Biostatistics (2008), **9**, 1, *pp.* 137–151
doi:10.1093/biostatistics/kxm018
Advance Access publication on June 12, 2007

Combining assays for estimating prevalence of human herpesvirus 8 infection using multivariate mixture models

RUTH M. PFEIFFER*

Biostatistics Branch, National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd, EPS/8030, Bethesda, MD 20892-7244, USA
pfeiffer@mail.nih.gov

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, TX 77843-3141, USA

WILLIAM WHEELER

Information Management Services Inc., Rockville, MD 20852, USA

DENISE WHITBY, SAM MBULAITEYE

Viral Epidemiology Branch, National Cancer Institute, DCEG, 6120 Executive Blvd, Bethesda, MD 20892-7244, USA

Summary

For many diseases, it is difficult or impossible to establish a definitive diagnosis because a perfect "gold standard" may not exist or may be too costly to obtain. In this paper, we propose a method to use continuous test results to estimate prevalence of disease in a given population and to estimate the effects of factors that may influence prevalence. Motivated by a study of human herpesvirus 8 among children with sickle-cell anemia in Uganda, where 2 enzyme immunoassays were used to assess infection status, we fit 2-component multivariate mixture models. We model the component densities using parametric densities that include data transformation as well as flexible transformed models. In addition, we model the mixing proportion, the probability of a latent variable corresponding to the true unknown infection status, via a logistic regression to incorporate covariates. This model includes mixtures of multivariate normal densities as a special case and is able to accommodate unusual shapes and skewness in the data. We assess model performance in simulations and present results from applying various parameterizations of the model to the Ugandan study.

Keywords: Diagnostic tests; Mixture models; Semi-nonparametric densities; Semiparametrics; Sensitivity; Specificity; Transformations.

*To whom correspondence should be addressed.

Published by Oxford University Press 2007.

1. INTRODUCTION

This article was motivated by the problem of estimating prevalence of infection with human herpesvirus 8 (HHV-8) and quantifying the effects of factors that influence prevalence of HHV-8 infection in several African populations. HHV-8, also called Kaposi's sarcoma (KS)–associated herpesvirus, is the generally accepted infectious cause of KS (Chang *and others*, 1994). Prevalence of HHV-8 infection and KS risk shows distinctive geographical variations. They are highest in sub-Saharan Africa, intermediate in Mediterranean countries, and lowest in the United States and northern Europe, where most KS cases are AIDS related (Martin, 2003). The modes of HHV-8 transmission are not well understood and appear to differ in high- and low-HHV-8 incidence regions. In African and Mediterranean countries, HHV-8 infection occurs during childhood, most likely via nonsexual modes of transmission, but in the United States and northern Europe, infection is essentially restricted to homosexual men and associated with sexual exposures (Martin, 2003).

Infection status with HHV-8 can be assessed by several serological assays, but it is impossible to establish a definitive diagnosis of infection status as a perfect gold standard measure does not exist. Standard statistical approaches to investigating factors that affect the prevalence of HHV-8 infection, such as contingency table analyses and logistic regression, employ an operational definition of "infected", namely that the optical density (OD) reading of a given assay exceeds a prespecified cutoff value. Cutoff values are commonly determined based on previous experimental results and a visual inspection of histograms of the OD readings for the given study. To avoid having to use predefined cutoff values for an operational definition of infected, Pfeiffer *and others* (2000) fitted a 2-component mixture model to the results of continuous assay readings to estimate the prevalence of infection with *Helicobacter pylori*, with the components corresponding to "infected" and "uninfected" subpopulations. Letting y to denote the OD readings for immunoglobulin G, the model that treats the true infection status as a latent variable has a density function given by

$$g(y) = (1 - p) f_0(y) + p f_1(y), \tag{1.1}$$

where f_1 is the density function corresponding to the test results for infected and f_0 for uninfected subjects. To assess the $\mathbf{x}$ factors that influence infection with *H. pylori*, Pfeiffer *and others* modeled the mixing probability p, the probability of being infected, by a logistic regression, $p = p(\mathbf{x}; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{x})/\{1 + \exp(\boldsymbol{\beta}'\mathbf{x})\}$.

We will extend model (1.1) to the multivariate setting to address a second problem, that is, how to combine the information from several assays that capture different but not necessarily independent indicators of HHV-8 infection. This is important because a combination of assays may provide a better diagnostic tool and yield more accurate estimates of prevalence. Developments in the area of multivariate mixtures are mostly concentrated on mixtures of multivariate normals because of their computational convenience (McLachlan *and others*, 2003). However, the fit of multivariate normal densities to the log-transformed assay readings in our data was poor, leading to unreasonably large estimates of the key parameter of interest p, the prevalence, in model (1.1). To improve the fit and obtain unbiased estimates of p, we developed more flexible models to accommodate skewness and heavy tails in the OD readings. First, we incorporate the parameters of the Box–Cox transformation into the model and estimation. Second, we use densities from a flexible class introduced by Gallant and Nychka (1987) as the mixture components. This model includes multivariate normal mixture models as a special case. Covariates are incorporated into the mixing probability via logistic regression.

In Section 2, we define the bivariate logistic mixture models before assessing the performance of the models in simulations (Section 3). We apply the models to data from a cross-sectional study of blood-borne transmission of HHV-8 in Ugandan children afflicted with sickle-cell anemia (Section 4). We compare prevalence estimates from the bivariate mixture model to prevalence estimates obtained by averaging estimates from univariate mixtures fitted to each assay separately. We also compare logistic estimates

from the bivariate mixture to those based on predefined cutoff values for the assays. Section 5 contains concluding remarks.

2. The data and model formulation

Inference is based on cross-sectional data on disease status and covariates at the time of examination. The data are $(\mathbf{Y}_j, \mathbf{X}_j)$ for $j = 1, \ldots, n$, where $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jl})$ and Y_{jk} denotes the observed measurement for the kth assay on the jth subject. The $p \times 1$ vector $\mathbf{X}_j$ contains measured covariates. In our application, 2 immunoassays were used to determine HHV-8 infection status, and hence $\mathbf{Y}_j = (Y_{j1}, Y_{j2})$. The first component, Y_{j1}, stands for measurements on the K8.1 assay that detects antibodies expressed during lytic infection, in serum. The second component, Y_{j2}, corresponds to serological measurements on the orf73 assay that tests for antibodies expressed during latency (Mbulaiteye *and others*, 2003).

