Statistics and Its Interface Volume 14 (2021) 151–163

Bayesian zero-inflated growth mixture models with application to health risk behavior data

Si Yang and Gavino Puggioni*

This paper focuses on developing latent class models for longitudinal data with zero-inflated count response variables. The goals are to model discrete longitudinal patterns of rare events counts (for instance, health-risky behavior), and to identify individual-specific covariates associated with latent class probabilities. Two discrete latent structures are present in this type of model: a latent categorical variable that classifies subgroups with distinct developmental trajectories and a latent binary variable that identifies whether an observation is from a zero-inflation process or a regular count process. Within each class, two sets of covariates are used to separately model the probability of structural zeros and the mean trajectories of the count process. The estimation of the latent variables and regression parameters are carried jointly in a hierarchical Bayesian framework. Our methods are validated through a simulation study and then applied to cigarette smoking data, obtained from the National Longitudinal Study of Adolescent Health.

AMS 2000 subject classifications: Primary 62P15, 62C10; secondary 62C12.
Keywords and phrases: Latent Class Models, Finite mixtures, Longitudinal methods, Adolescent smoking, Behavioral sciences, Hierarchical models.

1. INTRODUCTION

Latent Class Models (LCMs), also known as finite mixture models, are a class of flexible methods used to model unobserved heterogeneity in a population. A LCM assumes that a heterogeneous group can be reduced to several homogeneous subgroups by minimizing the association among responses across multiple variables. The goal is to categorize participants into groups, each one containing participants who are similar to each other and different from participants in other groups [24]. A latent categorical variable is often used to label the group membership. The latent classification has a variety of interpretations under a wide range of applications. For instance, in medical diagnosis, it classifies patients with or without a certain disease when an accurate diagnosis is unavailable; in behavioral and health science, subgroups could involve different behavioral patterns (e.g. drinkers and abstainers); LCMs have also been applied to identify phenotypes or genetic susceptibility for diseases based on clinical and biological data [16, 24, 31, 34, 42].

LCMs have been extended to accommodate longitudinally observed data to identify distinct groups of changing trajectories within a population. Using a semi-parametric strategy, Nagin [26] developed a group-based approach for estimating trajectories for longitudinal data with different types of outcomes. The developmental trajectories are modeled through time dependent parameters. In many applications, it is customary to assume that the difference among several trajectory classes is associated with some stable individual characteristics or background variables. This type of LCM extension has been referred to as Latent Class Growth Model (LCGM). Using a frequentist approach, parameters from LCGM can be estimated through the SAS procedure TRAJ written by Jones et al. [15]. In this setting, the inferential interest focus is on a) estimating the proportion of the population in each subgroup, b) relating group membership probabilities to individual characteristics, and c) profiling the characteristics of individuals within subgroups [26]. More specifically, time invariant risk factors can be incorporated in the model by assuming they influence the probability of being in a certain class and time varying covariates can also be included to directly affect the observed outcome. A further extension, the Growth Mixture Models (GMMs), is based on the structural equations framework, and it can be described as a combination of Latent Growth Curve Model (LGCM) and a LCM [20, 25, 24]. In a LGCM, the initial status and slope of change for the outcome variables are considered as random continuous latent growth factors. Thus, a GMM estimates a mean growth curve for each class and also allows individual variations within classes, whereas a LCGM assumes variation in growth patterns within each class is zero. A detailed description of LCGM and GMM was given by Muthén [23].

When the observed outcome of interest is a count variable, often a high incidence of zero counts is encountered. As an illustration, consider a dataset of cigarettes smoking from the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a longitudinal, nationally representative, and school based study of U.S. adolescents in grades 7 through 12. In 1995–1996, the first wave in-home interviews were conducted on students aged 11–21 years. Further waves were collected in 1996, 2001–2002, and 2007–2008 when the

*Corresponding author.

