Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data

BO ZHANG*, ZHEN CHEN, PAUL S. ALBERT

Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892, USA
bo.zhang@nih.gov

SUMMARY

High-dimensional biomarker data are often collected in epidemiological studies when assessing the association between biomarkers and human disease is of interest. We develop a latent class modeling approach for joint analysis of high-dimensional semicontinuous biomarker data and a binary disease outcome. To model the relationship between complex biomarker expression patterns and disease risk, we use latent risk classes to link the 2 modeling components. We characterize complex biomarker-specific differences through biomarker-specific random effects, so that different biomarkers can have different baseline (low-risk) values as well as different between-class differences. The proposed approach also accommodates data features that are common in environmental toxicology and other biomarker exposure data, including a large number of biomarkers, numerous zero values, and complex mean–variance relationship in the biomarkers levels. A Monte Carlo EM (MCEM) algorithm is proposed for parameter estimation. Both the MCEM algorithm and model selection procedures are shown to work well in simulations and applications. In applying the proposed approach to an epidemiological study that examined the relationship between exposures to 62 individual polychlorinated biphenyl (PCB) congeners and endometriosis, we identified a highly significant overall effect of PCB concentrations on the risk of endometriosis.

Keywords: Categorical data; Chemical exposure biomarkers; Latent variables; Monte Carlo EM algorithm; Random effects.

1. INTRODUCTION

In many epidemiological studies, it is of interest to study the association between biological markers and a binary disease outcome. When only a few biomarkers are considered, traditional statistical modeling strategies can be directly used. Challenges arise, however, when a large number of biomarkers are involved. This paper proposes joint latent class models for high-dimensional biomarker data where numerous biomarkers and disease prevalence are linked through underlying latent classes of disease risk.

As an example, we consider an epidemiological study that assessed the relationship between exposures to 62 individual polychlorinated biphenyl (PCB) congeners (chemicals with a similar structure) and endometriosis, an estrogen-dependent gynecological disease (Buck Louis and others, 2005). These PCBs

*To whom correspondence should be addressed.
can be divided into 3 groups: estrogenic PCBs, antiestrogenic PCBs, and other PCBs. The study recruited a cohort of 79 eligible participants (between 18 and 40 years of age) with serum measurements who underwent laparoscopy. Participants were classified as having endometriosis or not based on laparoscopic visualization. For every participant, the concentration of each of the 62 PCB congeners was quantified by dual-column gas chromatography with electron capture from her serum specimens. The primary goal of the study was to investigate the relationship between exposure to PCB congeners and the risk of endometriosis. The top and bottom panels in Figure 1 show boxplots for each of 62 PCBs for subjects with and without endometriosis, respectively. Differences between the 2 profiles appear subtle and may be complex. For example, out of 62 PCB markers, approximately 82% have higher median exposure levels for subjects with endometriosis as compared with the subjects without endometriosis. Therefore, it is desirable to develop a flexible approach, which can identify potentially complex associations between high-dimensional PCB exposures and endometriosis diagnosis.

Fig. 1. Boxplots of each of 62 PCBs for subjects with and without endometriosis.
Several challenges arise in analyzing these data. (1) "High dimensionality of biomarker data." With 62 PCB congeners and a relatively small number of participants ($I = 79$), it may be difficult to assess the complex relationship between PCB concentrations and the risk of endometriosis with traditional statistical techniques such as logistic regression. (2) "Collective effect of numerous biomarkers." PCB congeners may act on the risk of endometriosis in complex ways. For example, small changes in all or many PCBs may act simultaneously in increasing the risk of endometriosis. Alternatively, larger changes in a few PCBs may influence risk. It is desired to develop methodology that allows for these flexible association structures in risk modeling. (3) "Zero exposure level." For many of the PCBs, there is a sizable proportion of zero values. These values can result from either true zero concentrations of PCB congeners or PCB values that are too small to detect in the laboratory. Moreover, the occurrence of zero PCB values is potentially related to the means of the positive values. Figure 2A shows the rate of observing nonzero PCB concentrations versus their means. It suggests that the probability of observing nonzero exposures increases with the nonzero PCB means in a linear fashion. (4) "Variance depending on mean." Figure 2B shows the standard deviations (SDs) of the observed nonzero PCB concentration levels versus their means. It suggests that the PCB SDs are related in a linear fashion to their corresponding means. These features of the data need to be appropriately accounted for in the analysis.

In this paper, we propose joint latent class models with random effects to account for the aforementioned data analysis challenges. In particular, we consider a class of joint latent class models for associating the pattern of high-dimensional exposure data with the risk of disease through the incorporation of latent classes. Specifically, patterns in PCB exposures and risk of endometriosis are associated through assumed latent classes, which are shared between the 2 modeling components. The latent class structure allows for the flexible identification of important patterns of exposure, which can place women at high risk for endometriosis. The formulation of the joint latent class models for semicontinuous biomarkers (i.e. continuous value with numerous zeros) link the patterns in the biomarker and the risk of disease through unobserved latent classes. The pattern in the mean biomarker values varies by class and is allowed to change across biomarkers by the incorporation of random effects (Laird and Ware, 1982). Our work has close connection to the latent class models discussed in Bartholomew (1987) and Bandeen-Roche.

Fig. 2. Features of the PCBs and endometriosis data. (A) The rate of nonzero PCB concentrations versus the mean of nonzero PCB concentrations and (B) SD of nonzero PCB concentrations versus the mean of nonzero PCB concentrations. Each point in the plot represents a PCB congener. The solid lines are the fitted lines from a linear regression; the dotted lines are the fitted lines from a LOWESS smoother (Cleveland, 1979).
and others (1997). These models, which are recognized as categorical variable analogs to factor analysis, postulate the joint distribution of multiple observed discrete variables as a mixture of a few distributions conditional on the value of a latent categorical variable. However, our model is for the disease biomarker data where the disease indicator is binary and the biomarker levels are all semicontinuous. Another related approach is the finite Gaussian mixture model discussed in Fraley and Raftery (2002) and McLachlan and Peel (2002). The finite Gaussian mixture model assumes that continuous data follow a finite normal mixture model with each component corresponding to a latent cluster or latent class. This model can be viewed as a latent variable model since the unobserved categorical variable characterizes classification of observations into clusters. But unlike the proposed latent class joint model, the finite Gaussian mixture model is a model-based clustering technique for biomarkers where clustering is not informed by the disease outcome. Our work is also related to Lin and others (2002) who used latent classes to link longitudinal biomarker data and event times.

