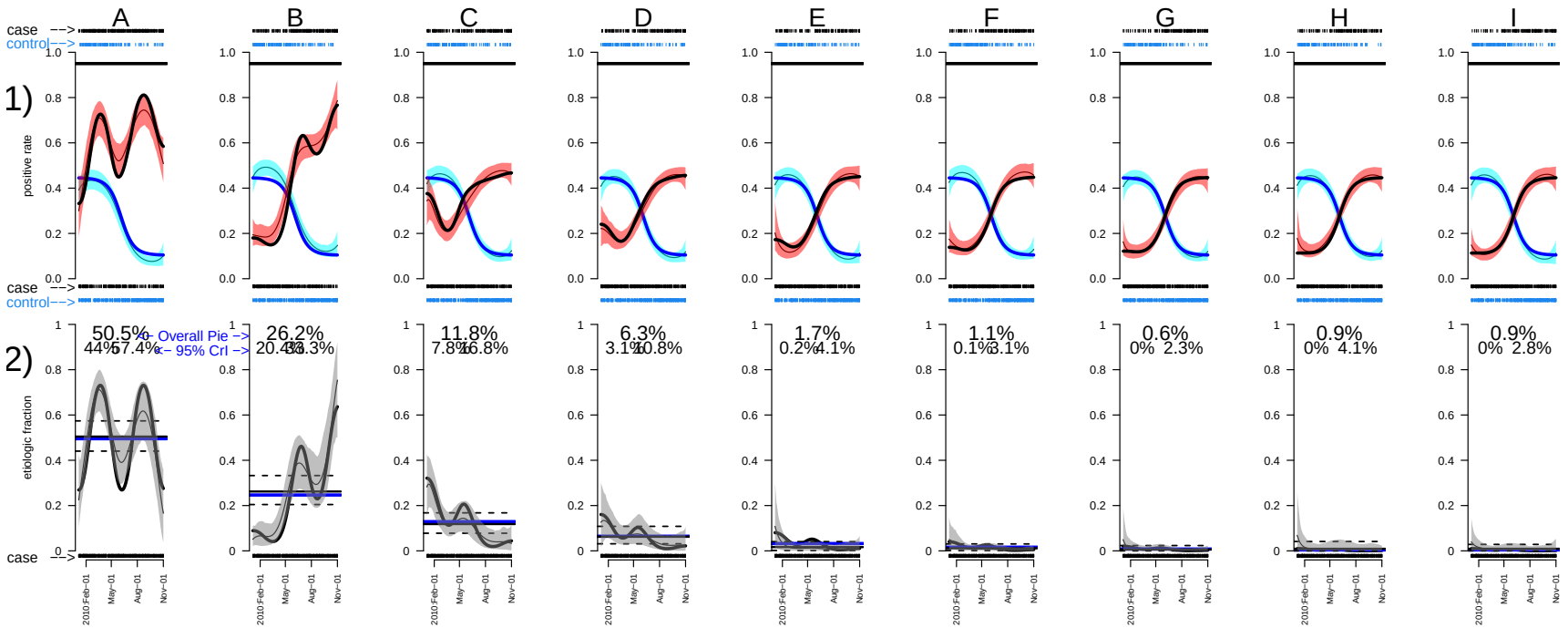
Simulation I Results

- $N_d = 500$ 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

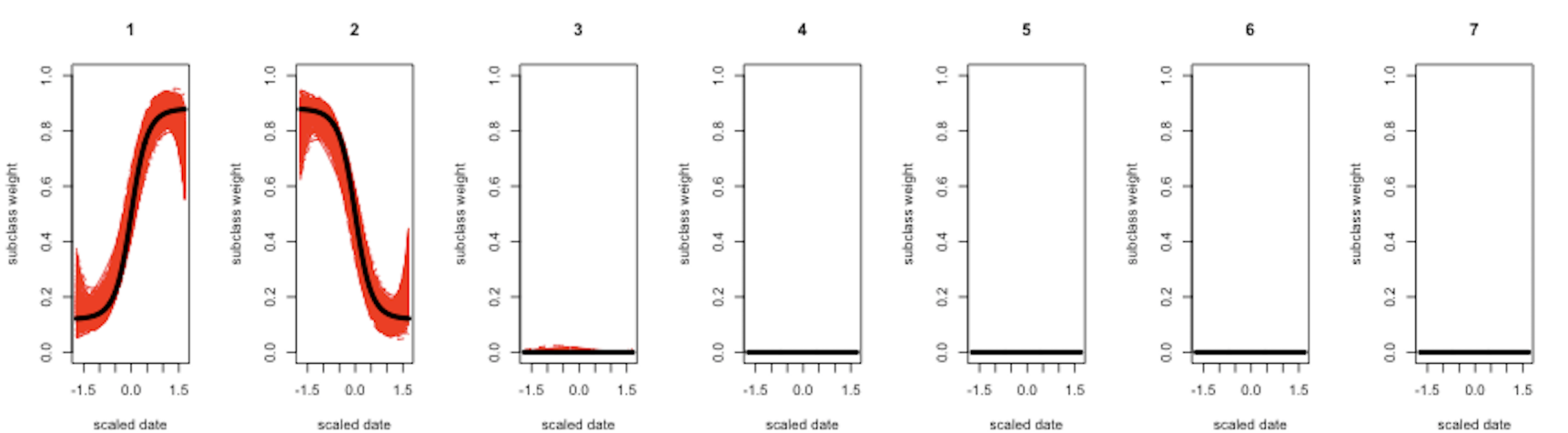


Simulation I: Recovery of Truth $\pi_\ell^0(t, S = s)$

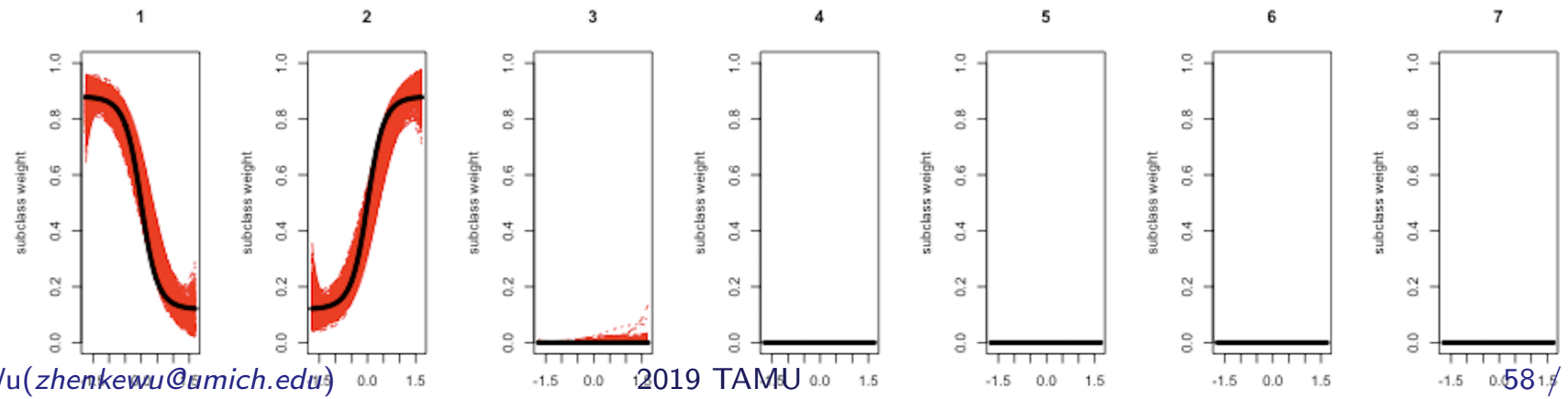


Simulation I: Recovery of $\nu_k(t)$ and $\eta_k(t)$

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 CSCF variability across X: i) $\boldsymbol{\beta}_0^{\mathrm{i}} = (0, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_1^{\mathrm{i}} = (-1.5, 0, -1.5, -1.5, 0, -1.5)$; ii) $\boldsymbol{\beta}_0^{\mathrm{ii}} = (1, 0, 1, 1, 0, 1)$ and $\boldsymbol{\beta}_1^{\mathrm{ii}} = (-1.5, 1, -1.5, -1.5, 1, -1.5)$
- 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 $j = 1, \ldots, J$.
- $\nu_k(W) = \eta_k(W) = \mathit{logit}^{-1}\left(\gamma_{k0} + \gamma_{k1}\,\mathbb{I}\{W = 2\}\right)$ where $(\gamma_{10}, \gamma_{11}) = (-0.5, 1.5)$ and $(\gamma_{20}, \gamma_{21}) = (1, -1.5)$.

