Biostatistics (2015), **0**, 0, *pp.* 1–30
doi:10.1093/biostatistics/nplcm˙20151028

Nested Partially-Latent Class Models for Dependent Binary Data; Estimating Disease Etiology

ZHENKE WU*,[1], MARIA DELORIA-KNOLL[2], SCOTT L. ZEGER[1]

[1] *Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205*

[2] *Department of International Health, Johns Hopkins University, Baltimore, MD 21205*

zhwu@jhu.edu

Summary

The Pneumonia Etiology Research for Child Health (PERCH) study seeks to use modern measurement technology to infer the causes of pneumonia for which gold-standard evidence is unavailable. The paper describes a latent variable model designed to infer from case-control data the etiology distribution for the population of cases, and for an individual case given his or her measurements. We assume each observation is drawn from a mixture model for which each component represents one cause or disease class. The model addresses a major limitation of the traditional latent class approach by taking account of residual dependence among multivariate binary outcome given disease class, hence reduces estimation bias, retains efficiency and offers more valid inference. Such "local dependence" on a single subject is induced in the model by nesting latent subclasses within each disease class. Measurement precision and covariation can be estimated using the control sample for whom the class is known. In a Bayesian framework, we use stick-breaking priors on the subclass indicators for model-averaged inference across different

*To whom correspondence should be addressed.

numbers of subclasses. Assessment of model fit and individual diagnosis are done using posterior samples drawn by Gibbs sampling. We demonstrate the utility of the method on simulated and on the motivating PERCH data.

Key words: Bayesian methods; Case-control studies; Local dependence; Latent class model; Measurement error; Disease etiology.

1. Introduction

Clinicians routinely use measurements to differentially diagnose a patient's unknown disease etiology and then choose a treatment from among those available. More often than not, the differential diagnosis is a qualitative process based on judgment and experience. As clinical measurements become more precise and complex and as the number of possible known etiologies grows, such qualitative processes are less likely to be optimal. An important question therefore is whether formal probabalistic calculations can improve clinical decisions when the relevant information is quantitative. For example, in the Pneumonia Etiology Research for Child Health (PERCH) study of childhood pneumonia (Levine *and others*, 2012), a vector of presence/absence indicators for a large number of pathogens is measured on each child by polymerase chain reaction (PCR) using specimens from the nasopharyngeal cavity. A clinical goal is to use the multivariate binary response to infer the pathogen in the child's lung causing pneumonia.

In addition, public health researchers are interested in estimating the population fraction of cases caused by each pathogen, referred to as the *etiologic fractions* or *population etiology distribution* (Feikin *and others*, 2014). Knowledge of the etiology distribution is essential for planning prevention and treatment programs. Because the lung cannot be directly sampled, except in cases of critical illness, imperfect measurements from the periphery are used to infer the *latent state* of the disease.

Figure 1 summarizes the relations among measurements, covariates and lung infection for an individual case. PERCH intends to infer her latent lung infection status (I_i, the latent state) by collecting multivariate binary measurements $\boldsymbol{M}_i$ from the periphery. The joint distribution for $\boldsymbol{M}_i$ is characterized by the true- and false- positive rates and the distribution of the latent disease-causing infection. Covariates such as age and HIV status can also influence the chance for each pathogen causing her disease.

In general terms, the PERCH scientific questions require inference about latent random variables (e.g. Bollen, 2002). The same is true for many other problems, for example, biomarkers for disease diagnosis (e.g. Jokinen and Scott, 2010), words for learning topics of a text (e.g. Hofmann, 2001), and questionnaire items for evaluating severity of depression (e.g. Kroenke and Spitzer, 2002). One way of classifying latent variable models is by the discrete or continuous nature of their latent and manifest (observed) variables. Among them, "latent class" models (LCM) for discrete latent and discrete manifest variables were developed and widely applied since the 1950s (e.g. Lazarsfeld, 1950; Anderson, 1954; Lazarsfeld *and others*, 1968; Goodman, 1974).

LCMs constitute a family of distributions for correlated discrete measurements. The conventional LCM generally makes *local independence* (LI) assumption that manifest variables are independent of one another given the latent class (Lord, 1952; Lazarsfeld, 1959; McDonald, 1981; Bartholomew *and others*, 2011). In the multivariate binary case, individual i's measurement vector, $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})'$, is linked to her latent class (I_i) by the simple product likelihood $\mathbb{P}(\boldsymbol{M}_i \mid I_i = \ell, \boldsymbol{\theta}) = \prod_{j=1}^{J} \mathbb{P}(M_{ij} \mid I_i = \ell, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the collection of measurement parameters — sensitivities and specificities. We then obtain the observed likelihood by summing over all the possible values of I_i, i.e., $\mathbb{P}(\boldsymbol{M}_i \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{\ell=1}^{L} \pi_\ell \prod_{j=1}^{J} \mathbb{P}(M_{ij} \mid I_i = \ell, \boldsymbol{\theta})$, where $\boldsymbol{\pi}$ is a vector of mixing weights of length L. The LI assumption implies that the latent memberships I_i completely explains the marginal dependence in $\boldsymbol{M}_i$. Under local identifiability conditions (Jones *and others*, 2010), we can estimate $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ by the values that optimally reduce the observed dependence among measurements given latent class. The expectation-maximization (EM) algorithm (Goodman, 1974) is one popular approach. Individual classification can then proceed by applying Bayes rule using the estimated parameters.

When classes are observed for some subjects, for example, motivated by the known control infection status $I_i = 0$, Wu *and others* (2015) introduced a "partially-latent" class model (pLCM). The control sample provides the requisite information to estimate the specificities of the measurements. In the original formulation, they assumed LI for the multivariate binary measurements within each class. However, within cases or controls, several pairs of pathogens had observed log odds ratios that are inconsistent with their model-based predictive distributions. To address this lack of fit in the covariances, one approach is to extend pLCM by introducing dependence among measurements for persons within the same class. These associations have scientific value in their own right, for example, to study patterns of pathogen-pathogen stimulation or inhibition.

Deviations from LI, or "local dependence" (LD) can occur in many applications, for example, in medical diagnostic tests when most severely diseased patients and the healthiest patients are easiest to correctly classify (Albert *and others*, 2001), or when tests target on similar genetic molecules (Qu and Hadgu, 1998). Many authors have noted that not accounting for LD can bias estimates of model parameters (e.g. Vacek, 1985; Torrance-Rynard and Walter, 1997; Pepe and Janes, 2007). Therefore, in many applications where the LI model for $[\boldsymbol{M}_i \mid I_i]$ is assumed, model adequacy is studied to ensure valid model-based conclusions (e.g. Garrett and Zeger, 2000; Wu *and others*, 2015).

Ideas for relaxing LI can be distinguished by whether or not extra latent variables are introduced. Without doing so, Harper (1972) modeled associations between pairs, triples, and higher order combinations of variables given latent class. Haberman (1979) and Espeland and Handelman (1989) used log-linear models to extend LCM viewing the latent class as one of the category variables. See also Hagenaars (1988) and Yang and Becker (1997).

The second approach allows for dependence by using extra latent variables of continuous or discrete types or a mixture. For example, Qu *and others* (1996) used Gaussian random intercepts to induce within-subject symmetric and positive correlations among multiple diagnostic tests. Xu and Craig (2009) used probit latent class models for more complex LD structures. Albert *and others* (2001) proposed to nest one extra unobserved subclass within each of two latent classes (diseased or non-diseased) to represent subjects measured without error. Dendukuri *and others* (2009) hierarchically layered extra mixed latent variables in a Bayesian framework. Adding extra latent variables can account for LD because any multivariate discrete distribution can be represented by a locally independent LCM with sufficiently many latent classes (Dunson and Xing, 2009, Corollary 1). However, when a satisfactory fit requires many classes — especially when the dimension of manifest variables is high — interpreting inferred classes remains a difficult task.

In this paper, we build on the second strategy and develop a novel latent variable model for multivariate binary data obtained from a *case-control* study. Using control data with a known class and assuming the covariation among control measurements is shared among the other latent classes for cases, we extend the traditional latent class approach to avoid the LI assumption. The proposed model is a natural extension of pLCM (Wu *and others*, 2015) and can be used to test its LI assumption.

We assume each child's measurements comprise an observation from a mixture model with component classes that represent the L different pathogens that can cause her pneumonia. One primary goal of analysis is to estimate the probability distribution for these classes. To allow for LD, we introduce *latent subclasses* nested within each of the $L+1$ (L case, 1 control) disease classes. Measurements within a subclass are assumed independent. We refer to the model as a "nested partially-latent class model" or npLCM and use a Bayesian penalty to encourage small but variable numbers of subclasses that parsimoniously approximate the multivariate discrete dependence and avoid overfitting (Section 2.5).