The remainder of the paper is organized as follows. Section 2 presents the joint latent class models for high-dimensional semicontinuous biomarker data and the binary disease outcome. Section 3 derives the log likelihood of the proposed models and discusses strategies for estimation. In Section 4, we apply the proposed model to data from a study examining the effects of environmental PCB concentrations on the risk of endometriosis. In Section 5, simulation studies are performed to demonstrate the performance of the estimation and model selection methods. Finally, we conclude in Section 6.

2. JOINT LATENT CLASS MODELS

2.1 Joint latent class models with 2 latent classes

Let \( Y_i \) be a binary variable representing disease status for participant \( i, i = 1, 2, \ldots, I \), where \( Y_i = 1 \) indicates the \( i \)th participant has disease and \( Y_i = 0 \) indicates no disease. Denote \( X_{ij} \) as the level of the \( j \)th biomarker measured from the \( i \)th participant, \( j = 1, 2, \ldots, J \). In the motivating example of PCB concentrations and endometriosis, \( Y_i \) represents endometriosis status of the \( i \)th participant and \( X_{ij} \) represents the concentration level of the \( j \)th PCB congener measured from the \( i \)th participant. Let \( L_i \) denote the latent disease risk class for the \( i \)th participant. We first consider the case with only 2 latent classes: low-risk class, denoted by \( L_i = 0 \), and high-risk class, denoted by \( L_i = 1 \). In Section 2.2, we extend the model to the case of more than 2 classes using an ordinal latent variable. Let \( \pi_1 = P(L_i = 1) \) be the probability of a participant belonging to the high-risk class and let \( \pi_0 = 1 - \pi_1 \). The probability of disease given the latent disease class is modeled as:

\[
\logit P(Y_i = 1|L_i) = \beta_0 + \beta_1 L_i, \quad \beta_1 > 0, \tag{2.1}
\]

where \( \beta_0 \) and \( \beta_1 \) are regression coefficients characterizing the relationship between the latent variable and the disease risk. In (2.1), the coefficient \( \beta_1 \) is the positive increase of the log odds of disease prevalence from low-risk to high-risk class, and the positivity constraint ensures model identifiability related to the “label switching” problem common in latent class modeling.

Each biomarker level \( X_{ij} \), conditional on the latent class, is treated as a semicontinuous measurement (Olsen and Schafer, 2001) and can be denoted by 2 variables:

\[
U_{ij} = \begin{cases} 
1, & \text{if } X_{ij} \neq 0 \\
0, & \text{if } X_{ij} = 0 
\end{cases} \quad \text{and} \quad V_{ij} = \begin{cases} 
X_{ij}, & \text{if } X_{ij} \neq 0 \\
\text{irrelevant,} & \text{if } X_{ij} = 0 
\end{cases}, \tag{2.2}
\]

where \( U_{ij} \) is the binary nonzero biomarker value indicator and \( V_{ij} \) is the value of the biomarker when it is nonzero. We model \( V_{ij} \) using the normal distribution:

\[
V_{ij}|L_i, b_j \sim N(\mu_{ij}(L_i, b_j), g^2(\mu_{ij}(L_i, b_j), \xi)), \tag{2.3}
\]
where $\mu_{ij}(L_i, b_j) = E(V_{ij}|L_i, b_j)$ is the conditional mean of $V_{ij}$ that depends on latent class indicator $L_i$ and random effects vector $b_j$, and $g^2$ is a variance function that depends on $\xi$, a vector of variance parameters. In (2.3), we allow the variance of nonzero biomarker values to depend on the mean biomarker value through the variance function $g^2(\mu_{ij}(L_i, b_j), \xi)$. In the application in Section 4, the SD of PCB concentration is proportional to PCB mean (see Figure 2B). Thus, in this case, it is reasonable to let $g^2(\mu_{ij}(L_i, b_j), \xi) = \xi^2 \mu_{ij}^2(L_i, b_j)$.

For the mean structure of each biomarker, we assume

$$\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i + b_{j0} + b_{j1} L_i,$$

where $a_0$ and $a_1$ are parameters defining the mean levels of the biomarkers in the low- and high-risk groups, and $b_j = (b_{j0}, b_{j1})' \sim N((0, 0)', \begin{pmatrix} \sigma_0^2 & \rho \sigma_0 \sigma_1 \\ \rho \sigma_0 \sigma_1 & \sigma_1^2 \end{pmatrix})$ is a random effects vector that follows a multivariate normal distribution. The inclusion of the latent classes $L_i$ allows the multiple biomarkers from the same participant to be correlated, while the random effects $b_{j0}$ and $b_{j1}$ allow for individual biomarkers to deviate from the overall mean. If the 2 random effects degenerate, or equivalently $\sigma_0^2 = \sigma_1^2 = \rho = 0$, then $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i$, which means the average concentration of each biomarker is $a_0$ for participants in the low-risk class and $a_0 + a_1$ for participants in the high-risk class (see Figure 3A). If $\sigma_0^2 \neq 0$ and $\sigma_1^2 = \rho = 0$, then $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i + b_{j0}$. In this case, there is a biomarker-specific shift $b_{j0}$ in the average concentration of each biomarker, but the difference in biomarker in classes is the same across biomarker (see Figure 3B). With 2 nondegenerated random effects $b_{j0}$ and $b_{j1}$, or equivalently $\sigma_0^2 \neq 0$ and $\sigma_1^2 \neq 0$, we allow both biomarker-specific baseline shift $b_{j0}$ and biomarker-specific difference $a_1 + b_{j1}$ between the 2 classes to vary across the $J$ biomarkers (see Figure 3C). When $a_1$ is large and $\sigma_1^2$ is small, the modeling approach corresponds to the situation where cumulative PCB effects transmit risk. In this case, a simple logistic regression model where the independent variable is the sum of PCB exposures describes the association. However, when $a_1$ is small and $\sigma_1^2$ is large, the association is complex and a simple cumulative logistic model will miss this association. We illustrate this with a simulation study in the supplementary material available at Biostatistics online.