Appendix

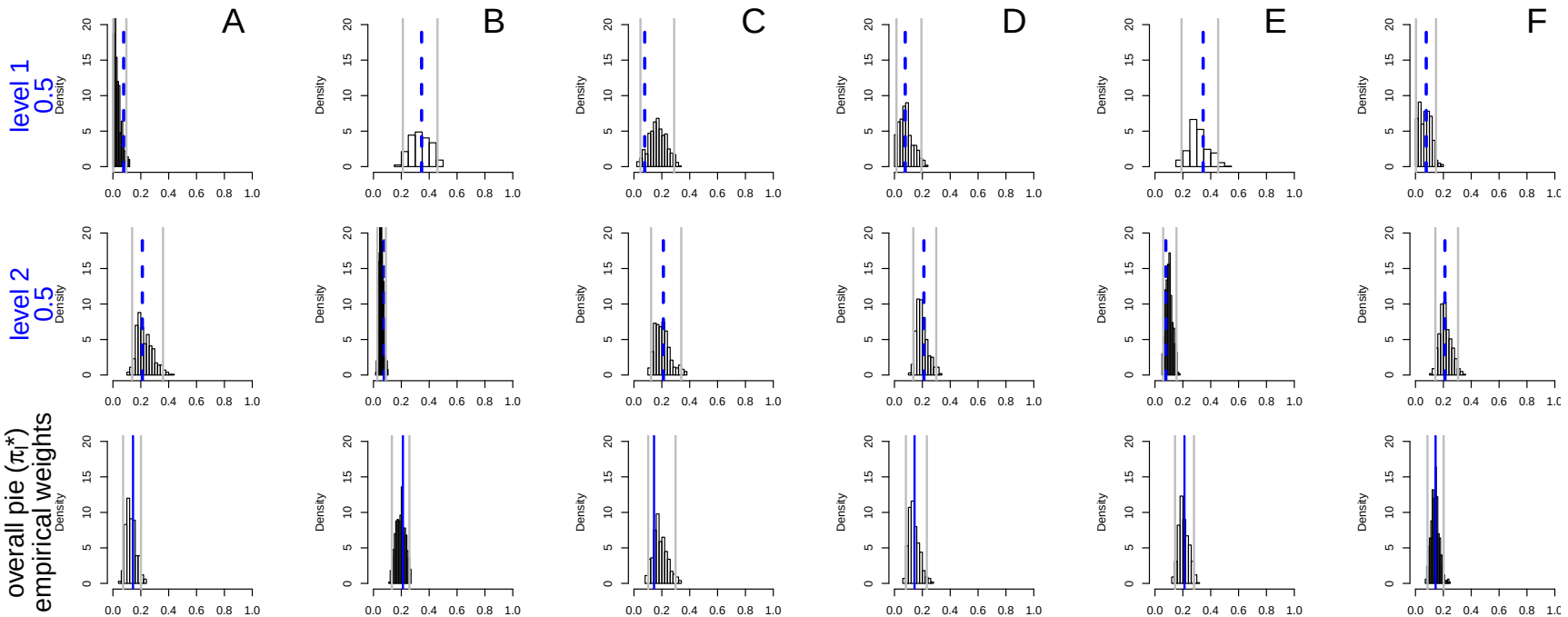


Figure: Posterior distributions of the stratum-specific (Row 1 and 2) and the overall (Bottom Row) CSCFs based on a simulation with a two-level discrete covariate and $L = J = 6$ causes. The vertical gray lines indicate the 2.5% and 97.5% posterior quantiles, respectively; The truths are indicated by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