We show that the proposed model is partially-identifiable (Gustafson, 2015) and incorporate prior knowledge about measurement sensitivities to facilitate Bayesian estimation of the etiologic fractions. The npLCM model is estimated via Markov chain Monte Carlo (MCMC) with designed precision to approximate the posterior distributions of the population etiologic fractions, individual latent state, as well as functions of them, such as the fraction of pneumonia cases caused by bacteria.

In Section 2, we formulate our model and discuss its statistical properties. Section 3 provides details on the posterior sampling algorithm to draw inference based on our model. Section 4 illustrates through asymptotic evaluations and finite-sample simulations the benefits of the new model relative to a version that ignores LD. Section 5 applies the proposed method to PERCH study data. Section 6 concludes with remarks on the method's advantages, limitations, and future extensions.

2. Nested Partially-Latent Class Model

In this section, we specify the nested partially-latent class model (npLCM) and consider its statistical properties using the PERCH study example to make the ideas concrete. Let $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})'$ comprise a J-dimensional multivariate binary measurement collected for subjects $i = 1, ..., n_1 + n_0$, where the first n_1 subjects are cases and the remaining n_0 are controls. Let $Y_i = 1$ denote a case and $Y_i = 0$ denote a control.

2.1 *Measurement Likelihood*

Figure 2 pictures the general structure of the npLCM with $J = 5$ measurements, one pathogen per row in the matrix. With 5 pathogens, there are 6 classes: one for the control state (pathogen-free) on the left of the dashed vertical line; and $L = 5$ case states, one for each possible cause on the right. In the figure, the control measurements have joint distribution that is approximated by a mixture of $K = 2$ subclasses, with K-dimensional mixing weights $\boldsymbol{\nu} = (\nu_1, ..., \nu_K)'$. Here $\boldsymbol{\psi}_k = \{\psi_k^{(j)}\}_{1 \leq j \leq J}$ is the column vector of false positive rates for measurements $j = 1, ..., J$, for subclass $k = 1, ..., K$. The mixing weights of the K subclasses in the case population (right of dashed line) are assumed to be $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)'$. The *etiologic fractions* are the mixing weights for the $L(= J)$ classes in the case population, denoted $\boldsymbol{\pi} = (\pi_1, ..., \pi_L)'$ with $\sum_{\ell=1}^{L} \pi_\ell = 1$.

Throughout the paper, we rely on the scientific assumption that each child's pneumonia is caused by a single primary pathogen. The more general case where disease can be attributed to multiple pathogens is a natural extension (Section 6).

2.2 *Control Likelihood*

The control measurement distribution is assumed to take the form in Goodman (1974). Mutual dependence is induced by the existence of multiple subclasses, with each subclass having possibly distinct positive rate profiles. Given an unobserved subclass, measurements are assumed to be mutually independent. Marginalizing over the latent subclasses produces dependence for pathogens with different rates across subclasses. The formulation is natural for PERCH given the heterogeneity in the health status of controls.

For control i, we introduce subclass indicator Z_i that takes value in $\{1, ..., K\}$ and let

$$\textsf{sample subclass indicator}: \quad Z_i \sim \textsf{Multinomial}(\{1, ..., K\}, \boldsymbol{\nu}) \tag{2.1}$$

$$\textsf{generate measurements}: \quad M_{ij} \mid Z_i = k \sim \textsf{Bernoulli}(\psi_k^{(j)}), \text{ independently for } j = 1, ..., J, \tag{2.2}$$

where $\nu_k = \mathbb{P}(Z_i = k \mid Y_i = 0)$ and $\psi_k^{(j)} = \mathbb{P}(M_{ij} = 1 \mid Z_i = k, Y_i = 0)$. Here $\boldsymbol{\nu}$ comprises of the probabilities of a control falling in the subclasses; $\psi_k^{(j)}$ is the probability of a positive response within subclass k viewed as an event of false detection for controls and hence is termed the false positive rate (FPR); the FPRs for subclass k are collected in the FPR profile vector $\boldsymbol{\psi}_k$ which is then combined by column into the matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 | ... | \boldsymbol{\psi}_K]$ for all subclasses. The control distribution of the 2^J measurement patterns ($\forall \boldsymbol{m} \in \{0,1\}^J$) are then given by

$$\boldsymbol{P}^0(\boldsymbol{m}) = \mathbb{P}(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{\nu}, \boldsymbol{\Psi}, Y_i = 0) = \sum_{k=1}^{K} \nu_k \prod_{j=1}^{J} \left\{\psi_k^{(j)}\right\}^{m_j} \left\{1 - \psi_k^{(j)}\right\}^{1-m_j}. \tag{2.3}$$

2.3 *Case Likelihood*

For a case with known cause, her vector of binary measurements is again assumed to be generated from a latent K-subclass model as for the controls. In PERCH context, motivated by the observation that cases and controls have similar correlation patterns for many pathogen pairs (e.g., Appendix Figure 2), we let the cases share controls' measurement characteristics. To be more precise, given a case's disease class $I_i = \ell_0 \in \{1, \ldots, L\}$, with $L = J$, she falls into subclass k with probability η_k, for $k = 1, ..., K$. Then subclass k's response probabilities are assumed equal to $\psi_k^{(j)}$ as in controls for $j \neq \ell_0$, and equal to a new parameter $\theta_k^{(j)}$ for $j = \ell_0$. That is, an infection by pathogen j may alter the response probabilities in the j-th dimension but not others. Since the disease for cases i is in fact unknown, her measurement distribution is a mixture across all L states given by $\boldsymbol{P}^1(\boldsymbol{m}) = \mathbb{P}(\boldsymbol{M}_i = \boldsymbol{m} \mid \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, Y_i = 1)$, $\forall \boldsymbol{m} \in \{0,1\}^J$,

$$\boldsymbol{P}^1(\boldsymbol{m}) = \sum_{\ell=1}^{L} \pi_\ell \sum_{k=1}^{K} \left[\eta_k \left\{\theta_k^{(\ell)}\right\}^{m_\ell} \left\{1 - \theta_k^{(\ell}\right\}^{1-m_\ell} \prod_{j \neq \ell} \left\{\psi_k^{(j)}\right\}^{m_j} \left\{1 - \psi_k^{(j)}\right\}^{1-m_j}\right], \tag{2.4}$$

where $\boldsymbol{\Theta}$ is a parameter matrix with (j,k)-th element $\theta_k^{(j)}$.

We can reformulate (2.4) by a three-stage generative process similar to (2.1-2.2) by indicators of case disease classes I_i and the nested subclasses Z_i:

$$\text{sample class indicator :} \quad I_i \mid Y_i = 1 \sim \text{Multinomial}(\{1, ..., L\}, \boldsymbol{\pi}), \tag{2.5}$$

$$\text{sample subclass indicator :} \quad Z_i \mid I_i = \ell \sim \text{Multinomial}(\{1, ..., K\}, \boldsymbol{\eta}), \ell = 1, ..., L \tag{2.6}$$

$$\text{generate measurements :} \quad M_{ij} \mid Z_i = k, I_i \sim \text{Bernoulli}\left(\theta_k^{(j)} \mathbf{1}_{\{I_i = j\}} + \psi_k^{(j)} \mathbf{1}_{\{I_i \neq j\}}\right), \tag{2.7}$$

independently for $1 \leq j \leq J$. At the first stage, the vector $\boldsymbol{\pi}$ comprises probabilities of a case in class 1 to L and is the primary target of inference in this paper. Then, the cases' subclass mixing weights $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$ determines the probability of a case falling into each subclass. The final stage generates the measurement at the j-th dimension: positive with probability $\theta_k^{(j)}$ or $\psi_k^{(j)}$ according as the realized values of I_i and Z_i in previous steps. Because $\theta_k^{(j)}$ is the probability of true detection for infections caused by pathogen j, we term it true positive rate (TPR) and collect them in $\boldsymbol{\theta}_k = (\theta_k^{(1)}, \ldots, \theta_k^{(J)})'$ for subclass k.

Importantly, case and controls' subclass mixing weights ($\boldsymbol{\eta}$ and $\boldsymbol{\nu}$) need not be identical. This admits a measurement dependence structure for cases different from that in controls, such as pathogen increased or reduced interactions due to the former's lung infection. We refer to the special case of $\boldsymbol{\eta} = \boldsymbol{\nu}$ (element-wise equality) as *non-interference submodels*, under which controls and cases of class j have identical distributions of the leave-one-dimension-out measurement vector $\boldsymbol{M}_{i[-j]}$. Further setting $\eta_1 = \nu_1 = 1$, or $K = 1$, gives the pLCM.

We have assumed cases' latent states categories take value from a complete list of J measured pathogens (i.e., $L = J$). The case likelihood (2.4) can be extended to account for *other* causes by adding an extra term: $\pi_{J+1} \sum_{k=1}^{K} \eta_k \left(\prod_{j=1}^{J} \{\psi_k^{(j)}\}^{m_j} \{1 - \psi_k^{(j)}\}^{1-m_j} \right)$, where $\pi_{J+1} = \mathbb{P}(I_i = J+1)$ is the total etiology fraction of other causes. For a clinically-confirmed pneumonia case, the negative responses on J pathogens by highly-sensitive assays indicate the possibility of other etiologic pathogens.