2.1 *Mixture model*

There is an extensive literature on mixture models (McLachlan and Peel, 2000; Fraley and Raftery, 2002) that were developed to analyze the data that arise from 2 or more distinct data-generation processes. One major problem in mixture models concerns estimation of the number of component densities. However, in our application, we are confident that there are precisely 2 distinct populations that give rise to the data, an infected and an uninfected population. Thus, we assume that each person is in one of the 2 latent true infection states, which we label as state $I_j = 1$ (infected) and state $I_j = 0$ (uninfected) with $p = \mathrm{pr}(I_j = 1)$ for the jth subject. The probability of infection can depend on covariates $\mathbf{x}_j$, for example, through logistic regression models $\mathrm{pr}(I_j = 1|\mathbf{x}_j) = p(\mathbf{x}_j; \boldsymbol{\beta})$. Other parameterizations of p have been used in the context of survival data in Pfeiffer *and others* (2004). Given $\mathbf{x}_j$, the probability density function of $\mathbf{Y}_j$ is modeled as

$$g(\mathbf{y}_j|\mathbf{x}_j, \theta) = f(\mathbf{y}_j; \boldsymbol{\alpha}_0)\{1 - p(\mathbf{x}_j; \boldsymbol{\beta})\} + f(\mathbf{y}_j; \boldsymbol{\alpha}_1)p(\mathbf{x}_j; \boldsymbol{\beta}), \tag{2.1}$$

where $f(\cdot; \boldsymbol{\alpha}_0)$ is a bivariate parametric density function that corresponds to the OD measurements of the uninfected subpopulation and $f(\cdot; \boldsymbol{\alpha}_1)$ is the density of the OD readings for the infected subpopulation.

2.2 *Choice of component densities*

In previous work (Pfeiffer *and others*, 2000), we log-transformed the positive OD readings to remove asymmetry in the measurements and used normal densities for the components in model (2.1). However, the transformation was determined by visual inspection. We now incorporate the data transformation indexed by an unknown parameter $\boldsymbol{\lambda}$ into the likelihood. We choose the Box–Cox power transformation

$$y_i^{(\lambda_i)} = \begin{cases} (y_i^{\lambda_i} - 1)/\lambda_i, & \lambda_i \neq 0, \\ \log(y_i), & \lambda_i = 0, \end{cases} \tag{2.2}$$

for the components of $\mathbf{y} = (y_1, y_2)$. In the univariate density setting, coupling the Box–Cox power transformation to likelihood methods for normal mixtures has been used previously (e.g. Gutierrez *and others*, 1995). We use the same transformation for both components of the mixture for each assay, that is $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$. Transformation ignores the scale of the observed data, and thus for different values of $\boldsymbol{\lambda}$, the parameters of the component densities are not directly comparable (Carroll and Ruppert, 1981). Unlike parameters of the component densities, the mixing proportion $p(\mathbf{x}; \boldsymbol{\beta})$ in (2.1) has a physical meaning independent of the transformed scales, namely the percentage of the subpopulation with $\mathbf{x}$ that is infected.

To allow for more flexible shapes compared to normal densities, we choose the component densities in model (2.1) from a general class introduced by Gallant and Nychka (1987), called the semi-nonparametric densities. These densities have been studied, for example, by Zhang and Davidian (2001) and are defined as follows. Let $\varphi(\mathbf{y}, \boldsymbol{\mu}, \Sigma)$ be the bivariate normal density with mean vector $\boldsymbol{\mu}$, covariance matrix Σ, and argument $\mathbf{y} = (y_1, y_2)$. Then, the semi-nonparametric density is

$$f(\mathbf{y}, \boldsymbol{\mu}, \Sigma, \mathbf{a}) = \left(\sum_{0 \leqslant i+j \leqslant K} a_{ij} y_1^i y_2^j \right)^2 \varphi(\mathbf{y}, \boldsymbol{\mu}, \Sigma). \tag{2.3}$$

In (2.3), $K = 0$ reduces to the bivariate normal density. For $K \geqslant 1$, the polynomial part of the density has $d = (K+1)(K+2)/2$ distinct terms. Using the standard normal density and $z = \Sigma^{-1/2}(y - \boldsymbol{\mu})$ in (2.3), Zhang and Davidian (2001) showed that $\int f(\mathbf{y})\mathrm{d}\mathbf{y} = 1$ can be guaranteed by imposing the condition $a'Aa = 1$ on the coefficients $a = (a_{ij})$ of (2.3), where A is a matrix with (i, j)th element $E(U_1^{i_1+j_1})E(U_2^{i_2+j_2})$ for 2 standard normal variables U_1 and U_2 and the superscripts correspond to a_i and a_j. Because A is a positive definite matrix, there exists a matrix B such that $A = B^2$, and letting $c = Ba$, the constraint $a'Aa = 1$ reduces to $c'c = 1$. They represent c in terms of polar coordinates as $c_1 = \sin(\phi_1), c_2 = \cos(\phi_1)\sin(\phi_2), \ldots, c_d = \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-1})$, for $-\pi/2 \leqslant \phi < \pi/2$. Note that the dimension of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{d-1})$ is now $d-1$. The constraints are automatically satisfied, and standard unconstrained optimization techniques can be used to find the maximum likelihood estimates of the parameters.

Combining (2.2) and (2.3), the semi-nonparametric mixtures (model I) are

$$g(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \sum_{d=0}^{1} J(\boldsymbol{\lambda}) p^{1-d}(\mathbf{x}; \boldsymbol{\beta})\{1 - p(\mathbf{x}; \boldsymbol{\beta})\}^d f(\mathbf{y}^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_d, \Sigma_d, \boldsymbol{\phi}_d), \tag{2.4}$$

where $J(\boldsymbol{\lambda}) = y_1^{\lambda_1 - 1} y_2^{\lambda_2 - 1}$ is the Jacobian of the transformation $\mathbf{y} \to \mathbf{y}^{(\boldsymbol{\lambda})}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}_0, \Sigma_0, \boldsymbol{\phi}_0, \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\phi}_1)$. Following the recommendations by Zhang and Davidian (2001), we limit the model to $K \leqslant 2$. This model contains the mixture of multivariate normal densities as a special case. Two other special cases of interest are the situation of model I, $K = 0$, that results in a model that combines the Box–Cox transformation with multivariate normal densities, and a model that fits the mixture (2.4) with $K \geqslant 1$ to untransformed data (model II):

$$g(\mathbf{y}|\mathbf{x}, \theta) = \{1 - p(\mathbf{x}; \boldsymbol{\beta})\} f(\mathbf{y}, \boldsymbol{\mu}_0, \Sigma_0, \boldsymbol{\phi}_0) + p(\mathbf{x}; \boldsymbol{\beta}) f(\mathbf{y}, \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\phi}_1). \tag{2.5}$$

For a fixed number of mixing components, models (2.4) with increasing K are nested, and thus formal likelihood-ratio tests can be applied to assess goodness of fit. Testing models of increasing complexity allows one to choose a parsimonious but well-fitting model within the class. While model (2.5) is also nested within the more general model (2.4) for the same K, model I, $K = 0$, and (2.5) are not nested within each other, and we thus also use the Akaike information criterion (AIC) for model comparison.

2.3 *Estimation*

The log-likelihood for n individuals based on model (2.4) is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} g(\mathbf{y}_i|\mathbf{x}_i, \theta) = \sum_{i=1}^{n} \{\log[\{1 - p(\mathbf{x}_i; \boldsymbol{\beta})\} f(\mathbf{y}_i^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_0, \Sigma_0, \boldsymbol{\phi}_0)$$
$$+ p(\mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{y}_i^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\phi}_1)] + (\lambda_1 - 1)\log(y_{i1}) + (\lambda_2 - 1)\log(y_{i2})\}.$$

We maximized $L(\boldsymbol{\theta})$ directly with respect to $\boldsymbol{\theta}$ using a dual quasi-Newton method (NLPQN in PROC IML, SAS 8.2), subject to the constraints $\det(\Sigma_i) > 0, i = 0, 1$. We also implemented an Expectation and Maximization (EM) algorithm (supplemental material available at *Biostatistics* online). In every case we tested, the EM and quasi-Newton method agreed. To obtain convergence to a global maximum, we choose 100, 600, and 1000 starting values for the optimization for model I, $K = 0$, 1, and 2, respectively. For model II, 200 ($K = 1$) and 300 ($K = 2$) starting values were used.