Finally, the model for the positive biomarker indicator $U_{ij}$ is

$$\logit P(U_{ij} = 1|L_i, b_j) = \eta_0 + \eta_1 h(\mu_{ij}(L_i, b_j)),$$

in which $h(\mu_{ij}(L_i, b_j))$ can be specified according to data features. The inclusion of $\mu_{ij}(L_i, b_j)$ in function $h$ allows for a flexible yet parsimonious way to model the probability of a positive biomarker concentration as a function of the mean of the nonzero biomarker values. The parameter $\eta_1$ and function $h$ in (2.5) characterize the relationship between the occurrence of a positive measurement and the mean of the measurement given that it is positive. In our motivating example, there is a positive linear relationship between the occurrence of a positive value and the mean of these values (see Figure 2A). This motivates the specification $h(\mu_{ij}(L_i, b_j)) = \mu_{ij}(L_i, b_j)$.

Equations (2.1–2.5) characterize joint latent class models with random effects for high-dimensional biomarker data and disease prevalence, where the disease prevalence and biomarker levels are linked and jointly modeled through binary latent classes among participants. In the next section, we extend the model to the situation in which the latent risk classes are ordinal.

### 2.2 Joint latent class models with multiple latent classes

When more than 2 latent classes are considered, a variety of parameterizations could be pursued to extend the proposed 2-class model. A natural specification would be to treat each latent class as a categorical
Latent class models for joint analysis of disease biomarker data

Fig. 3. Simulated patterns of mean PCB concentrations with different random effects specifications in $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i + b_{j0} + b_{j1} L_i$: (A) $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i$; (B) $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i + b_{j0}$; and (C) $\mu_{ij}(L_i, b_j) = a_0 + a_1 L_i + b_{j0} + b_{j1} L_i$. Plots are based on simulated data with 2 latent disease class (0 for low-risk class and 1 for high-risk class) and 62 PCB congeners. Parameter values: $a_0 = 0.015$, $a_1 = 0.035$, $b_{j0} \sim N(0, 0.005)$, and $b_{j1} \sim N(0, 0.01)$, with $\text{cov}(b_{j0}, b_{j1}) = 0$.

variable and to treat the latent classes as dummy variables in the disease prevalence model (2.1), the mean structure model (2.4), and the model for a positive biomarker value (2.5). This would require introducing fixed and random effects of dimension equal to the number of latent classes in each of the model components. Although rich and flexible, this modeling strategy is complicated and computationally demanding.
Further, this approach does not exploit the inherent ordinal nature of disease severity. As a result, we build a parsimonious extension by treating \( L_i = 0, 1, \ldots, K - 1 \) as an ordinal variable in (2.1–2.5). In essence, we model both the logit of disease prevalence and biomarker levels with linear functions of the ordinal latent class variable \( L_i \). For example, if 3 latent classes are considered, the model then yields probabilities

\[
P(Y_i = 1|L_i = 0) = \exp(\beta_0)/[1 + \exp(\beta_0)],
\]

\[
P(Y_i = 1|L_i = 1) = \exp(\beta_0 + \beta_1)/[1 + \exp(\beta_0 + \beta_1)],
\]

\[
P(Y_i = 1|L_i = 2) = \exp(\beta_0 + 2\beta_1)/[1 + \exp(\beta_0 + 2\beta_1)].
\]

These correspond to the disease prevalence of low-, middle-, and high-risk classes, respectively. Similarly, the means of biomarker levels for the 3 latent classes are \( \alpha_0, \alpha_0 + \alpha_1 \), and \( \alpha_0 + 2\alpha_1 \), respectively. We focus on this parameterization to achieve a parsimonious modeling framework for the case of multiple disease classes.

### 2.3 Covariates for joint latent class models

Participant-specific or biomarker-specific covariates can be incorporated into the disease probability (2.1) and biomarker distribution (2.3) as follows:

\[
\logit P(Y_i = 1|L_i) = \beta_0 + \beta_1 L_i + \mathbf{W}_i \mathbf{y}, \quad \beta_1 > 0 \tag{2.6}
\]

and

\[
V_{ij}|L_i, \mathbf{b}_j \sim N(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), \mathbf{g}^2(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), \xi)), \tag{2.7}
\]

where \( \mathbf{W}_i \) is a vector of participant-specific covariates that may affect disease prevalence and \( \mathbf{Z}_{ij} \) is a vector of variance covariates whose components are either participant specific or biomarker specific. In our example described in Section 1, there are 3 groups of biomarkers among the 62 PCB congeners: estrogenic PCBs, antiestrogenic PCBs, and other PCBs. Thus, in Section 4, we incorporate the PCB grouping information into (2.7) with \( \mathbf{Z}_{ij} \) comprising 2 indicator variables.

Covariates can also be introduced into (2.5):

\[
\logit P(U_{ij} = 1|L_i, \mathbf{b}_j) = \eta_0 + \eta_1 h(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), \mathbf{T}_{ij}, \xi), \tag{2.8}
\]

with covariate vector \( \mathbf{T}_{ij} \) and parameter vector \( \xi \).

### 3. Maximum likelihood estimation

#### 3.1 The likelihood

The proposed joint latent class model (with extended covariates in Section 2.3) contains unknown parameters: \( \beta_0, \beta_1, \mathbf{y} \) in (2.6), \( \xi \), \( \alpha_0, \alpha_1, \sigma_0, \sigma_1, \rho \) in (2.7) and (2.4), and \( \eta_0, \eta_1, \xi \) in (2.8). If we define the probability that a participant belongs to the \( k \text{th} \) class as \( \pi_k = P(L_i = k) \), \( k = 0, 1, \ldots, K - 1 \), then the components of \( \mathbf{y} = (\pi_0, \pi_1, \ldots, \pi_{K-1})' \) are also unknown parameters. Let \( y_i, u_{ij}, v_{ij}, w_i, z_{ij}, \) and \( t_{ij} \) be the observed values of \( y_i, u_{ij}, v_{ij}, w_i, z_{ij}, \) and \( t_{ij} \), respectively, and let \( \theta = (\beta_0, \beta_1, \mathbf{y}', \xi', \alpha_0, \alpha_1, \sigma_0, \sigma_1, \rho, \eta_0, \eta_1, \xi', \xi')'. \) Note that the latent variable \( L_i \) is crossed with the random effects \( \mathbf{b}_j \) in the proposed model. Therefore, the likelihood of the proposed model is given by