Appendix

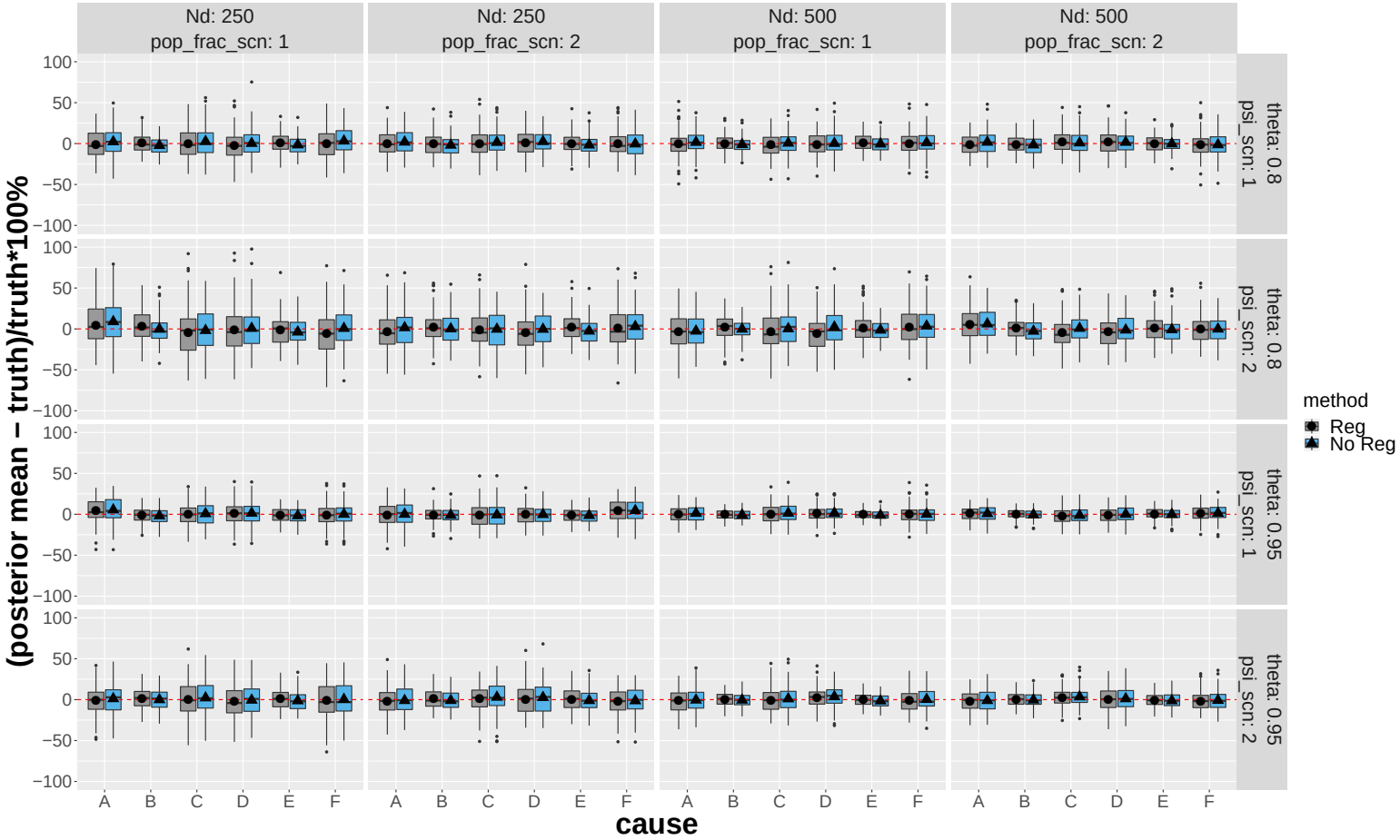