For independent and identically distributed $(\mathbf{Y}_j, \mathbf{X}_j)$, the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \longrightarrow \text{Normal}(\underline{0}, Q^{-1})$, where $\boldsymbol{\theta}$ denotes the true parameters and $Q = -E[\partial^2 \log\{g(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})\}/(\partial\theta_i \partial\theta_j)]$. The expectation is taken with respect to the joint distribution of $(\mathbf{Y}, \mathbf{X})$. We estimate Q by $\widehat{Q}_n = n^{-1} \sum_{i=1}^{n} H_i$, where H_i denotes the negative Hessian of $\log\{g(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})\}$ obtained through numerical differentiation at $\widehat{\boldsymbol{\theta}}$.

Local identifiability of model I (2.4) can be shown to hold at a given point in the inside of the parameter space as the information matrix at any point is nonsingular under the correctly specified model (Rothenberg, 1971). However, issues relating to global identifiability and stability of the models can still arise. For $K = 2$, a single semi-nonparametric density can accommodate heavy tails and skewness as well as multiple modes. This very flexibility can lead to identifiability problems, for example, when multiple modes are present in the data. For a specific realization, either component density, the one corresponding to the infected and the one corresponding to the uninfected population, can capture modes located roughly in the center, which would greatly affect estimates of the mixing proportion. A second related issue is that a single density alone may provide an excellent fit to the data. To our knowledge, there are no published references that address what general shapes the semi-nonparametric densities can accommodate. We aimed to address identifiability of model (2.4) in several numerical experiments based on simulated data (supplemental material available at *Biostatistics* online) and the real data (Section 4).

3. Simulations

To assess the performance of the various models and numerical issues, we fit the models (2.4) and (2.5) to data from several simulated scenarios. One hundred data sets with 1000 or 2000 data points were generated for 3 sets of simulations presented in this section (further simulations are in the supplemental material available at *Biostatistics* online). The tables show mean estimates of $\boldsymbol{\lambda}$ and p over the simulations that converged.

In Table 1, we study the performance for estimating p based on data that arose from a mixture of bivariate normal distributions with constant mixing proportion, $p = 0.25$, and to assess the robustness of the models to small p, $p = 0.05$. This case corresponds to a linear Box–Cox transformation, that is, $\lambda_1 = \lambda_2 = 1$. The means of the normal components were $\boldsymbol{\mu}_0 = (10, 10)$ and $\boldsymbol{\mu}_1 = (11, 11)$ and the covariance matrices were $\Sigma_0 = [0.25\ 0.1;\ 0.1\ 0.25]$ and $\Sigma_1 = [0.25\ 0.05;\ 0.05\ 0.5]$, where $\Sigma = [(\Sigma)_{11}\ (\Sigma)_{12};\ (\Sigma)_{12}\ (\Sigma)_{22}]$. For these 2 settings, we also compared the coverage of a 95% confidence interval for p based on asymptotic normality of the estimate to likelihood-ratio test-based confidence intervals.

For the simulations with $p = 0.25$, the corresponding mean estimates of p (with empirical standard errors in parenthesis) were 0.25(0.06), 0.25(0.06), and 0.27(0.11) for models I, $K = 0, 1, 2$, and 0.25(0.07) and 0.10(0.14) for models II, $K = 1$ and $K = 2$, respectively. While all models yielded unbiased estimates of p, the standard error of p for model I, $K = 2$, was nearly twice as large as the standard error of the simpler models I, $K = 0$ and $K = 1$. The coverage of the confidence intervals was close to the nominal 95% level for models I, $K = 0$ and $K = 1$ and for model II, $K = 1$, for which p was well estimated, but it was low for model I, $K = 2$, and model II, $K = 2$, ranging from 71% to 75% (Table 1). This illustrates the need to choose a parsimonious but well-fitting model. Based on the likelihood-ratio

Table 1. *Mean estimates of p and $\boldsymbol{\lambda}$ for models I and II over* 100 *simulations for $N = 1000$ observations simulated from "a mixture of* 2 *bivariate normal distributions with constant p."* $\boldsymbol{\mu}_0 = (10, 10)$, $\boldsymbol{\mu}_1 = (11, 11)$, $\Sigma_0 = [0.25\ 0.1;\ 0.1\ 0.25]$, *and* $\Sigma_1 = [0.25\ 0.05;\ 0.05\ 0.50]$. *Empirical standard errors are given in parenthesis*

Model	$\lambda_1 = 1$	$\lambda_2 = 1$	$p = 0.25$	Coverage† for p	Coverage‡ for p	Log-likelihood
I, $K = 0$	1.00 (0.03)	1.00 (0.02)	0.25 (0.06)	0.93	0.94	−2010.71 (32.9)
I, $K = 1$	1.00 (0.02)	1.02 (0.01)	0.25 (0.06)	0.91	0.93	−2009.69 (32.7)
I, $K = 2$	1.00 (0.02)	1.00 (0.02)	0.27 (0.11)	0.72	0.71	−2006.68 (32.8)
II, $K = 1$			0.25 (0.07)	0.93	0.90	−2009.50 (32.6)
II, $K = 2$			0.10 (0.14)	0.71	0.75	−2007.83 (30.9)
Model	$\lambda_1 = 1$	$\lambda_2 = 1$	$p = 0.05$	Coverage for p	Coverage for p	Log-likelihood
I, $K = 0$	1.00 (0.02)	1.00 (0.02)	0.06 (0.06)	0.78	0.89	−1865.70 (31.0)
I, $K = 1$	1.00 (0.02)	1.02 (0.02)	0.06 (0.05)	0.85	0.84	−1864.69 (30.9)
I, $K = 2$	1.00 (0.02)	1.00 (0.02)	0.08 (0.10)	0.64	0.68	−1860.14 (31.2)
II, $K = 1$			0.08 (0.10)	0.54	0.50	−1860.14 (31.2)
II, $K = 2$			0.08 (0.17)	NA	0.35	−1862.83 (31.3)

†Coverage of confidence intervals based on the asymptotic normality of $\hat{p}$.
‡Coverage of likelihood-ratio test-based confidence intervals.

test, in 85/100 simulations, both models I, $K = 1$ and $K = 2$, did not fit the data statistically significantly better than the simpler model I, $K = 0$. Similarly, model II, $K = 2$, did not provide a better fit than model II, $K = 1$. Not surprisingly, as the data were generated from a mixture of normals, the simplest model provided unbiased estimates of p with smaller variance and the best fit in most of the runs, as is also reflected by the mean log-likelihood values for each model (Table 1).