\[
\int \cdots \int \sum_{L_i=0}^{K-1} \cdots \sum_{L_i=0}^{K-1} \prod_{i=1}^{J} \pi_{L_i} \left( \frac{e^{\beta_0 + \beta_1 L_i + \mathbf{W}_i \mathbf{y}}}{1 + e^{\beta_0 + \beta_1 L_i + \mathbf{W}_i \mathbf{y}}} \right)^{y_i} \left( \frac{1}{1 + e^{\beta_0 + \beta_1 L_i + \mathbf{W}_i \mathbf{y}}} \right)^{1-y_i} \times \prod_{j=1}^{J} \left( \frac{e^{\beta_0 + \eta_1 h(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), t_{ij}, \xi)}}{1 + e^{\beta_0 + \eta_1 h(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), t_{ij}, \xi)}} \right)^{u_{ij}} \left( \frac{1}{1 + e^{\beta_0 + \eta_1 h(\mu_{ij}(L_i, \mathbf{b}_j, \mathbf{Z}_{ij}), t_{ij}, \xi)}} \right)^{1-u_{ij}} \times f_{V_{ij}|L_i, \mathbf{b}_j}(v_{ij}) \prod_{j=1}^{J} f_{\mathbf{Y}_j}(\mathbf{b}_j) \mathbf{db}_j, \tag{3.1}
\]
where \( f_{\mu_j}(l, b, \zeta) \) is the probability density function of a univariate normal distribution with mean \( \mu_j(l, b) \) and variance \( \sigma_j^2 g^2(\mu_j(l, b), \zeta) \) as in (2.3), \( \mu_j(l, b) = a_0 + \alpha_1 l_1 + b_{j0} + b_{j1} L \) as in (2.4), and \( f_b \) is the bivariate normal distribution probability density function of \( b_j \) as described in Section 2.

### 3.2 Monte Carlo EM algorithm

We consider an expectation-maximization (EM) algorithm (Dempster and others, 1977) as follows. We treat the latent class variable \( L = (L_1, \ldots, L_J)' \) and the random effects \( b = (b_1, \ldots, b_J)' \) as missing data in the EM algorithm. We denote \( X_{obs} \) as the observed data, \( Z_{mis} = (L, b)' \) as the missing data, and let \( Y_{com} = (X_{obs}, Z_{mis})' \). At the \( (t+1) \)th iteration of the EM algorithm, the E-step involves calculation of the \( Q \)-function \( Q(\theta(t)) = E[\log f(y_{com}|\theta) | x_{obs}, \theta(t)] = \int \log f(y_{com}|\theta) f(z_{mis}|x_{obs}, \theta(t)) dz_{mis} \), where \( \theta(t) \) denotes the parameter value from the \( t \)th iteration, \( f(z_{mis}|x_{obs}, \theta(t)) \) is the conditional distribution of \( (L, b) \) given the observed data and \( \theta(t) \), and \( f(y_{com}|\theta) \) is the full likelihood. The M-step involves maximizing \( Q(\theta(t)) \) with respect to \( \theta \) to yield the new update \( \theta(t+1) \). The process is iterated from a starting value \( \theta(0) \) to convergence. Under regularity conditions, the value at convergence maximizes (3.1).

The dimensionality of integration and summations in likelihood function (3.1) increases with the number of study participants and biomarkers. As a result, the integration in the E-step is intractable with numerical integration techniques. Therefore, we conduct estimation using Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990). In this algorithm, we form the Monte Carlo approximation to the integrals in the likelihood function are of high dimension. An alternative approach is to use the Metropolis–Hastings algorithm, as in McCulloch (1997), to sample from the conditional distribution using the density of latent variables as the proposal distribution. Here, we use the Metropolis–Hastings algorithm but with an adaptive proposal distribution. To generate \( N \) values \( (l(n), b(n)) \), \( n = 1, 2, \ldots, N \), from the conditional distribution \( f(l, b|x_{obs}, \theta(t)) \) using the Metropolis–Hastings algorithm, we start from the starting values \( (l(0), b(0)) \) with \( f(l(0), b(0)) > 0 \), and then, at the \( nth \) iteration of the MCEM step, sample a candidate \( (l^*, b^*) \) from \( f(t)(l, b) \). Let \( (l(n+1), b(n+1)) = (l^*, b^*) \) with probability \( \min(1, R) \), where \( R = \frac{f(l^*, b^*|x_{obs}, \theta(t)) f(t)(l(n), b(n))}{f(l(n), b(n)|x_{obs}, \theta(t)) f(t)(l^*, b^*)} \), and \( (l(n+1), b(n+1)) = (l(n), b(n)) \) otherwise. The proposal distribution at the \( t \)th iteration of the EM algorithm is constructed to be \( \bar{f}(t)(l, b) = \prod_{i=1}^{l} \tilde{f}(l_i) \prod_{j=1}^{j} \tilde{f}(b_j) \), where \( \tilde{f}(l_i) \) is the empirical distribution of \( l_i \)'s sampled from the \( (t-1) \)th MCEM iteration and \( \tilde{f}(b_j) = \tilde{f}(b_{j0}) \tilde{f}(b_{j1}) \) is the product of 2 Student's \( t \)-distribution centered, respectively, at the means of \( b_{j0} \)'s and \( b_{j1} \)'s sampled from the \( (t-1) \)th MCEM iteration (Gelman and others, 1995). The proposal distribution presented here dynamically incorporates the obtained information of conditional distribution \( f(l, b|x_{obs}, \theta(t)) \) from the previous MCEM step and thus greatly improves the performance of the Metropolis–Hastings Markov chain. The starting values \( (l(0), b(0)) \) of the Metropolis–Hastings algorithm are chosen to be the \( t \) M-step estimates of the previous step of the MCEM algorithm. We run 100 steps in the EM algorithm and 10 000 Monte Carlo iterations in each MCEM step. To incorporate the constraint \( \beta_1 > 0 \) in the MCEM algorithm, we let \( \beta_1 = e^{\beta_1^*} \) with \( -\infty < \beta_1^* < +\infty \) and implement the MCEM algorithm with respect to the new parameter \( \beta_1^* \).