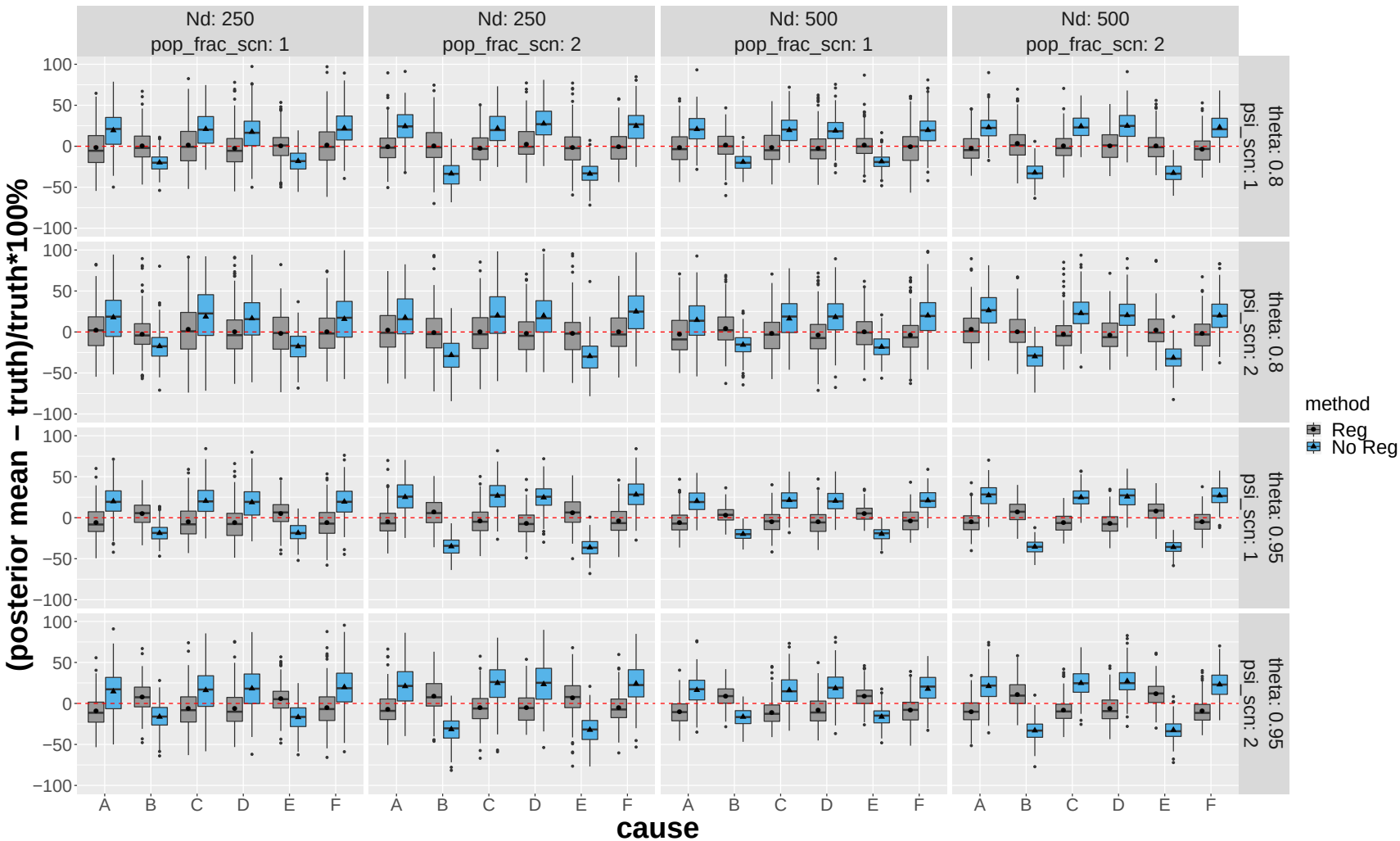