Combining (2.3) and (2.4), the joint likelihood across independent subjects is given by

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\eta}; \mathcal{D}) = \prod_{i:Y_i=0} \boldsymbol{P}^0(\boldsymbol{M}_i) \prod_{i:Y_{i'}=1} \boldsymbol{P}^1(\boldsymbol{M}_{i'}), \tag{2.8}$$

where $\mathcal{D}$ collects all the measurement data.

2.4 *Properties*

The proposed model extends pLCM in Wu *and others* (2015) by adding $(3J+1)(K-1)$ additional parameters compared to the original formulation with the total number of parameters linear in J when $K \ll J$ providing a parsimonious approximation to the case and control joint distributions that require $2(2^J-1)$ parameters in a saturated model. We further reduce the effective number of parameters using a penalty prior (Section 2.5).

We assumed that the LD of measurements within each case class can be explained by allowing the same number of LI subclasses as in the controls, so that the case subclass measurement parameters can be partly informed by their control counterparts (see (2.7)). Additional case subclasses can be included once I_i is directly observed for some cases.

In Appendix A, we provide expressions of the marginal means and pairwise associations for multivariate binary measurements given the npLCM likelihood. These formulas are used to study the magnitude of dependence given true parameters and to generate marginal posterior distributions for observables used in model checking, as illustrated in Section 4.1 and 5.

2.5 *Prior Specifications*

For the npLCM, we specify the prior distributions on unknown parameters as follows:

$$\boldsymbol{\pi} \sim \mathsf{Dirichlet}(a_1,\ldots,a_L), \tag{2.9}$$

$$\psi_k^{(j)} \sim \mathsf{Beta}(b_{1kj},b_{2kj}), j=1,...,J; k=1,...,\infty, \tag{2.10}$$

$$\theta_k^{(j)} \sim \mathsf{Beta}(c_{1kj},c_{2kj}), j=1,...,J; k=1,...,\infty, \tag{2.11}$$

$$Z_i \mid Y_i=1 \sim \sum_{k=1}^{\infty} U_k \prod_{s<k}\left[1-U_s\right]\delta_k, \qquad U_k \sim \mathsf{Beta}(1,\alpha_1), i=1,...,n_1, \tag{2.12}$$

$$Z_i \mid Y_i=0 \sim \sum_{k=1}^{\infty} V_k \prod_{s<k}[1-V_s]\delta_k, \qquad V_k \sim \mathsf{Beta}(1,\alpha_0), i=n_1+1,...,n_1+n_0, \tag{2.13}$$

$$\alpha_0,\alpha_1 \sim \mathsf{Gamma}(0.25,0.25), \tag{2.14}$$

where δ_k is a point mass on k, and prior independence is also assumed among these parameters. As discussed in more detail by Wu *and others* (2015), the npLCM is partially identified (Jones *and others*, 2010). Specifically, the TPRs $\boldsymbol{\Theta}$ are not fully identified by the model likelihood (2.8). Therefore, we choose $(c_{1kj},c_{2kj}),\forall k,j$, so that the 2.5% and 97.5% quantiles of the Beta distribution with parameters (c_{1kj},c_{2kj}) match the prior minimum and maximum TPR values elicited from pneumonia experts (Section 5). Otherwise, we use the default value of 1s for the Beta hyperparameters. Hyperparameters for the etiology prior, $(a_1, ..., a_J)'$, are usually 1s to denote equal and non-informative prior weights for each pathogen if expert prior knowledge is unavailable.

Because our goal is to estimate the etiology fractions, $\boldsymbol{\pi}$, after marginalizing over subclass indicators (Z_i), the parameters for the dependence structure within each disease class are nuisance parameters. Therefore, rather than fixing K, we let K be a random positive integer and perform model averaging using a prior that encourages small values of K to incorporate its uncertainty into the inference about $\boldsymbol{\pi}$ in a parsimonious way. This prevents model overfitting in finite samples when the observed contingency table for the multivariate binary PERCH measurements has mostly empty cells. In (2.12) and (2.13), we have actually specified stick-breaking priors for both $\boldsymbol{\eta} = \left\{U_k \prod_{s<k} [1 - U_s]\right\}_{k=1,2,...}$ and $\boldsymbol{\nu} = \left\{V_k \prod_{s<k} [1 - V_s]\right\}_{k=1,2,...}$ that on average place decreasing weights on the kth subclass as k increases (Sethuraman, 1994). Appendix B further discusses the use of stick-breaking prior in our model. The priors above are conjugate to the likelihood of unknown parameters, making the Gibbs sampler in Section 3 conveniently constructed.

3. Posterior Computations

The posterior distributions of the population etiology fraction vector $(\boldsymbol{\pi})$, TPRs $(\boldsymbol{\Theta})$ and FPRs $(\boldsymbol{\Psi})$ can be estimated by simulating approximating samples from the joint posterior via Markov chain Monte Carlo (MCMC) algorithms (Brooks *and others*, 2011). Appendix Figure 1 presents the directed acyclic graph (DAG) for the model structure with observed and latent variables in the npLCM. For posterior computation involving stick-breaking priors, without truncation on the number of stick segments, Walker (2007) and Papaspiliopoulos and Roberts (2008) proposed the slice sampler and retrospective MCMC, respectively. In the following, we develop a simple and efficient blocked Gibbs sampler relying on truncation approximation to the stick-breaking prior distribution (e.g., Ishwaran and James, 2001; Gelfand and Kottas, 2002). We also include in the sampling algorithms two sets of auxiliary variables, the partially-latent individual class indicator (I_i) the nested subclass indicator (Z_i). Appendix C shows the algorithm step-by-step.

All model estimations are performed by the `R` package "`baker`" (https://github.com/zhenkewu/baker) that interfaces with freely available software `JAGS 3.4.0` (http://mcmc-jags.sourceforge.net/). Convergence was monitored via MCMC chain histories, auto-correlations, kernel density plots, and Brooks-Gelman-Rubin statistics (Brooks and Gelman, 1998). The statistical results below are based on $10,000$ iterations of burn-in followed by $50,000$ production samples from each of three parallel chains. Samples from every 50 iterations are retained for inference.

4. Asymptotic and Simulation Studies of Nested Partially-Latent Class Models

This section presents asymptotic and simulation studies to show that for cases like PERCH 1) when the LI assumption is incorrect, a working LI model will estimate $\boldsymbol{\pi}$ with asymptotic bias; 2) fitting the LD model to data generated with LI does not lose too much efficiency using sparse priors on subclass indicators; and 3) compared to the LI model, the LD model produces 95% credible intervals for $\boldsymbol{\pi}$ with better actual coverage rates.

4.1 *Asymptotic Bias Evaluations*

We first evaluate the asymptotic bias of using a working LI model (pLCM) in the estimation of $\boldsymbol{\pi}$ for J causes, i.e., $L = J$. Under the LI assumption, let the maximum likelihood estimator be $\widehat{\boldsymbol{\pi}}_N = (\widehat{\pi}_{N,1}, \ldots, \widehat{\pi}_{N,J-1}, \widehat{\pi}_{N,J})'$, where the J-th etiologic fraction is estimated as $\widehat{\pi}_{N,J} = 1 - \sum_{\ell \neq J} \widehat{\pi}_{N,\ell}$ and $N = n_1 + n_0$ is the total sample size. The estimator $\{\widehat{\boldsymbol{\pi}}_{N,\ell}\}_{1 \leq \ell \leq J-1}$ will converge to the first $L-1$ components of the parameter vector $\boldsymbol{\omega}^* = (\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_{J-1}^*, \boldsymbol{\psi}^{*\mathsf{M}})'$ that minimizes the