### 3.3 Empirical Bayes estimates of the random effects

Our estimation procedure consists of 2 steps: frequentist maximum likelihood estimation described in Section 3.2 for the unknown fixed parameters and empirical Bayes estimation for the random effects \( b_j \) in
this section. The maximum likelihood estimation procedure is not Bayesian since no priors are assigned to the fixed effects. In obtaining the empirical Bayes estimates of the random effects, the random effects distributions are treated as priors. This is commonly done in the literature of mixed effects models; see Laird and Ware (1982) for reference.

Viewing the bivariate normal distribution of \( b_j \) as a prior distribution, approximate inferences for \( b_j \) can be obtained by calculating a posterior mean \( E(b_j|\theta, x_{\text{obs}}) \) and covariance matrix \( \text{var}(b_j|\theta, x_{\text{obs}}) \) with the unknown parameter vector \( \theta \) replaced by its maximum likelihood estimates \( \hat{\theta} \). Consider the posterior expectation of some function \( M \) of \( b_j \), which are of the form \( E(M(b_j)|\theta, x_{\text{obs}}) = \int M(b_j)p(b_j|\theta, x_{\text{obs}}) \, db_j \). The approximate inference for \( b_j \) can be carried out by calculating \( E(M(b_j)|\theta, x_{\text{obs}}) \) with some form of \( M \). For instance, \( M(b_j) = b_j \) results in the posterior mean \( E(b_j|\theta, x_{\text{obs}}) \). The calculation is simply a by-product of the MCEM algorithm.

4. AN ANALYSIS OF THE ENDOMETRIOSIS AND PCB CONCENTRATION DATA

We fit the joint latent class models to data from the epidemiological study that examines the effect of environmental PCB concentrations on the risk of endometriosis. Details of the study have been previously reported in Buck Louis and others (2005) and have been summarized in Section 1. The primary interests of the data analysis are in examining the association between complex patterns of PCB concentrations and the risk of endometriosis. Using the methodology developed in the previous sections, we fit a series of models with different numbers of latent classes. As discussed in Section 2, there is evidence to suggest \( h(\mu_{ij}(L_i, b_j, Z_{ij}), T_{ij}, \xi) = \mu_{ij}(L_i, b_j, Z_{ij}) \) (Figure 2A) and \( g^2(\mu_{ij}(L_i, b_j, Z_{ij}), \xi) = \xi^2\mu_{ij}^2(L_i, b_j, Z_{ij}) \) (Figure 2B) when fitting the models. Two dummy variables \( I_{j,e} \) and \( I_{j,a} \) are created to represent estrogenic PCBs (\( I_{j,e} = 1 \) if a PCB is estrogenic; \( I_{j,e} = 0 \) otherwise) and antiestrogenic PCBs (\( I_{j,a} = 1 \) if a PCB is antiestrogenic; \( I_{j,a} = 0 \) otherwise), respectively. Thus, the PCB mean structure is:

\[
\mu_{ij}(L_i, b_j, Z_{ij}) = (a_0 + a_{0,e}I_{j,e} + a_{0,a}I_{j,a} + b_j0) + (a_1 + a_{1,e}I_{j,e} + a_{1,a}I_{j,a} + b_{j1})L_i.
\]

The MCEM algorithm was implemented as described in Section 3 and standard errors (SEs) were estimated using the nonparametric bootstrap method with 500 bootstrap samples (Efron and Tibshirani, 1993).

Table 1 shows parameter estimates and SEs of 4 random effects latent class models, with the number of latent classes being 2, 3, 4, and 5, respectively. The table also reports Bayesian information criterion (BIC) type ICQ values (see Ibrahim and others, 2008 and the supplementary material available at Biostatistics online) for the 4 candidate models. The model selection criterion ICQ chose the 3-class model as the best among the competing ones. This model divides participants into 3 disease classes according to their risks with respect to endometriosis: 71.40% of the participants belonged to the low-risk group, 25.27% to the middle-risk group, and 3.33% to the high-risk group. The endometriosis prevalences in the 3 groups were estimated as 30.42%, 53.16%, and 74.66%, respectively. Since \( \beta_1 \) was restricted to be positive, we performed a one-sided Wald test of \( H_0 : \beta_1 = 0 \) versus \( H_1 : \beta_1 > 0 \) and obtained a \( p \) value of 0.0186. Tests of whether \( a_1 \neq 0 \), \( a_1 + a_{1,e} \neq 0 \), or \( a_1 + a_{1,a} \neq 0 \) examine whether the mean biomarker values vary with latent risk class in the other, estrogenic, and antiestrogenic PCBs, respectively; these 3 tests were all highly significant (less than 0.0001). These tests indicate that the PCB concentration pattern is associated with the risk of endometriosis, and that the overall means of PCBs are significantly different between classes. The parameter estimates \( \hat{\eta}_0 = -0.8892 \) and \( \hat{\eta}_1 = 42.8010 \) provide information on the probabilities of observing zero PCB concentration levels. The positive \( \eta_1 \) estimate indicates that, for a given PCB and participant, the probability of observing a positive PCB level increases with the mean. We found a statistically significant relationship between the pattern of PCB concentrations and the risk of endometriosis using the proposed model. We now estimate features of the pattern in each of the risk groups.
Table 1. Analysis of the endometriosis and PCB concentrations data: parameter estimates (SEs) and values of model selection criterion ICQ from joint latent class models with different number of latent classes†