For the simulations with $p = 0.05$, the estimates of p were 0.06(0.06), 0.06(0.05), and 0.10(0.11) for models I, $K = 0, 1, 2$, and 0.08(0.10) and 0.08(0.17) for models II, $K = 1$ and $K = 2$, respectively. The estimates of the parameters of the component densities were nearly unbiased for all models (data not shown). Based on the likelihood-ratio test, for 83/100 simulations, the more complex models did not provide a better fit than the simplest model I, $K = 0$. While the models estimated the small mixing probabilities without bias, the coverage of all confidence intervals was below the nominal 95% level, with the likelihood-ratio-based confidence intervals yielding slightly better coverage for models I, $K = 0$ and $K = 2$. We attribute the lower-than-nominal coverage to the fact that p was close to the boundary zero of the parameter space. The asymptotic normal confidence intervals for model II, $K = 2$, are not shown as 90/100 runs resulted in singular Hessian matrices.

For the second set of simulations (Table 2), we generated data from the same bivariate normal distributions as Table 1, but with mixing probability $p = \exp(\beta_0 + \beta_1 X)/\{1 + \exp(\beta_0 + \beta_1 X)\}$, with $\beta_0 = -2$, $\beta_1 = 1$, and a Bernoulli covariate $X \in \{0, 1\}$ with probability 0.5. For these parameters, $E(p) = 0.19$.

For the logistic mixture, all models estimated the parameters $\boldsymbol{\beta}$ of the mixing proportion as close to $(-2.0, 1.0)$, with similar standard errors for $N = 1000$, with the exception of model II, $K = 2$. For $N = 2000$ data points, the estimates of $\boldsymbol{\beta}$ were virtually unbiased with comparable standard errors for all models. Models I, $K = 0$, $K = 1$, and $K = 2$, also resulted in similar parameter estimates of $\boldsymbol{\lambda}$. The estimated mean vectors and covariance matrices were also nearly unbiased for all models. The asymptotic normal confidence intervals had approximately 95% coverage for all models with the exception of model I, $K = 2$, where the coverage for β_1 was only 75% for $N = 1000$. For $N = 2000$, however, all models but

Table 2. *Mean estimates for models I and II over* 100 *simulations for* $N = 1000$ *or* $N = 5000$ *observations simulated from "a mixture of* 2 *bivariate normal distributions with logistic* $p(X, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x)/\{1 + \exp(\beta_0 + \beta_1 x)\}$*,"* $X \in \{0, 1\}$ *with probability* 0.5. $\boldsymbol{\mu}_0 = (10, 10)$, $\boldsymbol{\mu}_1 = (11, 11)$, $\Sigma_0 = [0.25\ 0.1;\ 0.1\ 0.25]$, *and* $\Sigma_1 = [0.25\ 0.05;\ 0.05\ 0.50]$. *Empirical standard errors are given in parenthesis*

Model	$\lambda_1 = 1$	$\lambda_2 = 1$	$\beta_0 = -2.0$	$\beta_1 = 1.0$	Coverage† (β_0, β_1)	Log-likelihood
$N = 1000$						
I, $K = 0$	1.02 (0.01)	1.01 (0.09)	−1.89 (0.53)	1.27 (1.038)	(0.94, 0.96)	−1976.37 (32.1)
I, $K = 1$	1.02 (0.01)	1.03 (0.01)	−1.85 (0.69)	1.02 (0.25)	(0.97, 0.98)	−1975.20 (32.6)
I, $K = 2$	1.01 (0.09)	1.01 (0.02)	−1.81 (0.69)	1.22 (0.95)	(0.75, 1.00)	−1971.33 (34.9)
II, $K = 1$			−1.88 (0.52)	1.11 (0.089)	(0.92, 0.96)	−1975.36 (31.5)
II, $K = 2$			−1.96 (1.53)	1.23 (1.38)	NA	−1970.26 (33.8)
$N = 2000$						
I, $K = 0$	0.99 (0.03)	1.00 (0.003)	−1.98 (0.21)	1.00 (0.15)	(0.98, 0.95)	−3960.01 (42.8)
I, $K = 1$	1.00 (0.003)	1.00 (0.004)	−2.00 (0.21)	0.99 (0.15)	(0.99, 0.94)	−3959.96 (42.8)
I, $K = 2$	1.00 (0.003)	1.00 (0.004)	−2.00 (0.23)	1.00 (0.15)	(0.97, 0.97)	−3959.62 (42.9)
I, $K = 1$			−1.99 (0.27)	1.00 (0.17)	(0.97, 0.95)	−3959.77 (42.7)
I, $K = 2$			−1.96 (0.35)	1.01 (0.17)	NA	−3959.11 (42.9)

†Coverage of confidence intervals based on the asymptotic normality of $\hat{p}$.

model II, $K = 2$, had nominal coverage. For model II, $K = 2$, again nearly all runs resulted in singular Hessian matrices. Based on likelihood-ratio test, model I, $K = 0$, was preferable to the more complex models in nearly all simulations.

The last set of simulations (Table 3) assessed the performance of the models for highly skewed data with constant mixing proportions $p = 0.5$ and $p = 0.05$. The components y_{01} and y_{02} of the first subpopulation were independent χ_1^2 variables, and the components of the second subpopulation y_{11} and y_{12} were independent χ_3^2 variables. For the data simulated with $p = 0.5$, the mean estimated mixing proportions were 0.51(0.05), 0.51(0.05), and 0.48(0.12) for models I with $K = 0$, $K = 1$, and $K = 2$, respectively, with the standard error for $K = 2$ being more than twice as large as for the simpler models. The estimates of p for model II, $K = 2$, and model II, $K = 1$, yielded a lower average estimate of p of 0.44(0.06). Models I, $K = 0$ and $K = 1$, resulted in very similar parameter estimates, and the estimates of the polynomial coefficients of models I, $K = 1$ and $K = 2$, did not provide evidence for a statistically significant polynomial component. The likelihood-ratio-based confidence intervals for models I, $K = 0$ and $K = 1$, had 94% coverage, while it was only 74% for model I, $K = 2$. For both models II, the coverage was much lower, 39% and 38% for $K = 1$ and $K = 2$, respectively. Again, these models did not fit the data as well as model I, $K = 0$, which has fewer parameters but allows for a Box–Cox transformation. All models correctly estimated the correlation terms in the mixing densities close to zero. All runs converged for models I, $K = 0$ and $K = 1$, and 97/100 runs converged for $K = 2$. For model II, 91/100 simulations converged for $K = 1$ and 89/100 for $K = 2$. The simulations with $p = 0.05$ illustrate well that a lack of fit of the mixing components can lead to severe bias in the estimates of p. The mean estimates of p were 0.04(0.05), 0.05(0.07), and 0.08(0.09) for models I with $K = 0$, $K = 1$, and $K = 2$, respectively, and, highly biased, 0.60(0.05) and 0.59(0.06) for models II, $K = 1$ and $K = 2$, respectively. The coverage of likelihood-ratio-based confidence intervals, however, was below the 95% nominal level for all models. For model I, 99/100 runs converged for $K = 0$, 95/100 for $K = 1$, and 94/100 for $K = 2$. For model II, all runs converged for $K = 1$ and 99/100 for $K = 2$. As indicated by the log-likelihood-ratio test, model

Table 3. *Mean estimates of p and $\boldsymbol{\lambda}$ for models I and II over* 100 *simulations for $N = 1000$ observations simulated from "a mixture of* 2 *chi-square distributions, constant p": $y_{01} \sim \chi^2_1$, $y_{02} \sim \chi^2_1$, $y_{11} \sim \chi^2_3$, and $y_{12} \sim \chi^2_3$. Empirical standard errors are given in parenthesis*