Kullback-Leibler information criterion, or equivalently, to the $\boldsymbol{\omega}^*$ that satisfies

$$\mathbb{E}_{\mathbf{\Omega}_0}\left\{\frac{\partial}{\partial \boldsymbol{\omega}} \log \mathbb{P}_{\boldsymbol{\omega}}(\boldsymbol{M}_i \mid Y_i)\Big|_{\boldsymbol{\omega}^*}\right\} = 0, \tag{4.15}$$

where $\mathbf{\Omega}_o = (\boldsymbol{\pi}_o, \boldsymbol{\Psi}_o, \mathbf{\Theta}_o)$ is the collection of etiologic fractions, FPRs and TPRs in the true data generating mechanism, and $\mathbb{P}_{\boldsymbol{\omega}}(\boldsymbol{M}_i \mid Y_i)$ is the likelihood of the pLCM given by

$$\sum_{\ell \neq J} \pi_\ell \cdot f_\ell(\boldsymbol{M}_i, \boldsymbol{\psi}^{\mathsf{M}}) + \left(1 - \sum_{\ell \neq J} \pi_\ell\right) \cdot f_J(\boldsymbol{M}_i, \boldsymbol{\psi}^{\mathsf{M}}), \tag{4.16}$$

where $f_\ell(\boldsymbol{m}, \boldsymbol{\psi}^{\mathsf{M}}) = (\theta_\ell^{\mathsf{M}})^{\boldsymbol{m}_\ell}(1-\theta_\ell^{\mathsf{M}})^{1-\boldsymbol{m}_\ell}\prod_{j\neq\ell}(\psi_j^{\mathsf{M}})^{\boldsymbol{m}_j}(1-\psi_j^{\mathsf{M}})^{1-\boldsymbol{m}_j}$ for cases $Y_i = 1$; and $\prod_{j=1}^{J}(\psi_j^{\mathsf{M}})^{\boldsymbol{m}_{ij}}(1-\psi_j^{\mathsf{M}})^{1-\boldsymbol{m}_{ij}}$ for controls $Y_i = 0$. We also fix at the true values the *marginal* sensitivities $\theta_j^{\mathsf{M}} = \sum_{k=1}^{K}\theta_{o,k}^{(j)}\eta_k, j = 1, \ldots, J$, to eliminate the partial-identifiability issue and to focus on asymptotic bias evaluations. Our calculation of the expectation in (4.15) assumed equal case and control sample sizes, and could be easily modified for other sampling ratios. Further, White (1982) also established the asymptotic normality of the estimator: $\sqrt{N}(\widehat{\boldsymbol{\omega}}_N - \boldsymbol{\omega}^*) \overset{d}{\to} \mathcal{N}\left(\mathbf{0}, A(\boldsymbol{\omega}^*)^{-1}B(\boldsymbol{\omega}^*)A(\boldsymbol{\omega}^*)^{-1}\right)$, where

$$A(\boldsymbol{\omega}^*) = -\mathbb{E}_{\mathbf{\Omega}_0}\left\{\frac{\partial^2}{\partial \boldsymbol{\omega}^2}\log\mathbb{P}_{\boldsymbol{\omega}}(\boldsymbol{M}_i \mid Y_i)\Big|_{\boldsymbol{\omega}^*}\right\}, B(\boldsymbol{\omega}^*) = \mathbb{V}_{\mathbf{\Omega}_0}\left\{\frac{\partial}{\partial \boldsymbol{\omega}}\log\mathbb{P}_{\boldsymbol{\omega}}(\boldsymbol{M}_i \mid Y_i)\Big|_{\boldsymbol{\omega}^*}\right\}. \tag{4.17}$$

The robust variance of $\widehat{\boldsymbol{\omega}}_N$ is calculated as $V_R^* = N^{-1}A^{-1}(\boldsymbol{\omega}^*)B(\boldsymbol{\omega}^*)A^{-1}(\boldsymbol{\omega}^*)$ by approximating the variance operator in $B(\boldsymbol{\omega}^*)$ using empirical samples. We use Monte Carlo method with 10^7 samples from $[\boldsymbol{M}_i \mid Y_i = y]$, for $y = 0, 1$, to evaluate the expectation in (4.15) and then numerically solve it for its root to obtain $\boldsymbol{\omega}^*$. We then calculate (4.17) by plugging in $\boldsymbol{\omega}^*$ and evaluating the expectation using 10^7 Monte Carlo samples.

The strength of LD given disease class determines the estimation bias. When the true data generating mechanism is close to independence, the working LI model estimates of $\boldsymbol{\pi}$ are close to being asymptotically unbiased. To illustrate, we quantify the asymptotic bias for $J = 5$ binary measures (pathogens A, B, C, D and E). We generate Monte Carlo samples from the true data generating mechanisms with varying degrees of LD, while fixing the etiologic fraction

$\boldsymbol{\pi}_o = (0.5, 0.2, 0.15, 0.1, 0.05)'$ to mimic what is seen in PERCH. We create associations among measurements by defining two subclasses ($K = 2$) for each of the 6 disease states (controls plus 5 disease classes for cases). We consider two scenarios of measurement parameters $(\boldsymbol{\Psi}, \boldsymbol{\Theta})$: ($\mathsf{I}$) little LD — small between-subclass differences in positive rates; (II) substantial LD — large differences (see Appendix F).

The subclass weights characterize the degree of LD. We assume controls and cases fall into the first subclass with probability $\nu_o = 0.5$ and $\eta_o \in [0, 1]$, respectively. Row (a) of Figure 3 summarizes both the marginal and within-class dependence for Scenario I and II. The marginal associations are stronger in Scenario II (solid curves). Note that the within-class odds ratio curves leave and return to 1 and remain above or below 1 as η_o increases from 0 to 1 (non-solid lines labelled by A-E in small panels), because when all the weight is on one of the two subclasses, the true data generating mechanism satisfies LI. In particular, the equality $\eta_o = \nu_o (= 0.5)$ represents identical LD structures (non-interference submodels) for cases and controls, with deviations from it indicating differential dependence patterns.

Row (b) of Figure 3 summarizes the results by the percent relative asymptotic bias (PRAB) at all η_o values for each etiologic fraction, $(\pi^*_\ell - \pi_{o,\ell})/\pi_{o,\ell} \times 100\%$. The working LI model produces PRABs less than 13% in magnitude in Scenario I. Given small asymptotic biases, we also obtain good estimates of precision produced by the working LI model with the ratios for model-based variance $V^*_M = N^{-1}A^{-1}(\boldsymbol{\omega}^*)$ versus the robust variance V^*_R between 0.97^2 and 1.05^2 for A-E. The two variances are mathematically identical at arbitrary parameter values if marginal FPRs $(\boldsymbol{\psi}^{\mathsf{M}})$ are known.

The asymptotic bias is large under strong LD as in Scenario II. For example, the working LI model overestimates π_{oC} with 121.3% relative bias at $\eta_o = 0$ for its failure to account for the strong LD among controls. When the case LD is more similar to controls at $\eta_o = 0.5$, the PRAB is 40.5%. This is because the measurement on C is negatively associated with the measurements on B, D, or E given disease class B, D, or E, i.e. mutual inhibition (see shaded cells in Figure 3, a-II), leading to the case pattern $\boldsymbol{M}_i = (1, 0, 1, 1, 0)'$ observed twice as frequently as expected by a working LI model. When they are further assigned with the highest likelihood to cause C under the working LI model, the upward bias results.

4.2 *Bayesian Fitting in Finite Samples*

In finite samples, one can fit the larger LD model that *a priori* encourages a small number of subclasses. Extra subclasses can be used if the measurements have rich multivariate associations. Through simulations, we compare Bayes estimates of etiologic fractions obtained from the npLCM and pLCM. We generate $T = 1,000$ datasets with sample size $n_1 = n_0 = 500$ under Scenario I and II. We fit the npLCM (truncation level $K^* = 5$ subclasses) and pLCM ($K = 1$) to each data set using informative Beta priors on the true positive rates ($\{\theta_k^{(j)}\}$) with 0.5 and 0.99 as the 2.5% and 97.5% quantiles mimicing PERCH study.

We view the Bayes estimates as functionals of data and assess their frequentist properties, such as bias and variance (e.g. Efron, 2015). We define the repeated-sampling bias of the posterior mean and its mean squared error (MSE) respectively as $\lim_T T^{-1} \sum_{t=1}^{T} \left\{ \overline{\pi}_\ell^{(t)} - \pi_{o,\ell} \right\}$, and $\lim_T T^{-1} \sum_{t=1}^{T} (\overline{\pi}_\ell^{(t)} - \pi_{o,\ell})^2, \ell = A, \ldots, E$, where $\overline{\pi}_\ell^{(t)} = \mathbb{E}\{\pi_\ell \mid \mathcal{D}^{(t)}, \mathcal{M}\}$ is the posterior mean taken with respect to the posterior distribution of $\boldsymbol{\pi}$ given the t-th simulated data set $\mathcal{D}^{(t)}$ and model $\mathcal{M}$.

The top panel of Table 1 compares the estimation bias by posterior means obtained from the two models. For a data set with finite sample size, estimation bias can arise from random sampling, model mis-specification or the prior, for which the first is averaged out by replication. The non-zero biases seen here reflect likelihood mis-specification and the influence of the prior. When the likelihood is correctly specified, only biases from priors remain. In Scenario II with strong LD, the npLCM performs much better. For example, the LI assumption (pLCM) results in an upward bias of 26.2% for C at $\eta_o = 0$, as well as other highlighted biases greater than 10% in magnitude. In Scenario I with weak LD, the biases from both models are negligible ($-1.9\% \sim 1.9\%$).