sample was aged 24–33 years. Participants were asked to report the average number of cigarettes smoked per day in the past 30 days each time they took the survey. Although the percentage of individuals who reported 0 cigarette use decreased as age increased, there were about 64%–77% of zero counts in these four waves of the data. In practice, the classic Poisson regression model is often of limited use because of its equality constraint on variance and mean. Zero-inflated Poisson (ZIP) model and zero-inflated negative binomial (ZINB) model are often used to analyze count data with excessive zeros. Zero-inflated models assume that there are two underlying processes generating zeros, one from the zero point mass (or structural zero) process and one from the Poisson or negative binomial process. When a value of zero is observed in the response, the process which it belongs to is unknown. A latent binary variable that follows a Bernoulli distribution is usually introduced to label structural zeros and non-structural zeros. Zero-inflation is particularly meaningful when there are theoretical justifications for modeling zeros in two separate processes. For instance, in public health and medical studies, we can assume that zeros to arise from at-risk (susceptible) and not-at-risk (non-susceptible) populations (e.g. a zero count of smoked cigarettes could come from a non-smoker or a smoker who reported zero cigarette during the period of study). For this reason, zero-inflated LCMs are a methodologically justified choice for the application in our paper, since several zeros are expected to represent a temporary attempt to quit smoking.

For count data, Shiyko et al. [37] applied Poisson GMM for modeling smoking cessation behavior among smokers. GMM was also used for modeling delinquent behavior of adolescents and the model was specified to be zero-inflated to account for a large amount of non-delinquent adolescents using a frequentist approach [32]. In the Bayesian framework, both latent class models and zero-inflated regression models have been widely used and applied. Ghosh et al. [12] first proposed a data augmentation method with Markov Chain Monte Carlo (MCMC) to generate posterior samples from zero-inflated models. Dagne [7], Fu et al. [10], and Neelon et al. [28] proposed analyses for correlated or clustered zero-inflated count data. Klein et al. [18] developed Bayesian generalized additive models for data with zero-inflation and over-dispersion. However, there is sparse literature on the Bayesian analysis of zero-inflated latent class models. To our knowledge the only study was conducted by Neelon et al. [27]. They propose a two-part latent class model to analyze the effect of a health care parity policy on mental health use and expenditures. Their data contained a large proportion of participants who did not use any mental health service. A binomial component was used to model the observed zeros and a lognormal to model the right skewed nonzero values. Three classes of participants were identified as low spenders, moderate spenders, and high spenders and they also found that the parity policy had an impact only on moderate spenders.

While Neelon et al. [27] focuses on a zero-inflated continuous dependent variable, our paper proposes a latent class model on longitudinal zero-inflated count responses. The application of interest is to model trajectories of smoking behavior from adolescent to adulthood. Although myriad studies have been conducted on smoking behaviors, many of them focus on adult populations using cross-sectional data. The pattern of cigarette smoking is commonly established during adolescence, and often carried into adulthood, affecting health and wellbeing in later life. Thus, a more detailed and sophisticated understanding of the initiation and establishment of smoking behaviors from adolescence to adulthood is particularly important. Only few researchers have studied development trajectories of smoking behavior using longitudinal data. For instance, Colder et al. [6] studied trajectories of adolescent smoking on a sample of 323 from 12–16 years old and found five distinct patterns for cigarette smoking: early rapid escalators, late moderate escalators, late slow escalators, stable light smokers, and stable puffers. White et al. [44] interviewed 374 participants five times from age 12 until age 30/31 about their smoking behavior and identified three classes of trajectory group: non/experimental smokers, occasional/maturing out smokers, and heavy/regular smokers and found sex differences in smoking developmental trajectories to be notable. From five cohorts of adolescents (ages 12–16 with a sample size of 3647) followed for 3 years, Bernat et al. [3] found six distinct trajectories of smoking: nonsmokers, triers, occasional users, early established, late established, and decliners. Chen and Jacobson [5] also used data from Add Health and modeled the overall developmental trajectories of substance use and found that levels of substance use, including smoking, increased from early adolescence to mid-20s, and then declined after.

Literature from the studies described above and other comparable studies on smoking trajectories [8, 19, 43] all suggest that first, there are diverse patterns of smoking behavior among the population; second, for those who smoke, they usually initiate the smoking behavior in early adolescence and tend to smoke more as they age, and when they reach their 20s or mid-20s, some choose to quit smoking and others become regular smokers; third, the classification of trajectories differ study by study and demographic variables such as gender and race play a role in trajectories of cigarette use; and fourth, most of the studies used a "two stage" estimation approach that cigarette outcomes were first used to categorize participants into different groups and then standard logistic regression analyses were used to test the cross-group difference by risk factors. The separate estimation fails to capture the uncertainty in class membership and often results underestimated standard errors [2]. In order to overcome this limitation, we propose a joint estimation of the latent class membership and risk factors. Gender, race, and some other smoking related risk factors are included in the model as covariates for smoking patterns. Polynomial functions are used to reflect curvilinear trends.