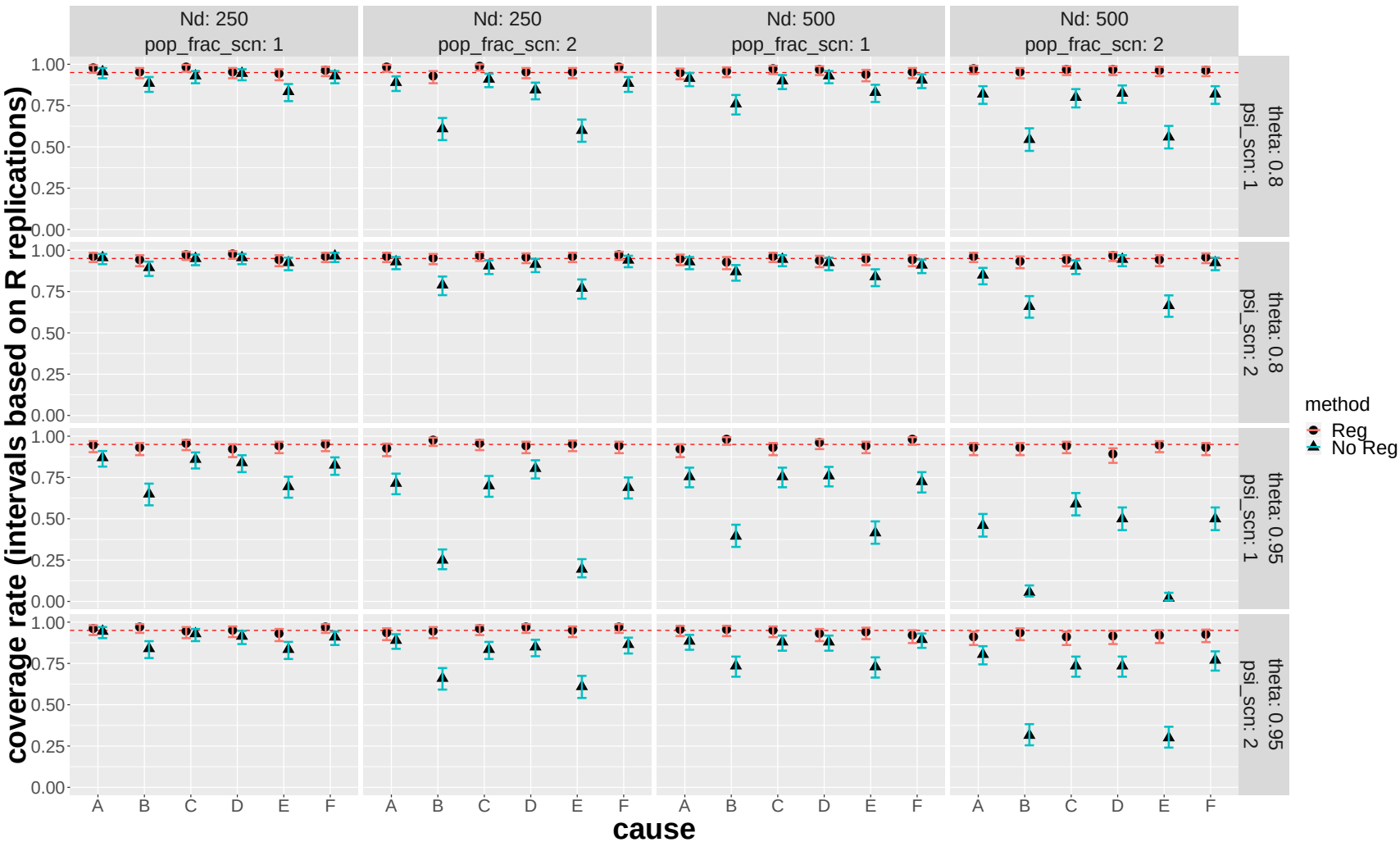