Model	λ_1	λ_2	$p = 0.5$	Coverage† for p	Log-likelihood
I, $K = 0$	0.20 (0.02)	0.20 (0.02)	0.51 (0.05)	0.94	−3152.49 (66.9)
I, $K = 1$	0.20 (0.02)	0.20 (0.02)	0.51 (0.05)	0.94	−3152.47 (66.9)
I, $K = 2$	0.19 (0.03)	0.18 (0.03)	0.48 (0.12)	0.74	−3138.60 (67.8)
II, $K = 1$			0.44 (0.06)	0.39	−3893.36 (56.6)
II, $K = 2$			0.44 (0.06)	0.38	−3862.35 (69.3)
Model	λ_1	λ_2	$p = 0.05$	Coverage† for p	Log-likelihood
I, $K = 0$	0.31 (0.03)	0.30 (0.03)	0.04 (0.05)	0.77	−4061.50 (39.3)
I, $K = 1$	0.30 (0.03)	0.30 (0.03)	0.05 (0.07)	0.67	−4061.64 (40.3)
I, $K = 2$	0.28 (0.06)	0.27 (0.06)	0.08 (0.09)	0.66	−4054.07 (40.7)
II, $K = 1$			0.60 (0.05)	0.00	−4327.55 (40.5)
II, $K = 2$			0.59 (0.06)	0.00	−4253.37 (40.6)

†Coverage for likelihood-ratio-based confidence intervals.

I, $K = 2$, fits the data better than model I, $K = 0$ or $K = 1$, for $p = 0.5$, but not for $p = 0.05$. For all values of K, however, model I fits the data significantly better than model II.

4. Application to the Ugandan HHV-8 study

We applied the bivariate mixture models (2.4) and (2.5) to data collected from 599 children aged 0–16 years, at Mulago Hospital, Kampala, from November 2001 to April 2002. Interviewers obtained a blood sample from each child for the K8.1 and the orf73 immunoassays. The main predictors of infection status were age (younger than 5 years, $5 \leqslant$ age < 10, older than 10 years), transfusion status (ever/never transfused), and water source (tap water versus surface water). Details about the study and related HHV-8 epidemiology are in Mbulaiteye *and others* (2003).

4.1 *Analysis using mixture models with constant p*

Table 4 shows the estimates of $\boldsymbol{\lambda}$, p, and the value of the log-likelihood for models I and II with constant mixing probability p and for single semi-nonparametric densities. Model I, $K = 2$, had a significantly better fit than all other models based on the likelihood-ratio test and also had the largest AIC value. Model I, $K = 1$, also fit the data significantly better than model I, $K = 0$, and both models II, based on the likelihood-ratio test. Histograms of the K8.1 and orf73 OD readings on the λ scales for model I and on the original scale for model II, with the corresponding fits from models I and II for $K = 1$ and $K = 2$ superimposed are presented in Figure 1, respectively. Model parameter estimates are presented in the supplemental material available at *Biostatistics* online. The prevalence estimates based on model I were $p = 0.19(0.03)$ for $K = 1$ and $p = 0.18(0.02)$ for $K = 2$, while they were much larger for the models with poor fit, $p = 0.49(0.06)$ for model I, $K = 0$, and $p = 0.44(0.03)$ and $p = 0.43(0.06)$ for model II with $K = 1$ and $K = 2$. Models with poorer fit exhibited higher collinearity in parameter estimates.

Table 4. *"Uganda HHV-8 study": Results for mixture models I and II with constant p for bivariate and univariate data, and single Semi-nonparametric density fit to bivariate data; model-based standard errors given in parenthesis*

Model	λ_1	λ_2	p	Log-likelihood	AIC
Bivariate models					
I, $K = 0$	0.43 (0.07)	0.35 (0.05)	0.49 (0.06)	−172.48	−370.96
I, $K = 1$	0.24 (0.03)	0.24 (0.04)	0.19 (0.03)	−160.13	−354.26
I, $K = 2$	0.27 (0.05)	0.10 (0.05)	0.18 (0.02)	−145.98	−337.96
II, $K = 1$			0.44 (0.03)	−258.25	−546.50
II, $K = 2$			0.43 (0.03)	−210.26	−462.52
Single density, $K = 0$	0.07 (0.16)	0.03 (0.03)		−232.39	−478.78
Single density, $K = 1$	0.09 (0.16)	0.03 (0.03)		−224.28	−466.56
Single density, $K = 2$	0.32 (0.04)	0.20 (0.05)		−181.10	−386.20
Univariate models for K8.1 assay					
I, $K = 0$	0.11 (0.05)		0.17 (0.03)	−302.36	−308.36
I, $K = 1$	0.04 (0.04)		0.16 (0.02)	−293.30	−301.30
I, $K = 2$	0.41 (0.08)		0.16 (0.03)	−288.59	−298.59
II, $K = 1$			0.39 (0.03)	−329.14	−336.14
II, $K = 2$			0.42 (0.03)	−293.26	−302.26
Univariate models for orf73 assay					
I, $K = 0$	0.28 (0.11)		0.35 (0.11)	−47.29	−53.29
I, $K = 1$	0.08 (0.06)		0.18 (0.04)	−43.01	−51.01
I, $K = 2$	0.24 (0.10)		0.35 (0.09)	−38.78	−48.78
II, $K = 1$			0.41 (0.03)	−81.72	−88.72
II, $K = 2$			0.42 (0.03)	−50.78	−59.78

The correlation of $\hat{p}$ with the other model parameters was largest for model I, $K = 0$; for example, the correlations of $\hat{p}$ with $\hat{\mu}_{00}$ and $\hat{\mu}_{10}$ were 0.71 and 0.85, respectively, which made the estimates of p very sensitive to the fit of the mixing components. For model I, $K = 2$, however, the largest correlation was 0.35 between $\hat{p}$ and $(\widehat{\Sigma}_1)_{11}$. Estimates of p were insensitive to the choice of starting values.

To study the stability and possible identifiability problems of the estimates of p in the Uganda data set, we sampled 100 data sets with replacement and fit models I, $K = 1$ and $K = 2$ with 600 and 1000 starting values, respectively. The mean estimates of p over 100 bootstrap repetitions (bootstrap standard deviation in parenthesis) were 0.19(0.05) and 0.18(0.03) for models I, $K = 1$ and $K = 2$, respectively. The standard deviations estimated from the bootstrap were very close to the model-based estimates of the standard deviations (Table 4), indicating that the information matrix was well defined and asymptotic theory could be used for inference. Histogram plots of the bootstrap $\hat{p}$ (Figure 2) showed a unimodal distribution very narrowly centered around 0.2 for models I, $K = 1$ and $K = 2$.

For all choices of K, and even for $K = 0$, the 2-component mixture fits the Uganda data better than a single semi-nonparametric density as assessed by the AIC (Table 4).

We compared the prevalence estimate from the bivariate model to estimates obtained by averaging prevalence estimates from univariate mixture models that were fitted separately to the K8.1 and orf73 assays (Table 4). To account for the dependence between the univariate estimates, we computed the standard errors of the averaged estimates using a bootstrap procedure, assuming that the weights were known and fixed for the inverse variance–weighted estimate.