<table>
<thead>
<tr>
<th>Estimates (SEs)</th>
<th>Two classes</th>
<th>Joint latent class models with</th>
<th>Four classes</th>
<th>Five classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>$-0.8052 (0.3329)$</td>
<td>$-0.8272 (0.3757)$</td>
<td>$-1.2187 (0.3929)$</td>
<td>$-1.2587 (0.4661)$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$0.8401 (0.4054)$</td>
<td>$0.9539 (0.4579)$</td>
<td>$0.8342 (0.3652)$</td>
<td>$0.6480 (0.2779)$</td>
</tr>
<tr>
<td>$a_0$</td>
<td>$0.0159 (0.0024)$</td>
<td>$0.0152 (0.0028)$</td>
<td>$0.0130 (0.0023)$</td>
<td>$0.0117 (0.0023)$</td>
</tr>
<tr>
<td>$a_{0,e}$</td>
<td>$0.0090 (0.0014)$</td>
<td>$0.0046 (0.0008)$</td>
<td>$0.0042 (0.0006)$</td>
<td>$0.0025 (0.0005)$</td>
</tr>
<tr>
<td>$a_{0,a}$</td>
<td>$-0.0043 (0.0007)$</td>
<td>$-0.0048 (0.0005)$</td>
<td>$-0.0053 (0.0010)$</td>
<td>$-0.0054 (0.0010)$</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$0.0289 (0.0045)$</td>
<td>$0.0218 (0.0042)$</td>
<td>$0.0156 (0.0024)$</td>
<td>$0.0132 (0.0026)$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$0.0177 (0.0020)$</td>
<td>$0.0075 (0.0009)$</td>
<td>$0.0077 (0.0013)$</td>
<td>$0.0079 (0.0014)$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$-0.0075 (0.0010)$</td>
<td>$-0.0067 (0.0012)$</td>
<td>$-0.0036 (0.0006)$</td>
<td>$-0.0027 (0.0003)$</td>
</tr>
<tr>
<td>$\tilde{\eta}$</td>
<td>$0.7263 (0.8600)$</td>
<td>$0.8361 (1.1114)$</td>
<td>$0.8800 (0.845)$</td>
<td>$0.9013 (1.147)$</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>$0.0088 (0.0011)$</td>
<td>$0.0063 (0.0012)$</td>
<td>$0.0053 (0.0010)$</td>
<td>$0.0048 (0.0007)$</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>$0.0190 (0.0033)$</td>
<td>$0.0127 (0.0021)$</td>
<td>$0.0094 (0.0015)$</td>
<td>$0.0087 (0.0013)$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$0.7788 (0.0749)$</td>
<td>$0.7259 (0.1195)$</td>
<td>$0.6599 (0.1148)$</td>
<td>$0.6464 (0.1240)$</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>$-0.8556 (0.0750)$</td>
<td>$-0.8892 (0.1625)$</td>
<td>$-0.9025 (0.1427)$</td>
<td>$-0.7844 (0.1195)$</td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>$37.9655 (3.5367)$</td>
<td>$42.8010 (4.6698)$</td>
<td>$45.2033 (4.4552)$</td>
<td>$45.6119 (4.6333)$</td>
</tr>
</tbody>
</table>

Model selection

| $IC_Q$         | $578.1685$ | $566.7670$ | $587.4758$ | $607.9914$ |

†The analysis is also performed under the assumption $L_i = 0$ for any $i$; this amounts to assuming no association between PCBs and endometriosis. The parameter estimates are $\beta_0 = -0.5537$ (SE 0.2339), $\hat{\alpha}_0 = 0.0225 (0.0035)$, $\hat{\alpha}_{0,e} = 0.0172 (0.0021)$, $\hat{\alpha}_{0,a} = -0.0113 (0.0017)$, $\tilde{\xi} = 0.6191 (0.0586)$, $\hat{\alpha}_0 = 0.0231 (0.0030)$, $\hat{\eta}_0 = -0.5858 (0.1217)$, $\hat{\eta}_1 = 35.5567 (3.5701)$. The $IC_Q$ value is 612.4074.

To investigate the features of the individual PCB congener concentration, we computed empirical Bayes estimators of random effects $\hat{b}_{j0}$ and $\hat{b}_{j1}$, $j = 1, 2, \ldots, 62$. The nonzero concentration levels for the $j$th PCB congener in the $k$th latent class can be estimated by replacing random effects with their empirical Bayes estimators and replacing unknown parameters with their estimates in (4.1). The estimated rate of observing zero concentration levels for the $j$th PCB can be similarly derived. Figures 4 shows, in the selected 3-class model, the nonzero concentration levels of each PCB congener. There is a positive relationship between PCB concentration mean and the risk class for all PCBs (i.e. the pattern in mean concentration for all PCBs is highest for high-risk participants and lowest for low-risk participants). This suggests that there is a combined contribution of all 62 PCB concentrations in assessing the risk of endometriosis. The differences in the PCB patterns from the low risk to the high risk are largest for estrogenic PCBs among the 3 PCB groups. Moreover, we see a reverse pattern for the probability of obtaining a zero PCB. The rate of zero varied across PCB concentrations and across classes (range 0–0.7).

The inferences for $\beta_1$, $\alpha_1$, $\alpha_{1,e}$, and $\alpha_{1,a}$ were similar across the different models (2- through 5-class models), suggesting that even though the data are best described by the 3-class model, important inferences are robust to model misspecification. Further, empirical Bayes estimators of concentration patterns in each of the risk groups (similar to Figure 4, but for 2-, 4-, 5-class models) showed similar results to those presented in Figure 4. Namely, concentration levels increased with an increasing risk
Fig. 4. Estimated means of nonzero PCB concentrations by the three-class latent class model in the analysis of endometriosis and PCB concentrations data. The numbers on the lines are the latent class: 0 for low-risk class, 1 for middle-risk class, and 2 for high-risk class. One PCB is both estrogenic and antiestrogenic.

group for all PCBs, and the increasing concentration levels were largest for estrogenic markers as compared to other markers (data not shown).

The joint latent class model provided strong evidence for an association between high-dimensional PCB exposure and the risk and endometriosis. The primary goal of the proposed latent class model is to provide a way to associate high-dimensional exposure data with the risk of endometriosis. The fact that the nature of the association ($\beta_1$ in Table 1 and parameters that characterize PCB biomarker exposures) is similar across the models with different numbers of latent class provides reassurance about the robustness of important inference to the number of latent classes. That said, the evidence for 3 latent classes of risk does suggest that there may be 3 types of exposure patterns that are associated with the risk of endometriosis. The fact that overall patterns are higher in the high-risk group and are lower in the other groups suggests that what distinguishes the different groups are small differences in a majority of PCBs.