When the truth is close to LI, the npLCM is comprablely efficient to pLCM for almost all settings. The bottom panel of Table 1 shows the the ratio of MSEs for pLCM versus npLCM. In Scenario I, the ratios are close to 1 indicating the npLCM has efficiently used stick-breaking to strike the balance between estimation bias and variance. In Scenario II, compared to the pLCM, the npLCM produced smaller MSEs for C at all η_o values, where the advantage is largely explained by smaller biases.

The npLCM also produces 95% credible intervals (CI) with near-nominal empirical coverage rates. For example, Appendix Table 1 highlights that the substantial under-coverages ($< 80\%$) only occurred when assuming LI. Because of the extra variability from the informative priors on the TPRs, the CIs are conservative in Scenario I for both models. The over-coverage of both models is largely due to the assumed variances in the TPR parameters.

5. Analysis of PERCH Data

The Pneumonia Etiology Research for Child Health (PERCH) study is a standardized and comprehensive case-control study that has enrolled over 4,000 patients hospitalized for severe or very severe pneumonia and over 5,000 controls selected randomly from the community, frequency-matched on age in each month. Its objective is to evaluate etiologic agents causing severe and very severe pneumonia among hospitalized children aged 1-59 months in seven low and middle income countries with a significant burden of childhood pneumonia and a range of epidemiologic characteristics (Levine *and others*, 2012). More details about the PERCH design can be found in Deloria-Knoll *and others* (2012).

Using preliminary PERCH data from one site, we focus on PCR assays on nasopharyngeal

(NP) specimens for cases and controls. We illustrate the advantage of the npLCM in accounting for measurement LD, with improved efficiency, better empirical fit, and more valid etiology estimation. Results for all seven countries will be reported elsewhere upon study completion. Included in the current illustrative analysis are NPPCR data for 592 cases and 613 controls on 6 species of pathogens (abbreviations and full names in Appendix F).

We have compared the population etiology fractions, $\boldsymbol{\pi}$, estimated separately by two methods: the pLCM and the npLCM with subclass truncation level $K^* = 10$. The npLCM results are similar when larger values of K^*s are used. As discussed in Section 2.5, we need expert prior knowledge on the sensitivities for posterior inference by both methods; we used elicited sensitivity priors from laboratory experts with range $50 \sim 99.5\%$. Given our focus on 6 leading pathogens, we include the "other" cause for completeness as discussed in Section 2.3.

Strong LD is present in the analyzed data, with statistically significant log odds ratios observed for 6 out of 30 pathogen pairs among cases and controls, ranging from -2.47 (s.e.: 1.01) to 1.67 (s.e.: 0.39), and also by noting that under LI assumption we expect $0.05 \times 30 = 1.5(\pm 2.4)$ such pairs. In addition, as noted in Berger and Sellke (1987) and Dunson and Xing (2009), the interval null hypothesis $H_0 : \max_k \eta_k > 1 - \epsilon$, is useful for detecting deviations from the point null of exact LI. We choose $\epsilon = 0.05$ based on experience in simulation studies and to permit deviations from LI so small as to be non-significant in our application. The largest subclass weight is estimated with 95% CI $(0.65, 0.89)$ for the cases and $(0.27, 0.75)$ for the controls, again suggesting non-negligible LD in the data.

Figure 4(a) compares the results obtained from the pLCM (left boxes) and npLCM (right boxes). Each vertical box-and-whisker shows the marginal posterior mean (solid dot) and median (segment within box), with 95% credible interval (CI; between whisker endpoints) and 50% CI (between top and bottom box edges) of the etiologic fraction for each pathogen listed on the horizontal bar. The two approaches produce differences in the posterior means of etiologic fractions between −9.9% and 9.5%. Half of the largest increase in RHINO, from 5.2 (95% CI: 0.3 ∼ 17.9)% to 15.1 (5.9 ∼ 27.5)% is explained by its increase in predicted individual etiologies for cases with the NPPCR data 000010 (Figure 5, bottom left).

The npLCM also provides a better empirical fit. We have compared the posterior predictive distributions (Gelman *and others*, 1996) of the frequencies of common NP measurement patterns to the observed values separately in the cases and the controls. Among cases (left panel in Figure 4(b)), for example, the npLCM adequately predicts the observed frequencies of the 2nd and 6th most common case patterns (000001: 12.5%; 000100: 5.4%) by accounting for the negative associations of RSV with other pathogens with the log odds ratios ranging from −3.37 to −0.12 (3 out of 5 statistically significant).

We also examine the pairwise associations by calculating the standardized LOR difference (SLORD) defined to be the observed LOR for a pair of measurements minus the mean LOR for the predictive distribution value from each method divided by the standard deviation of the LOR predictive distribution. Appendix Figure 3 shows 9 pairs of pathogens that have statistically significant deviations of model predicted LORs from the observed ones for the pLCM and only 3 pairs for the npLCM. A blank cell indicates a good model prediction for the observed pairwise LOR ($| SLORD |< 2$). The npLCM achieves a better fit by noting that, for a well-fitting model, we expect 1.5(±2.4) non-blank cells. The associations between pairs of measurements (HMPV-A/B,RSV) and (PARA-1,RSV) are not expected in either model, although npLCM does better. In the PERCH study, we observed that seasonal variation in the rate of detection for RSV, HMPV-A/B and PARA-1 were out of phase and seasonal regression adjustment, discussed elsewhere, can sensibly account for this negative association.

6. Discussion

In this paper, we derived and tested a nested pLCM to allow for local dependence among binary observations given class membership. We compare this new model with a special case that depends on local independence in terms of asymptotic and finite sample size properties. The npLCM reduces large-sample estimation bias, retains the estimation efficiency and gives more valid inferences about $\boldsymbol{\pi}$ than the pLCM. The npLCM family also makes it possible to study the sensitivity of scientific findings to the LI assumption when pLCM is used.

The model first approximates the probability distribution for the control measurements by a mixture of product Bernoulli distributions with mixing weights penalized towards a mixture with few components. The estimated control dependence structure is then applied to the case model with modifications that represent the influence of the latent disease state. This valuable information from controls may help distinguish competing models for the local dependence among measurements and warrants further studies (e.g. Albert *and others*, 2001).

In the analysis of 6 leading pathogens from the PERCH study, RSV is estimated to be the most prevalent infectious cause of childhood pneumonia except the "other" category. That evidence is robust to the LD assumption. Accounting for LD structure leads to notable increases in etiologic fraction estimates of two pathogens and decrease in another. The npLCM can also integrate extra measurements of better qualities, for example, blood culture tests for bacteria that have near-perfect specificities to inform TPRs and improve efficiency (Hammitt *and others*, 2012).

In this paper, we assumed a single primary cause for each pneumonia case in the npLCM. This framework can be extended from a single to multiple causes by using a latent vector for case i, $\boldsymbol{I}_i \in \{0,1\}^J$, where $I_{ij} = 1$ indicates pathogen j is a component cause. For example, Hoff (2005) used Dirichlet process mixture models to identify multiple abnormal genomic locations that are jointly responsible for each case's disease, but using case-only data with LI assumption. Alternatively, one can place an exponential penalty on the number of causes (e.g., Zhang and Liu, 2007), or use conditionally specified models $[I_{ij}=1 \mid \boldsymbol{I}_{i[-j]}, \boldsymbol{X}_{ij}]$ to characterize the interactions among pathogens (Besag, 1974), where $\boldsymbol{X}_{ij}$ is a vector of covariates predictive for pathogen j being a cause in case i. The computational cost to fit these models increases substantially because the search space for the latent vector $\boldsymbol{I}_i$ expands exponentially in J. Development of efficient and reliable posterior sampling algorithms can allow investigators to assess the evidence of multiple-pathogen etiologies as more measurements accrue.

A second extension of the npLCM family motivated by PERCH is to allow the etiology distribution and false positive rates to depend upon covariates. For example, season, child's age and HIV status. Regression versions for npLCM have been implemented and are the subject of current study.

Finally, Wu *and others* (2015) derived the pLCM model to be used with a combination of direct measurements of cases' lungs without error and peripheral measures of cases and controls with error. With gold-standard data, this analyses is an example of supervised learning. The npLCM can be used in the same way. In the PERCH application, we rely entirely on peripheral samples, so the analyses is largely unsupervised. Robustness of inferences to model assumptions is critical.

Supplementary Material

The reader is referred to the on-line Supplementary Materials for technical appendices, additional simulations referenced in Sections 2, 3, 4 and 5 at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank the members of the larger PERCH Study Group for discussions that helped shape the statistical approach presented herein, and the study participants. We also thank the members of PERCH Expert Group who provided external advice.

References

Albert, P.S., McShane, L.M. and Shih, J.H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57**(2), 610–619.

Anderson, Thomas W. (1954). On estimation of parameters in latent structure analysis. *Psychometrika* **19**(1), 1–10.

Bartholomew, David J, Knott, Martin and Moustaki, Irini. (2011). *Latent variable models and factor analysis: A unified approach*, Volume 904. John Wiley & Sons.