Figure: NPLCM analyses with or without regression perform similarly in terms of percent relative bias (top) and empirical coverage rates (bottom) over $R = 100$ replications in simulations where the case and control subclass weights *do not* vary by covariates. Each panel corresponds to one of 16 combinations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

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



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



Appendix

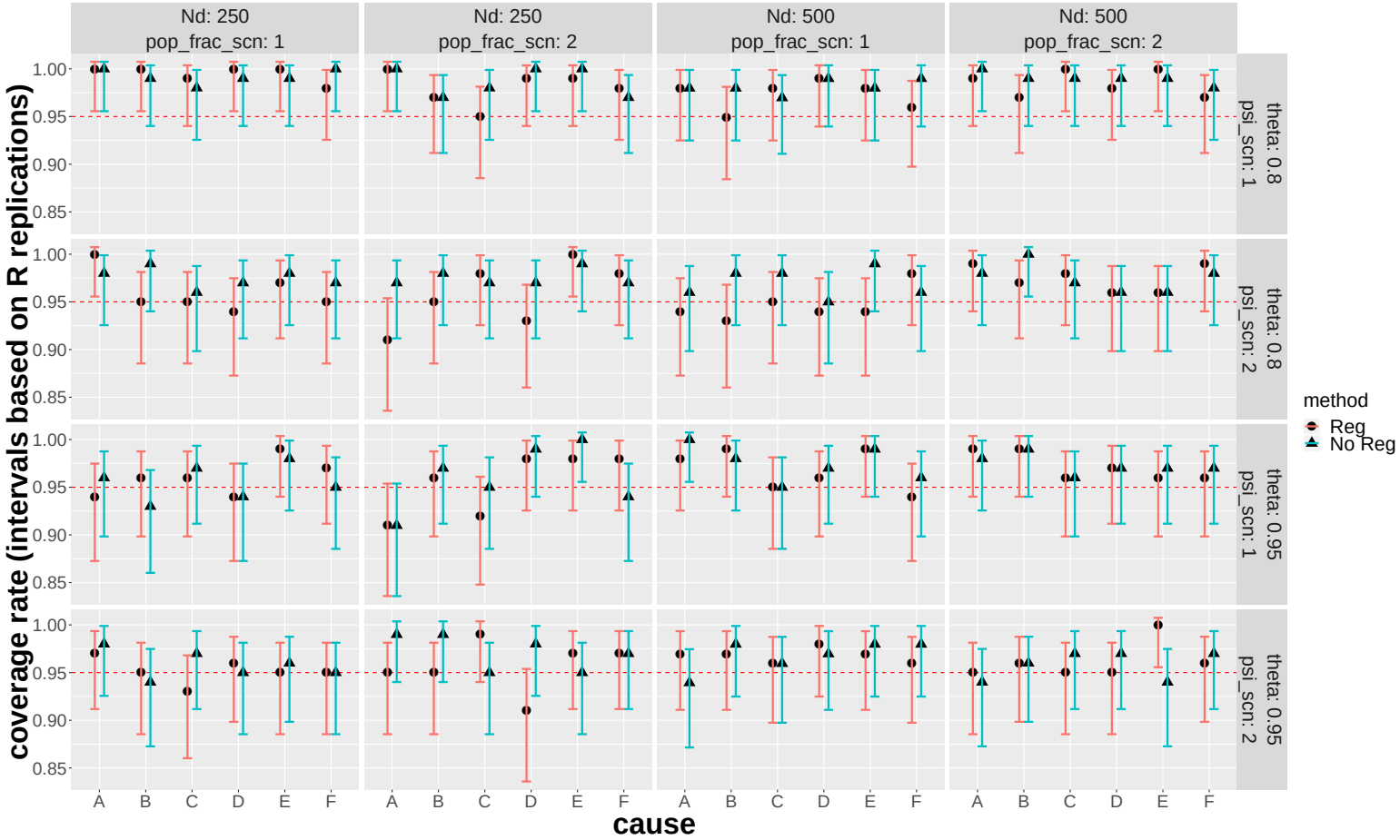
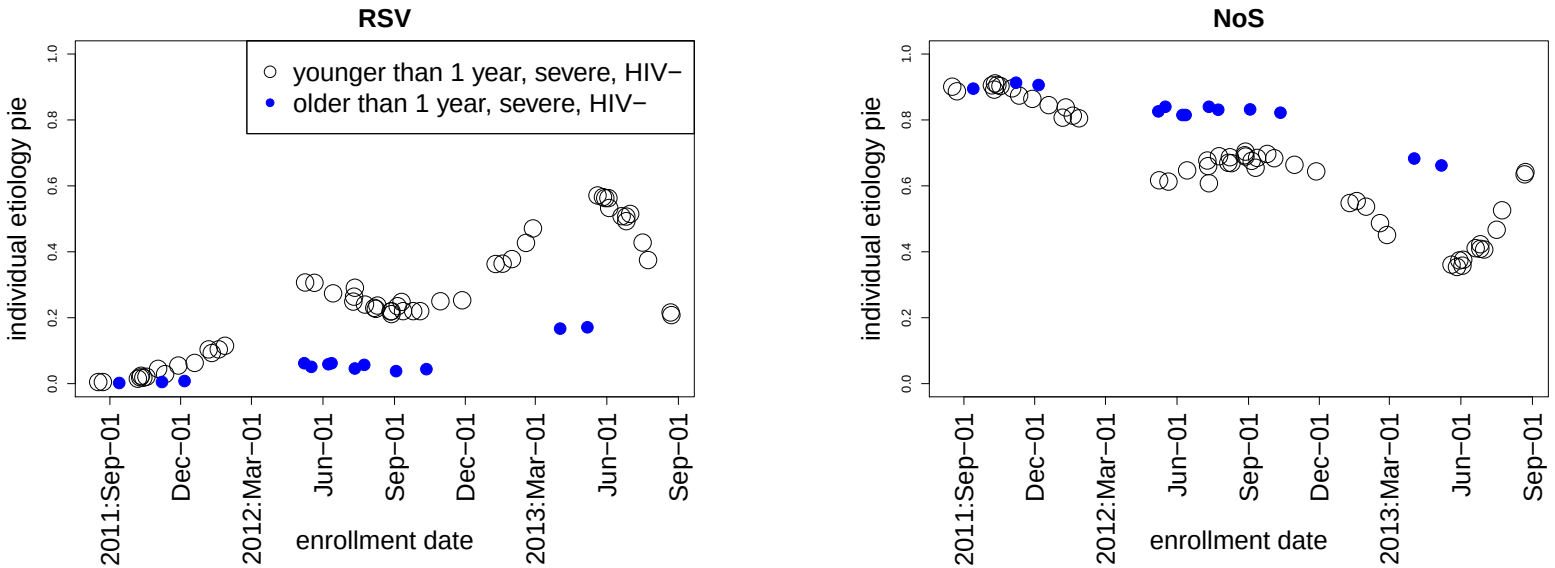


Figure: NPLCM analyses with or without regression perform similarly in terms of percent relative bias (top) and empirical coverage rates (bottom) over $R = 100$ replications in simulations where the case and control subclass weights *do not* vary by covariates. Each panel corresponds to one of 16 combinations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