5. SIMULATIONS

To evaluate the performance of the MCEM algorithm and the model selection criterion $\text{IC}_Q$ for the joint latent class modeling approach, we conducted a simulation study by generating data similar to that in our application. We generated data according to a 2-, 3-, 4-, or 5-class model. The true parameter values were chosen to be similar to the estimated values of the corresponding models in our application example. Also, to create a scenario similar to the application, we used 79 participants and 62 biomarkers in each simulation. A total of 500 data sets were generated from each of the 4 latent class models. We assumed the number of latent classes was known and then fitted the model to the generated data. The results are reported in Table 2. For each model, the table presents the average and SD of the point estimates.
Table 2. Simulation results for evaluating parameter estimation with MCEM in the joint latent class models, based on 500 simulation data sets†

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Average estimate</th>
<th>Estimate SD</th>
<th>Truth</th>
<th>Average estimate</th>
<th>Estimate SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two latent classes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-0.8052</td>
<td>-0.8097</td>
<td>0.3188</td>
<td>-0.8272</td>
<td>-0.8267</td>
<td>0.3823</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.8401</td>
<td>0.8382</td>
<td>0.3912</td>
<td>0.9539</td>
<td>0.9504</td>
<td>0.3681</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>0.0159</td>
<td>0.0159</td>
<td>0.0038</td>
<td>0.0152</td>
<td>0.0150</td>
<td>0.0035</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>0.0090</td>
<td>0.0092</td>
<td>0.0029</td>
<td>0.0046</td>
<td>0.0045</td>
<td>0.0010</td>
</tr>
<tr>
<td>( a_{0,a} )</td>
<td>-0.0043</td>
<td>-0.0043</td>
<td>0.0011</td>
<td>-0.0048</td>
<td>-0.0046</td>
<td>0.0014</td>
</tr>
<tr>
<td>( a_{1,a} )</td>
<td>0.0289</td>
<td>0.0287</td>
<td>0.0068</td>
<td>0.0218</td>
<td>0.0218</td>
<td>0.0049</td>
</tr>
<tr>
<td>( a_{0,e} )</td>
<td>0.0177</td>
<td>0.0178</td>
<td>0.0044</td>
<td>0.0075</td>
<td>0.0075</td>
<td>0.0020</td>
</tr>
<tr>
<td>( a_{1,e} )</td>
<td>-0.0075</td>
<td>-0.0077</td>
<td>0.0048</td>
<td>-0.0067</td>
<td>-0.0069</td>
<td>0.0016</td>
</tr>
<tr>
<td>( \xi )</td>
<td>0.7263</td>
<td>0.7214</td>
<td>0.1079</td>
<td>0.8361</td>
<td>0.8353</td>
<td>0.1525</td>
</tr>
<tr>
<td>( \sigma_1 )</td>
<td>0.0088</td>
<td>0.0086</td>
<td>0.0026</td>
<td>0.0063</td>
<td>0.0061</td>
<td>0.0019</td>
</tr>
<tr>
<td>( \sigma_2 )</td>
<td>0.0198</td>
<td>0.0198</td>
<td>0.0056</td>
<td>0.0127</td>
<td>0.0125</td>
<td>0.0024</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.7788</td>
<td>0.7817</td>
<td>0.1324</td>
<td>0.7259</td>
<td>0.7231</td>
<td>0.1521</td>
</tr>
<tr>
<td>( \eta_0 )</td>
<td>-0.8556</td>
<td>-0.8548</td>
<td>0.1383</td>
<td>-0.8892</td>
<td>-0.8934</td>
<td>0.1475</td>
</tr>
<tr>
<td>( \eta_1 )</td>
<td>37.9655</td>
<td>38.1683</td>
<td>5.0883</td>
<td>42.8010</td>
<td>43.0890</td>
<td>5.3585</td>
</tr>
<tr>
<td>( \pi_0 )</td>
<td>0.6924</td>
<td>0.6914</td>
<td>0.0538</td>
<td>0.7140</td>
<td>0.7126</td>
<td>0.0837</td>
</tr>
<tr>
<td>( \pi_1 )</td>
<td></td>
<td></td>
<td></td>
<td>0.2527</td>
<td>0.2539</td>
<td>0.0376</td>
</tr>
<tr>
<td><strong>Four latent classes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-1.2187</td>
<td>-1.2157</td>
<td>0.5768</td>
<td>-1.2587</td>
<td>-1.2604</td>
<td>0.6220</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.8342</td>
<td>0.8363</td>
<td>0.3995</td>
<td>0.6480</td>
<td>0.6513</td>
<td>0.2804</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>0.0130</td>
<td>0.0128</td>
<td>0.0032</td>
<td>0.0117</td>
<td>0.0120</td>
<td>0.0030</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>0.0042</td>
<td>0.0043</td>
<td>0.0016</td>
<td>0.0025</td>
<td>0.0026</td>
<td>0.0009</td>
</tr>
<tr>
<td>( a_{0,a} )</td>
<td>-0.0053</td>
<td>-0.0054</td>
<td>0.0017</td>
<td>-0.0054</td>
<td>-0.0055</td>
<td>0.0010</td>
</tr>
<tr>
<td>( a_{1,a} )</td>
<td>0.0156</td>
<td>0.0157</td>
<td>0.0040</td>
<td>0.0132</td>
<td>0.0133</td>
<td>0.0029</td>
</tr>
<tr>
<td>( a_{0,e} )</td>
<td>0.0077</td>
<td>0.0079</td>
<td>0.0026</td>
<td>0.0079</td>
<td>0.0077</td>
<td>0.0024</td>
</tr>
<tr>
<td>( a_{1,e} )</td>
<td>-0.0036</td>
<td>-0.0038</td>
<td>0.0019</td>
<td>-0.0027</td>
<td>-0.0026</td>
<td>0.0007</td>
</tr>
<tr>
<td>( \xi )</td>
<td>0.8800</td>
<td>0.8798</td>
<td>0.1348</td>
<td>0.9013</td>
<td>0.8971</td>
<td>0.1622</td>
</tr>
<tr>
<td>( \sigma_1 )</td>
<td>0.0053</td>
<td>0.0052</td>
<td>0.0016</td>
<td>0.0048</td>
<td>0.0044</td>
<td>0.0012</td>
</tr>
<tr>
<td>( \sigma_2 )</td>
<td>0.0094</td>
<td>0.0095</td>
<td>0.0019</td>
<td>0.0087</td>
<td>0.0086</td>
<td>0.0023</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.6599</td>
<td>0.6651</td>
<td>0.1950</td>
<td>0.6464</td>
<td>0.6476</td>
<td>0.1260</td>
</tr>
<tr>
<td>( \eta_0 )</td>
<td>-0.9025</td>
<td>-0.9017</td>
<td>0.1487</td>
<td>-0.7844</td>
<td>-0.7837</td>
<td>0.1384</td>
</tr>
<tr>
<td>( \eta_1 )</td>
<td>45.2033</td>
<td>45.2051</td>
<td>7.3796</td>
<td>45.6119</td>
<td>45.3604</td>
<td>7.9095</td>
</tr>
<tr>
<td>( \pi_0 )</td>
<td>0.3564</td>
<td>0.3514</td>
<td>0.0535</td>
<td>0.3050</td>
<td>0.3014</td>
<td>0.0565</td>
</tr>
<tr>
<td>( \pi_1 )</td>
<td>0.4497</td>
<td>0.4490</td>
<td>0.0681</td>
<td>0.3435</td>
<td>0.3389</td>
<td>0.0763</td>
</tr>
<tr>
<td>( \pi_2 )</td>
<td>0.1604</td>
<td>0.1595</td>
<td>0.0835</td>
<td>0.1877</td>
<td>0.1908</td>
<td>0.0531</td>
</tr>
<tr>
<td>( \pi_3 )</td>
<td></td>
<td></td>
<td></td>
<td>0.1189</td>
<td>0.1179</td>
<td>0.0310</td>
</tr>
</tbody>
</table>