Berger, James O and Sellke, Thomas. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American statistical Association* **82**(397), 112–122.

Besag, Julian. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 192–236.

Bollen, Kenneth A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology* **53**(1), 605–634.

Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**(4), 434–455.

Brooks, Steve, Gelman, Andrew, Jones, Galin and Meng, Xiao-Li. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Deloria-Knoll, M., Feikin, D.R., Scott, J.A.G., O'Brien, K.L., DeLuca, A.N., Driscoll, A.J., Levine, O.S. *and others*. (2012). Identification and selection of cases and controls in the pneumonia etiology research for child health project. *Clinical Infectious Diseases* **54**(suppl 2), S117–S123.

Dendukuri, Nandini, Hadgu, Alula and Wang, Liangliang. (2009). Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in medicine* **28**(3), 441–461.

Dunson, D.B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**(487), 1042–1051.

Efron, Bradley. (2015). Frequentist accuracy of bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(3), 617–646.

Espeland, Mark A and Handelman, Stanley L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*, 587–599.

Feikin, D.R., Scott, J.A.G. and Gessner, B.D. (2014). Use of vaccines as probes to define disease burden. *The Lancet* **383**(9930), 1762–1770.

Garrett, E.S. and Zeger, S.L. (2000). Latent class model diagnosis. *Biometrics* **56**(4), 1055–1067.

Gelfand, Alan E and Kottas, Athanasios. (2002). A computational approach for full nonparametric bayesian inference under dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **11**(2), 289–305.

Gelman, Andrew, Meng, Xiao-Li and Stern, Hal. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**(4), 733–760.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**(2), 215–231.

Gustafson, Paul. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*, Volume 140. CRC Press.

Haberman, Shelby J. (1979). *Analysis of Qualitative Data. Vol. 2, New Developments*. Academic Press.

Hagenaars, Jacques A. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research* **16**(3), 379–405.

Hammitt, L.L., Kazungu, S., Morpeth, S.C., Gibson, D.G., Mvera, B., Brent, A.J., Mwarumba, S., Onyango, C.O., Bett, A., Akech, D.O. *and others*. (2012). A preliminary study of pneumonia etiology among hospitalized children in kenya. *Clinical Infectious Diseases* **54**(suppl 2), S190–S199.

Harper, Dean. (1972). Local dependence latent structure models. *Psychometrika* **37**(1), 53–59.

Hoff, Peter D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics* **61**(4), 1027–1036.

Hofmann, Thomas. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* **42**(1-2), 177–196.

Ishwaran, Hemant and James, Lancelot F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**(453), 161–173.

Jokinen, Jukka and Scott, J Anthony G. (2010). Estimating the proportion of pneumonia attributable to pneumococcus in kenyan adults: latent class analysis. *Epidemiology (Cambridge, Mass.)* **21**(5), 719–725.

Jones, G., Johnson, W.O., Hanson, T.E. and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* **66**(3), 855–863.

Kroenke, Kurt and Spitzer, Robert L. (2002). The phq-9: a new depression diagnostic and severity measure. *Psychiatr Ann* **32**(9), 1–7.

Lazarsfeld, Paul F. (1950). *The logical and mathematical foundations of latent structure analysis*, Volume IV, Chapter The American Soldier: Studies in Social Psychology in World War II. Princeton, NJ: Princeton University Press, pp. 362–412.

Lazarsfeld, Paul F. (1959). *Latent structure analysis*, Chapter Psychology: A Study of Science. New York: McGraw-Hill, pp. 476–543.

Lazarsfeld, Paul Felix, Henry, Neil W and Anderson, Theodore Wilbur. (1968). *Latent structure analysis*. Houghton Mifflin Boston.

Levine, O.S., O'Brien, K.L., Deloria-Knoll, M., Murdoch, D.R., Feikin, D.R., DeLuca, A.N., Driscoll, A.J., Baggett, H.C., Brooks, W.A., Howie, S.R.C. *and others*. (2012). The pneumonia etiology research for child health project: A 21st century childhood pneumonia etiology study. *Clinical Infectious Diseases* **54**(suppl 2), S93–S101.

Lord, Frederic M. (1952). The relation of test score to the trait underlying the test. *ETS Research Bulletin Series* **1952**(2), 517–549.

McDonald, Roderick P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology* **34**(1), 100–117.

Papaspiliopoulos, Omiros and Roberts, Gareth O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika* **95**(1), 169–186.

Pepe, Margaret Sullivan and Janes, Holly. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8**(2), 474–484.

Qu, Y. and Hadgu, A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* **93**(443), 920–928.

Qu, Yinsheng, Tan, Ming and Kutner, Michael H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52**(3), 797–810.

Sethuraman, Jayaram. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**(2), 639–650.

Torrance-Rynard, Vicki L and Walter, Stephen D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in medicine* **16**(19), 2157–2175.

Vacek, Pamela M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 959–968.

Walker, Stephen G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36**(1), 45–54.

White, Halbert. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**(1), 1–25.

Wu, Zhenke, Deloria-Knoll, Maria, Hammitt, Laura L and Zeger, Scott L. (2015). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, DOI: 10.1111/rssc.12101.

Xu, Huiping and Craig, Bruce A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics* **65**(4), 1145–1155.

Yang, Ilsoon and Becker, Mark P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics*, 948–958.

Zhang, Yu and Liu, Jun S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**(9), 1167–1173.

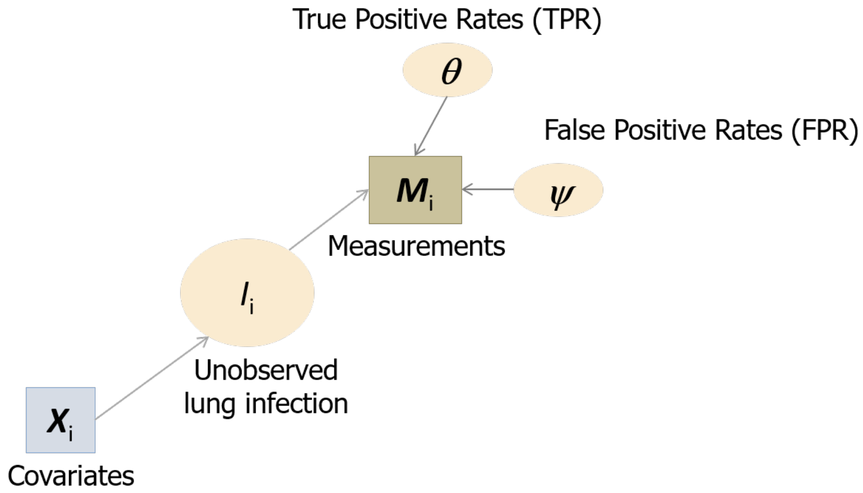
True Positive Rates (TPR)
θ
False Positive Rates (FPR)
M_i
ψ
Measurements
I_i
Unobserved lung infection
X_i
Covariates

Fig. 1: Schematic of PERCH problem for an individual case i. I_i for latent state; $\boldsymbol{M}_i$ for multivariate binary measurements; $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ for true- and false-positive rates; $\boldsymbol{X}_i$ for covariates. Quantities in rectangles are observed and those in ovals are unknown.

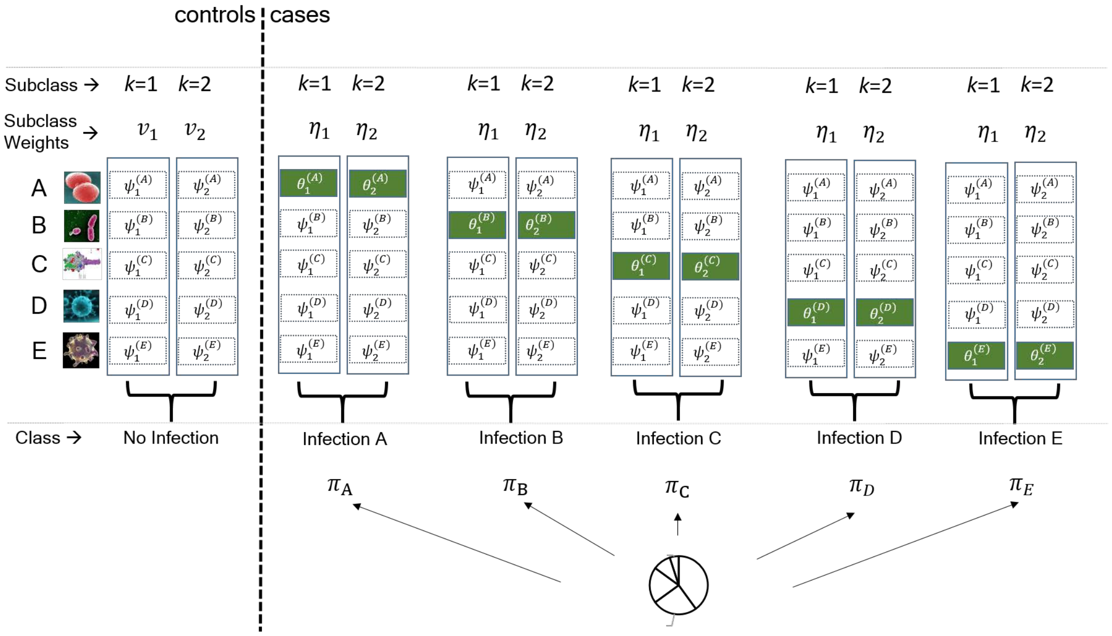
controls
cases
Subclass →
k=1 k=2
Subclass Weights →
ν_1 ν_2
η_1 η_2
A
B
C
D
E
$\psi_1^{(A)}$ $\psi_2^{(A)}$
$\psi_1^{(B)}$ $\psi_2^{(B)}$
$\psi_1^{(C)}$ $\psi_2^{(C)}$
$\psi_1^{(D)}$ $\psi_2^{(D)}$
$\psi_1^{(E)}$ $\psi_2^{(E)}$
$\theta_1^{(A)}$ $\theta_2^{(A)}$
$\theta_1^{(B)}$ $\theta_2^{(B)}$
$\theta_1^{(C)}$ $\theta_2^{(C)}$
$\theta_1^{(D)}$ $\theta_2^{(D)}$
$\theta_1^{(E)}$ $\theta_2^{(E)}$
Class →
No Infection
Infection A
Infection B
Infection C
Infection D
Infection E
π_A
π_B
π_C
π_D
π_E