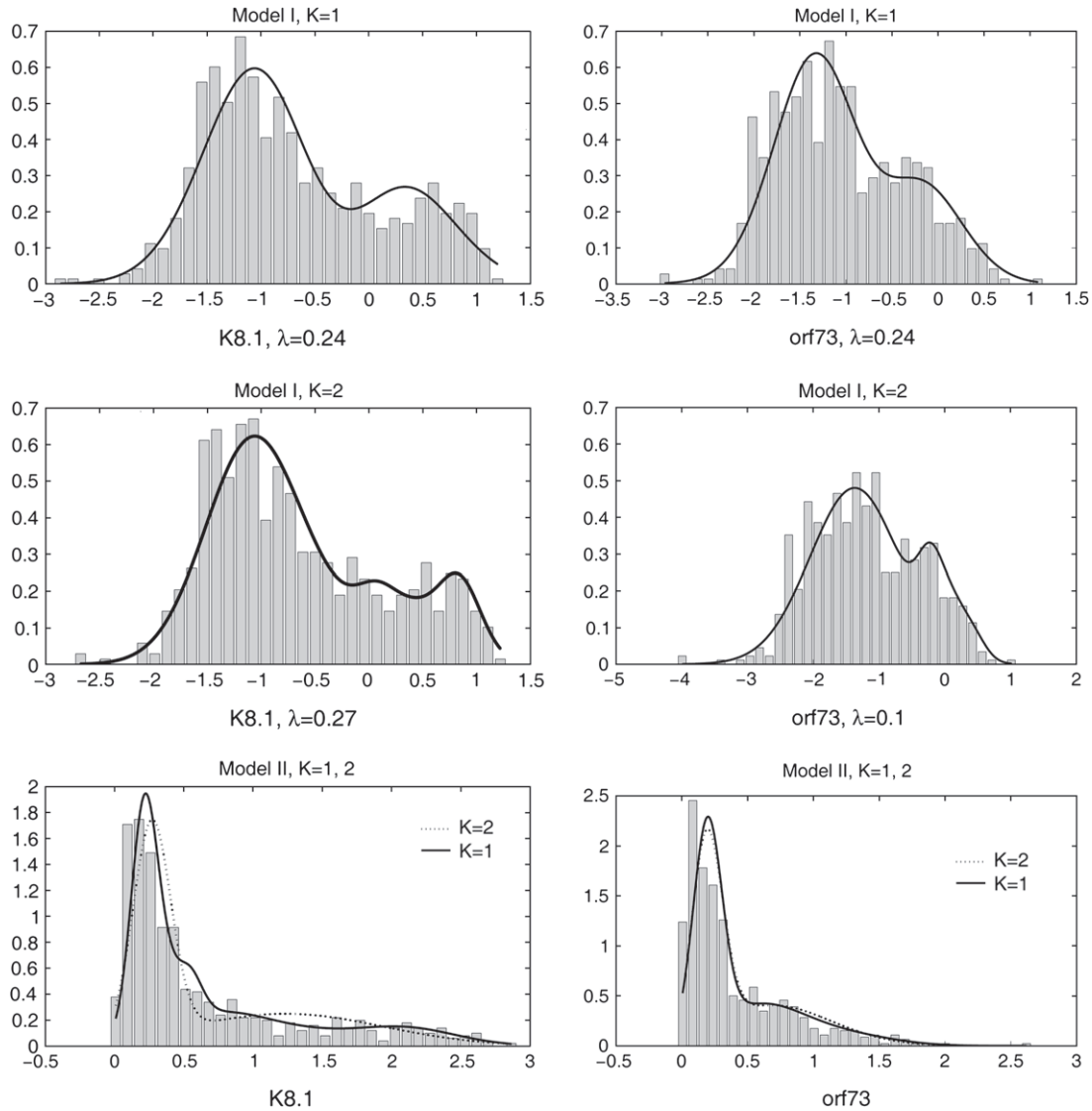

Fig. 1. Histograms and fits from bivariate mixture densities on λ scale for Uganda HHV-8 study.

Based on the likelihood-ratio test, model I, $K = 2$, had a significantly better fit than all other models and also the largest AIC value. The simple average of the prevalence estimates from the best fitting models was 0.26(0.08), while the inverse variance–weighted estimate was 0.18(0.06). The inverse variance–weighted estimate thus agreed with the prevalence estimate from the bivariate model. However, the bootstrap standard error of this estimate was 3 times larger than the estimated standard error of p from the bivariate mixture model and twice as large as the bootstrap estimate of the standard error of p from that model. The bivariate mixture model thus provided a much more precise estimate of prevalence.

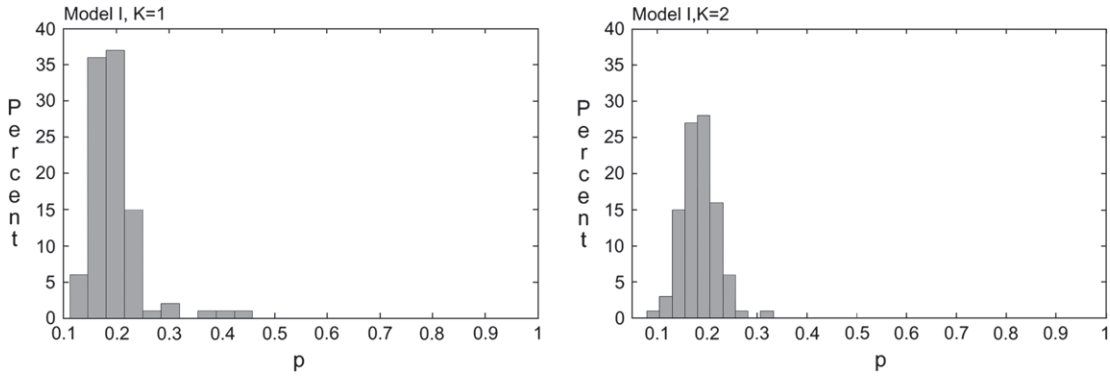

Fig. 2. Histogram of bootstrap estimates of p for Uganda HHV-8 data.