†The simulation is also performed under the assumption that there is no association between PCBs and endometriosis. As in other simulation studies, the true parameter values of simulated model is chosen to be the parameter estimates in the application example (see the Footnote of Table 1). The average of parameter estimates of \( \beta_0, a_0, a_{0,e}, a_{0,a}, \xi, \sigma_0, \eta_0, \) and \( \eta_1 \) are \(-0.5563 \text{ (SE 0.2037)}, 0.0222 \text{ (0.0038)}, 0.0177 \text{ (0.0051)}, -0.0113 \text{ (0.0021)}, 0.6214 \text{ (1.5353)}, 0.0230 \text{ (0.0039)}, -0.5954 \text{ (0.1581)}, \) and \(35.4520 \text{ (5.5767)}\), respectively.

We observed that the estimates were close to the true parameter values with reasonably small SDs. The parameters of most interest (e.g., \( \beta_1, a_1, a_{1,e}, \) and \( a_{1,a} \)) were nearly unbiased.

To examine the performance of model selection by IC\( _Q \), we fit models having 2, 3, 4, and 5 latent classes to each of the generated data sets and used BIC-type IC\( _Q \) to select the best model.
Supplementary Table 1 in the supplementary material available at *Biostatistics* online reports the frequencies of the selected models under each scenario. The correct selection rates are 92%, 81%, 79%, 77%, and 78%, respectively, for the no-association, 2-, 3-, 4-, and 5-class models. In addition, even when IC<sub>Q</sub> selected the wrong models, the majority of the misselected models are the adjacent ones (one fewer or more class). Furthermore, these simulations showed that even under a misspecified model (e.g. true model was a 3-class model and 2-class model was selected), estimates of β, α<sub>1</sub>, α<sub>1</sub>, e, and α<sub>1</sub> had little bias (data not shown).

6. Discussion

This paper proposed a joint latent class modeling approach for predicting a binary event from high-dimensional biomarker data. In this approach, the relationship between complex biomarker expression patterns and disease risk is established by incorporating ordinal latent risk classes, which link the 2 modeling components. Further, complex marker-specific differences are characterized through marker-specific random effects. The approach is flexible in that it easily incorporates a varying number of ordinal latent classes which will depend on the application of interest.

The joint latent class model presented in this paper supplements other methods for analyzing disease-biomarker data, especially when high-dimensionality and collective effects occur. Alternative analysis strategies have been proposed for analyzing the endometriosis and PCB concentration data set. Buck Louis and others (2005), Gennings and others (2010), and Herring (2010) analyzed the data with logistic–regression-based models, focusing on assessing the effect of individual PCB exposures on endometriosis risk. In general, these methods found only a small number of associations between the large number of individual PCB exposures and the risk of endometriosis. Our approach is more flexible in that we can identify complex patterns between the high-dimensional exposures and disease prevalence using latent class structure. The works of Buck Louis and others (2005), Gennings and others (2010), and Herring (2010) may have advantages when only a few biomarkers have very strong association with risk. However, when there is a large proportion of exposure measurements and each contributes a small amount to the risk, their methods may fail. Our approach is well suited for this latter case, while also allowing for the case where a few individual PCBs may contribute to the risk.

The application of the approach to the epidemiology study on PCB concentrations and the risk of endometriosis revealed some interesting findings from a substantive view point. First, we identified a highly significant overall relationship between PCB concentrations and the risk of endometriosis, a finding which was not found with other techniques. Second, the model allowed us to reveal features of the relationship, which were not available with other methods. Particularly, we found an overall positive association of all biomarkers with the risk of endometriosis, suggesting that all PCBs, to different degrees, are positively associated with the risk of endometriosis. Further, we identified that estrogenic PCBs may be particularly important in determining the risk of endometriosis.

The methodology was developed for a binary measure of risk assessment, which was consistent with the application presented here. However, the model could be adapted to other generalized linear model outcomes (e.g. counts, continuous, exponential). Our focus was primarily on assessing the role of high-dimensional biomarkers on disease risk and not on evaluating disease prediction per se. Future research will focus on comparing features of model performance such as the receiver operating characteristic curve and other measures of diagnostic accuracy between our proposed modeling framework and more conventional methods such as logistic regression with model selection.

ACKNOWLEDGMENTS

We thank associate editor and reviewers for their thoughtful and constructive comments, which have led to an improved article. We thank Drs Germaine Louis and Brian Whitcomb for their helpful comments and suggestions. We also thank the Center for Information Technology, the National Institutes of Health, for providing access to the high performance computational capabilities of the Biowulf Linux cluster. Conflict of Interest: None declared.

FUNDING

Intramural Research Program of the National Institutes of Health; Eunice Kennedy Shriver National Institute of Child Health and Human Development; National Institute of Environmental Health Sciences (Grant 1R01ES09044-01).

REFERENCES


[Received November 25, 2010; revised July 20, 2011; accepted for publication July 21, 2011]