Fig. 2: Borrowing measurement characteristics from controls to cases using $K = 2$ subclasses for each disease class. Five pathogens (A to E) are measured in this example.

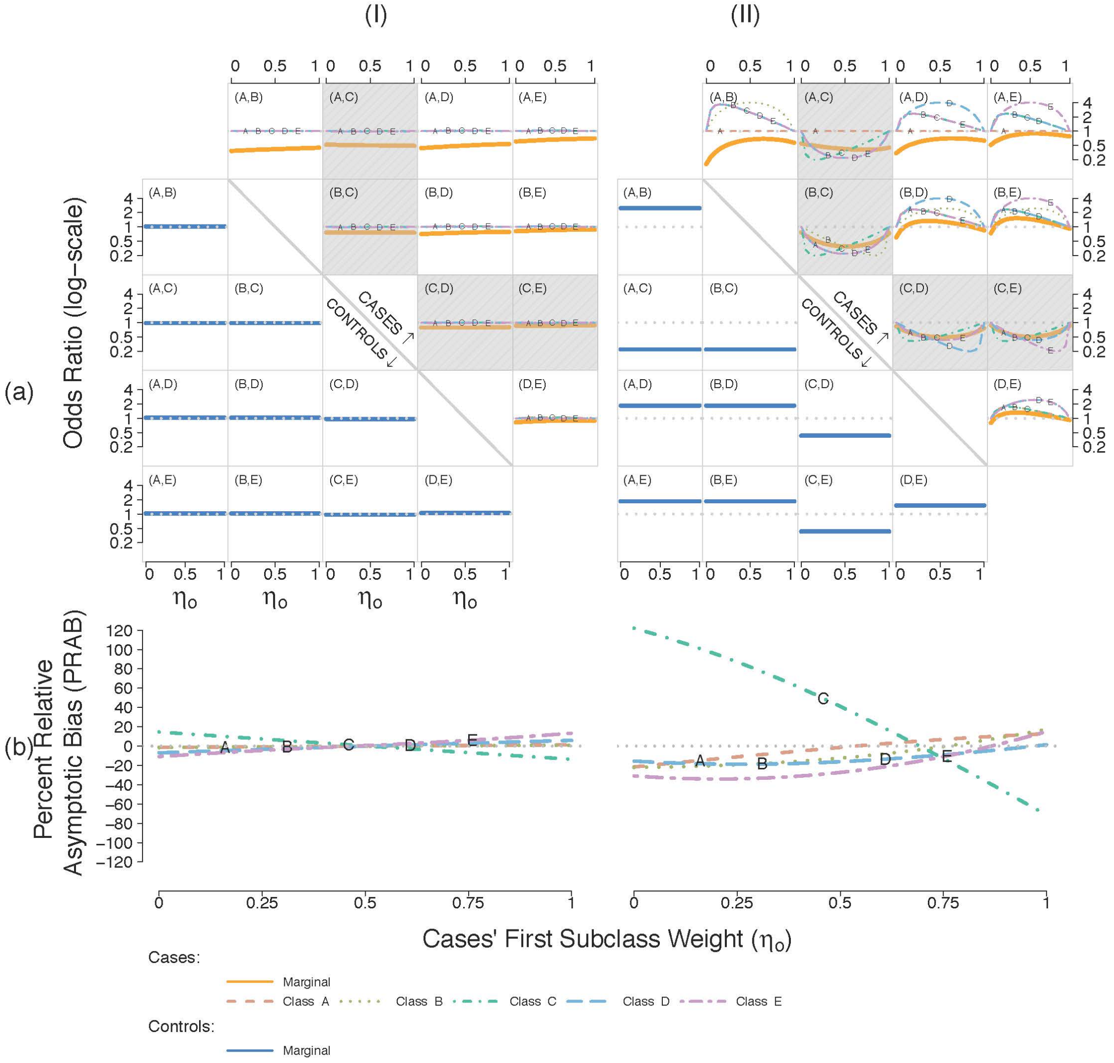

Fig. 3: In Scenario **I-II**,

Top (a): The true data generating mechanism summarized by pairwise odds ratios for cases (upper right, solid lines) and controls (lower left, solid lines) as the cases' first subclass weight (η_0) increases from 0 to 1. The pairwise odds ratios *within* each case class are shown by non-solid lines (legend at bottom). Pairwise independence is represented by the dotted horizontal lines for reference. The correlations of C with others are highlighted in shaded cells.

Bottom (b): Percent relative asymptotic bias (PRAB) for estimating etiology fractions using working local independence (LI) model when the truth varies across a range of local dependence (LD) settings parametrized by η_o.

Table 1: Comparison of Bayes estimates of etiology fractions obtained from npLCM (sf np) and pLCM (p). *Top*: direct bias of the posterior mean ($\overline{\pi}_\ell - \pi_{o,\ell}$); *Bottom*: ratio of mean squared errors (MSE) for pLCM vs npLCM. All numbers are averaged across $1,000$ replications and multiplied by 100.

			Truth: Cases' First Subclass Weight (η_o)				
		Model	0	0.25	0.5	0.75	1
	Class		100×Bias(Standard Error)				
I	A	np	-0.8(0.1)	-0.5(0.1)	-0.2(0.1)	0.1(0.1)	0.4(0.1)
		p	-1.1(0.1)	-0.7(0.1)	-0.3(0.1)	-0.1(0.1)	0.0(0.1)
	B	np	-0.6(0.1)	-0.5(0.1)	-0.4(0.1)	-0.5(0.1)	-0.4(0.1)
		p	-0.6(0.1)	-0.5(0.1)	-0.6(0.1)	-0.5(0.1)	-0.3(0.1)
	C	np	1.4(0.1)	0.7(0.1)	-0.1(0.1)	-0.9(0.1)	-1.7(0.1)
		p	1.9(0.1)	0.8(0.1)	-0.1(0.1)	-0.9(0.1)	-1.9(0.1)
	D	np	-0.1(0.1)	0.1(0.1)	0.4(0.1)	0.6(0.1)	0.9(0.1)
		p	-0.2(0.1)	0.3(0.1)	0.5(0.1)	0.7(0.1)	1.1(0.1)
	E	np	0.0(0.1)	0.2(0.1)	0.3(0.1)	0.6(0.1)	0.7(0.1)
		p	0.0(0.0)	0.2(0.1)	0.5(0.1)	0.8(0.1)	1.0(0.1)
II	A	np	4.5(0.1)	5.7(0.1)	5.5(0.1)	3.5(0.1)	0.5(0.1)
		p	-3.6(0.1)	0.2(0.1)	3.0(0.1)	5.0(0.1)	5.5(0.1)
	B	np	-5.7(0.1)	-6.1(0.1)	-4.9(0.1)	-2.1(0.1)	1.9(0.1)
		p	**-13.5**(0.1)	-8.5(0.1)	-4.3(0.1)	-0.3(0.1)	4.1(0.1)
	C	np	4.5(0.1)	4.1(0.1)	2.1(0.1)	-1.0(0.1)	-6.2(0.1)
		p	**26.2**(0.1)	**13.6**(0.1)	3.7(0.1)	-4.8(0.1)	**-12.5**(0.0)
	D	np	-2.4(0.1)	-2.5(0.1)	-1.7(0.1)	-0.4(0.1)	2.1(0.1)
		p	-5.8(0.0)	-3.3(0.1)	-1.6(0.1)	-0.2(0.1)	1.3(0.1)
	E	np	-1.0(0.0)	-1.3(0.0)	-1.0(0.0)	-0.1(0.1)	1.6(0.1)
		p	-3.2(0.0)	-1.9(0.0)	-0.8(0.1)	0.4(0.1)	1.7(0.1)
	Class		100×Ratio of MSE(Standard Error)				
I	A		94(6)	115(7)	100(6)	92(6)	91(6)
	B		105(6)	94(6)	98(6)	96(6)	91(6)
	C		114(7)	101(6)	93(5)	93(5)	90(5)
	D		104(6)	105(6)	96(6)	97(6)	108(7)
	E		97(4)	96(6)	124(7)	98(6)	119(7)
II	A		82(4)	25(1)	47(2)	115(6)	221(12)
	B		516(11)	177(5)	80(3)	62(4)	140(8)
	C		2379(77)	711(26)	131(7)	268(13)	357(8)
	D		397(14)	152(6)	94(5)	79(4)	60(4)
	E		357(13)	151(6)	102(5)	95(6)	82(5)