Table 5. *Uganda HHV-8 Study: Estimation of parameters in "logistic p" and logistic regression based on fixed cutoff values for the assays; model-based standard errors given in parentheses*

Parameter	Bivariate mixture model					Standard logistic regression Infected defined based on fixed cutoff values for		
	Model I			Model II				
	$K = 0$	$K = 1$	$K = 2$	$K = 1$	$K = 2$	K8.1	orf73	Combined
Intercept	−1.65	−1.73	−2.05	−1.65	−1.68	0.01	−0.35	0.15
	(0.28)	(0.29)	(0.39)	(0.25)	(0.24)	(0.23)	(0.23)	(0.21)
$5 \leqslant$ age < 10	1.56	1.55	1.72	1.32	1.35	0.94	1.49	1.17
	(0.28)	(0.28)	(0.33)	(0.24)	(0.24)	(0.30)	(0.32)	(0.26)
Age $\geqslant 10$	1.63	1.63	1.94	1.39	1.40	1.75	1.36	1.46
	(0.29)	(0.29)	(0.35)	(0.25)	(0.25)	(0.30)	(0.32)	(0.27)
Ever transfused	0.31	0.30	0.50	0.32	0.32	0.31	0.30	0.21
	(0.23)	(0.22)	(0.25)	(0.19)	(0.19)	(0.22)	(0.22)	(0.20)
Surface water	0.80	0.79	1.09	0.63	0.69	0.55	0.93	0.81
	(0.24)	(0.23)	(0.27)	(0.20)	(0.20)	(0.22)	(0.22)	(0.20)
Log-likelihood	−142.61	−136.67	−117.66	−231.32	−181.46	−271.00	−267.49	−310.21
AIC	−319.22	−315.34	−289.32	−500.64	−412.92	−372.92	−552.00	−544.98

4.2 *Analysis with logistic mixture models*

We then modeled the mixing component p by a logistic function that included age in 2 categories, transfusion status, and water source as covariates. We could incorporate covariates by regressing the means of the component densities on covariates, but in our problem, the covariates considered were thought to influence the chance of being infected, but not the antibody distributions conditional on infection status. To verify this, we first fit the constant p models to data stratified on the categories of age, transfusion status, and water source. The stratified component density estimates were very similar, and thus the more general model was not needed in our data.

The estimates for the parameters in the logistic component and the values of the likelihood and AIC are given in Table 5. Again, the fit of model I, $K = 2$, was significantly better than the fits of the other models based on the likelihood-ratio test. Model I, $K = 2$, also had the largest AIC value. The parameters of the mixing components for model I, $K = 0$, and model II did not change much compared to the model with constant p. For models I, $K = 1$ and $K = 2$, the parameters of the mixing components, however, were

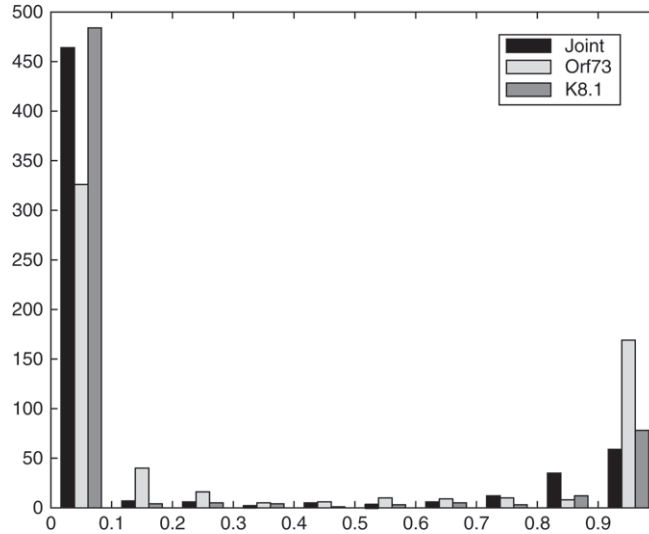

Fig. 3. Posterior probabilities of infection for the Uganda children.

more affected. For example, for $K = 1$, $\boldsymbol{\lambda}$ changed from (0.24, 0.24) for constant p to $\boldsymbol{\lambda} = (0.41, 0.33)$ for the logistic mixture. The parameters of the logistic mixing probability were similar: age and surface water source were associated with significantly elevated risk of infection for all models and transfusion status was not significantly associated. Model I, $K = 2$, resulted in slightly larger estimates for the log-odds parameters than the other models. Model I, $K = 2$, also provided the best univariate fits when applied separately to the K8.1 and orf73 OD readings (data not shown).

The mixture model allows one to calculate the posterior probability of infection, $I_j = 1$, given $\mathbf{x}_j$ and $\mathbf{y}_j$. Indeed, from (2.4), we get

$$\mathrm{pr}(I_j = 1|\mathbf{y}_j, \mathbf{x}_j) = \frac{p(\mathbf{x}; \boldsymbol{\beta}) f(\mathbf{y}^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\phi}_1)}{\{1 - p(\mathbf{x}; \boldsymbol{\beta})\} f(\mathbf{y}^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_0, \Sigma_0, \boldsymbol{\phi}_0) + p(\mathbf{x}; \boldsymbol{\beta}) f(\mathbf{y}^{(\boldsymbol{\lambda})}, \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\phi}_1)}.$$

Figure 3 shows histograms of the posterior probabilities of infection computed based on model I, $K = 2$, fit to the bivariate data as well as to the marginal K8.1 and orf73 data. To minimize the overall misclassification probability based on the mixture model when discriminating infected from uninfected subjects, one sets $I_j = 1$ if $\mathrm{pr}(I_j = 1|\mathbf{y}_j, \mathbf{x}_j) \geqslant 0.5$ and $I_j = 0$ otherwise. For all models, fewer than 5% of the estimates were between 0.3 and 0.7.

4.3 *Analysis with infection status assumed to be known*

We compared the estimates from the mixing probability of the multivariate mixture models with several marginal logistic models, based on operational definitions of infected. The coefficients in this model have a different interpretation than the parameters in the logistic part of the mixture, however. They are based on the observable events $T_i = I(y_i \geqslant c_i)$, $i = 1, 2$, whereas the mixture models estimate the probability of the latent, unobservable infection state. To define infected, we applied the cutoff points used by Mbulaiteye *and others* (2003), with OD for the K8.1 $\leqslant$0.90 corresponding to uninfected, OD $>$1.20 to infected, and, somewhat arbitrarily, OD reading in the range 0.90–1.20 labeled as "indeterminate". The prevalence of infection (excluding 38 indeterminate children) was 117/561= 20.9%. The operational definition for the

orf73 assay was that a child was HHV-8 negative for OD readings $\leqslant 0.5$, indeterminate for OD readings in the range 0.5–0.7, and infected if OD $\geqslant 0.7$. Fifty children were indeterminate and 127 (23.1%) of the remaining 549 were classified as infected, based on their orf73 OD readings. These estimated prevalences agree well with the estimates of $\hat{p}$ for models I ($K = 1, 2$) in Table 4. Often results from both assays are combined by assuming that a child that is positive on either one of the 2 assays is infected, $T = \max(T_1, T_2)$. After excluding 3 children who were indeterminate on both assays, the prevalence based on this criterion was 23%. The corresponding prevalence estimate based on the mixture model is the posterior probability of infection given that the assay readings were not in the indeterminate region, $\mathrm{pr}(I = 1 | y_1 \notin [0.9, 1.2], y_2 \notin [0.5, 0.7])$, estimated to be 0.19 and 0.18 for models I, $K = 1$ and $K = 2$, respectively. The estimates for the models with poor fit were again much higher, 0.48 for model I, $K = 0$, and 0.43 and 0.42 for models II, $K = 1$ and $K = 2$. As only 3 children had OD readings in the joint indeterminate region, these estimates differed only slightly from $\hat{p}$ in Table 4.

Estimates of log-odds ratios based on the various definitions for infected (Table 5) are close to those obtained from the mixture models, leading us to conclude that the operational definitions of infected capture the true latent infection status well.