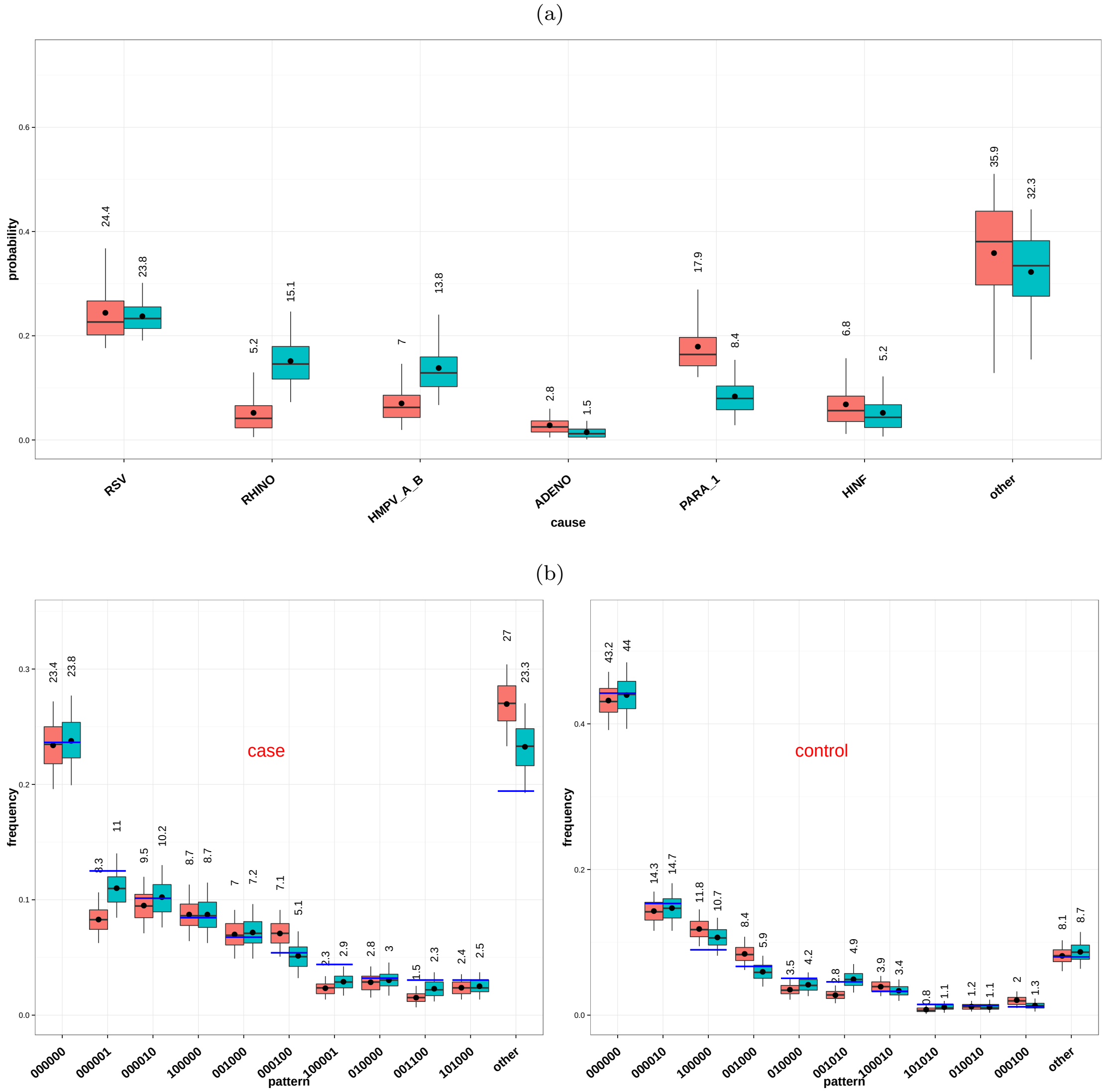

Fig. 4: *Top*: Comparison of the posterior distributions of $\boldsymbol{\pi}$ between the pLCM (left) and npLCM (right); The numbers above are the posterior means (×100). *Bottom*: Posterior predictive distributions (PPD) for 10 most frequent multivariate binary patterns separately for cases (left panel) and controls (right panel). The observed frequencies are overlayed as short segments across pairs of box-and-whiskers; the means of the PPDs (×100) are shown above them in actual numbers.

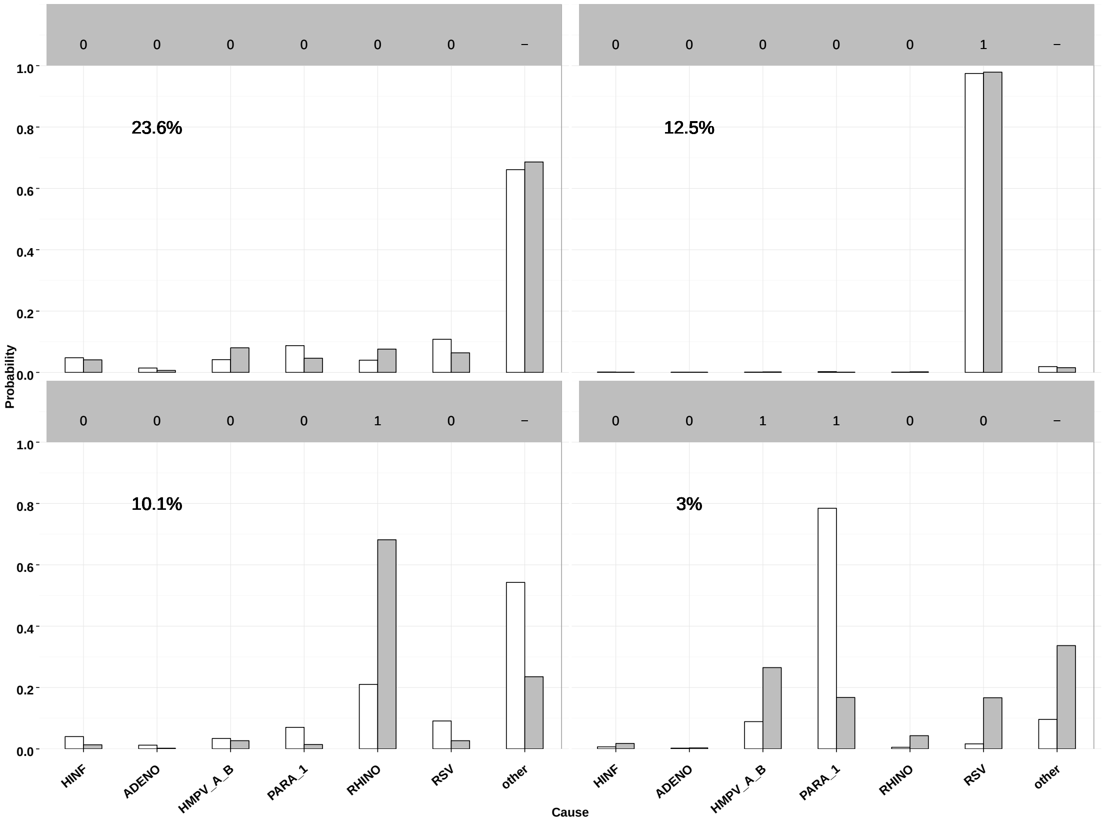

Fig. 5: Individual etiology distribution estimated by the empirical distribution of MCMC samples of the disease class indicator. Here four NPPCR data patterns are represented by the binary codes at the top (no measurements on "other" causes hence left as "-"), with its observed frequency marked beneath. The height of a bar represents the probability of a case caused by each of the 7 causes labelled on the horizontal axis. For each cause, paired bars compare the estimates from the pLCM (left) and the npLCM (right); Extra predictions are in Appendix Figure 4.