4.4 *Estimation and comparison of cutoff points*

The multivariate mixture can also be used to find the cutoff values that in some sense best separate the uninfected from the infected population. To determine optimal cutoff points for the assays, we minimize the probability of misclassification under the mixture model as a function of cut-points (c_1, c_2),

$$p \int_{c_1}^{\infty} \int_{c_2}^{\infty} f(\mathbf{y}; \alpha_0) \mathrm{d}y + (1 - p) \int_{-\infty}^{c_1} \int_{-\infty}^{c_2} f(\mathbf{y}; \alpha_1) \mathrm{d}y, \tag{4.1}$$

where the α_i, $i = 0, 1$, and p are replaced by their estimates. Reported on the original OD scale, minimizing (4.1) for model I, $K = 2$, yields $c_1 = 0.79$ and $c_2 = 0.79$ with misclassification probability 0.02. While this would not change the number of children that falls above the cutoff value for K8.1 compared to the cutoff values which the investigators used previously, 39 more children would be classified as infected based on the orf73 if $c_2 = 0.79$ was used.

5. Discussion

In this paper, we present a new class of multivariate mixture models that combines the Box–Cox transformation with a class of semiparametric densities for the mixing components. This class contains mixtures of normals as a special case and, for a fixed number of mixing components, allows for formal testing of models of increasing complexity. Covariates can be incorporated into the mixing probabilities by logistic regression or other generalized linear models. The motivation for our work was the desire to combine 2 different assays to assess infection status with HHV-8 and factors that influence prevalence. Although we are fairly certain that there are 2 distinct subpopulations in the data, one infected and the other uninfected, the assay measurements are not always well separated and the data have heavy tails. The main interest in our applications was the estimation of the parameters relating to the mixing proportion or prevalence, while the parameters of the mixing components were nuisance parameters in the model. An attractive feature of our model is that it can accommodate skewness and multimodality in the data, but this very flexibility can lead to identifiability problems. To study the identifiability aspects of the models, we examined the behavior and stability of estimates of p in simulations (supplemental material available at *Biostatistics* online), using a bootstrap procedure for the Uganda data. In summary, to avoid identifiability problems in using our proposed class of models, it is necessary to try numerous starting values for maximization to ensure convergence to a global maximum and, most importantly, to avoid overfitting by testing

nested models of increasing complexity as recommended for a single semi-nonparametric density (Liu and Zhang, 1998).

Choosing well-fitting but parsimonious models is also important as the use of more flexible models for the component densities, while reducing bias, has the potential to increase the variance of key parameters, such as prevalence. Such a tendency was seen in simulations (Tables 1 and 2) where all the models, even the simpler ones, fit the data well, but the simpler models yielded more precise estimates of prevalence. However, in data from the Uganda study, we found that more complex but better fitting models yielded more precise estimates of prevalence than a simpler, poorly fitting models (Table 4).

Application of the models to the Uganda data highlights the importance of including both the Box–Cox transformations and the polynomial components in the densities to provide adequate fit to the data and thus stable estimates of prevalence. Estimates of HHV-8 prevalence were insensitive to the choice of starting values for p, and bootstrap replications yielded a tight distribution of $\hat{p}$ centered about the original estimate. Bootstrap standard errors were close to model-based standard errors, indicating that p was well identified in our data and that inference based on asymptotic theory was valid.

We compared prevalence estimates p from the bivariate mixture model to estimates obtained by averaging prevalence estimates from univariate mixtures fit to each assay separately. While the inverse variance–weighted prevalence estimate was identical to p from the bivariate mixture model, the estimated standard error of the inverse variance–weighted estimate (assuming fixed and known weights) was 3 times larger than the model-based standard error of p from the bivariate mixture model and twice as large as its bootstrap standard error. Computing prevalence by averaging estimates from the marginal models thus resulted in an estimated loss of efficiency of at least 75% for this data set.

Our work relates to other approaches for evaluating diagnostic tests without gold standards. Rindskopf and Rindskopf (1986) among others fitted 2-component multivariate mixture models with the components corresponding to "diseased" and "non-diseased" subjects. However, the results of the k tests applied to the same person were assumed to be independent conditional on disease status. We relax the independence assumption by allowing joint densities for each component. In our application, 2 immunoassays that detect 2 different types of antibodies were used to assess infection status. In case of infection, both types of antibodies can be present, and thus independence likely does not hold.

The mixture model approach has several potential advantages compared to standard epidemiologic approaches to define infection status. We need not rely on an external definition of a cutoff value to classify each observation. The continuous nature of the data is used to its full extent, and we obtain a complete description of the distribution of the OD values (y_1, y_2) in the presence of covariates X. This enables us to calculate $\text{pr}(\text{infected}|y_1, y_2, X)$, the probability of being truly infected given the OD readings and covariates X.

ACKNOWLEDGMENTS

R. J. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). The Uganda study was funded partly by the National Cancer Institute (CO-12400). We thank Mitchell Gail for helpful comments. *Conflict of Interest:* None declared.

REFERENCES

CARROLL, R. J. AND RUPPERT, D. (1981). Prediction and the power transformation family. *Biometrika* **68**, 609–616.

CHANG, Y., CESARMAN, E., PESSING, M. S., LEE, F., CULPEPPER, J., KNOWLES, D. M. AND MOORE, P. S. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869.

FRALEY, C. AND RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

GALLANT, A. R. AND NYCHKA, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**, 363–390.

GUTIERREZ, R. G., CARROLL, R. J., WANG, N., LEE, G. H. AND TAYLOR, B. H. (1995). Analysis of tomato root initiation using a normal mixture distribution. *Biometrics* **51**, 1461–1468.

LIU, M. AND ZHANG, H. H. (1998). Overparameterization in the seminonparametric density estimation. *Economics Letters* **60**, 11–18.

MARTIN, J. N. (2003). Diagnosis and epidemiology of human herpesvirus 8 infection. *Seminars in Hematology* **40**, 133–142.

MBULAITEYE, S. M., BIGGAR, R. J., BAKAKAI, P. M., PFEIFFER, R. M., WHITBY, D., OWOR, A. M., KATONGOLE-MBIDDE, E., GOEDERT, J. J., NDUGWA, C. M. AND ENGELS, E.A. (2003). Human herpesvirus 8 infection and transfusion history in children with sickle-cell disease in Uganda. *Journal of the National Cancer Institute* **95**, 1330–1335.

MCLACHLAN, G. J. AND PEEL, D. (2000). *Finite Mixture Models*. New York: Wiley.

MCLACHLAN, G. J., PEEL, D. AND BEAN, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* **41**, 379–388.

PFEIFFER, R., GAIL, M. H. AND BROWN, L. (2000). A mixture model for the distribution of IgG antibodies to Helicobacter pylori: application to studying factors that affect prevalence. *Journal of Epidemiology and Biostatistics* **5**, 267–275.

PFEIFFER, R. M., MBULAITEYE, S. M. AND ENGELS, E. A. (2004). A model to estimate risk of infection with human herpesvirus 8 associated with transfusion from cross-sectional data. *Biometrics* **60**, 249–256.

RINDSKOPF, D. AND RINDSKOPF, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine* **5**, 21–27.

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39**, 577–591.

ZHANG, D. AND DAVIDIAN, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**, 795–802.

[*Received August 18, 2006; first revision January 17, 2007; second revision March 22, 2007; accepted for publication April 17, 2007*]