Estimation of the joint posterior distribution for the parameters is quite complex and it does not have a closed analytical form, thus estimation is performed with MCMC algorithms. The rest of this paper is organized as follows: Section 2 presents the proposed ZIP and ZINB LCMs. In Section 3, we have a choice of prior distributions for the model parameters and latent class variables and the MCMC algorithm is outlined. Criteria for model comparisons are also discussed. Section 4 provides a simulation study on a synthetic dataset generated from a three-class ZIP mixture process. Section 5 illustrates the procedure with real data. Section 6 summarizes our findings and discusses directions for future research.

2. ZERO-INFLATED LATENT CLASS GROWTH MODELS AND GROWTH MIXTURE MODELS

A zero-inflated model is a mixture model of a zero point mass and a distribution (in our case Poisson or negative binomial). Let y_{it} be a count measure for individual i measured at the t-th measurement. The probability mass function of a repeated measures ZIP model $f_{ZIP}(y_{it}; p_{it}, \mu_{it})$ and ZINB model $f_{ZINB}(y_{it}; p_{it}, \mu_{it}, \phi_{it})$ can be written, respectively as:

$$(1) \qquad \begin{cases} p_{it} + (1-p_{it})\dfrac{1}{e^{\mu_{it}}}, & \text{for} \quad y_{it}=0 \\ (1-p_{it})\dfrac{\mu_{it}^{y_{it}}}{y_{it}!e^{\mu_{it}}}, & \text{for} \quad y_{it}=1,2,\ldots, \end{cases}$$

$$(2) \qquad \begin{cases} p_{it} + (1-p_{it})\left(\dfrac{\phi}{\mu_{it}+\phi}\right)^{\phi}, & \text{for} \quad y_{it}=0 \\ \quad (1-p_{it})\dfrac{\Gamma(\phi+y_{it})}{y_{it}!\Gamma(\phi)}\times & \\ \times\left(\dfrac{\mu_{it}}{\mu_{it}+\phi}\right)^{y_{it}}\left(\dfrac{\phi}{\mu_{it}+\phi}\right)^{\phi}, & \text{for} \quad y_{it}=1,2,\ldots \end{cases}$$

Two kinds of zeros are thought to exist in the data: "structural zeros" (or true zeros) from a non-susceptible group (i.e., those that do not have the attribute or experience of interest, such as nonsmokers) and "random zeros" (or false zeros) for those from a susceptible group (e.g., those who smoke but may falsely indicate a count of zero). With p_{it} we denote the probability of being in a non-susceptible group and it can be estimated by information from covariates with a logistic link. If an individual is from the susceptible group, the observed count is a realization of a random variable distributed as a Poisson distribution with mean μ_{it} or from a negative binomial distribution with mean μ_{it} and dispersion parameter of ϕ, accounting for over-dispersion generated from positive counts. In this parametrization, ϕ takes only strictly positive values and a bigger ϕ indicates a higher degree of dispersion. In practice, a ZINB model with a value of ϕ close to zero is statistically indistinguishable from a ZIP model [13].

For our modeling purposes, we take mixtures of the distributions in (1) and (2). A random vector Y is said to arise from a finite mixture of ZIP or ZINB distributions, if the probability mass function takes the form of a mixture density for all $y \in Y$ as follows:

$$p(y|p_k, \mu_k) = \sum_{k=1}^{K} \pi_k f_{\text{ZIP}}(y; p_k, \mu_k),$$
$$p(y|p_k, \mu_k, \phi_k) = \sum_{k=1}^{K} \pi_k f_{\text{ZINB}}(y; p_k, \mu_k, \phi_k),$$

where $f_{\text{ZIP}}(y; p_k, \mu_k)$ or $f_{\text{ZINB}}(y; p_k, \mu_k, \phi_k)$ is a probability mass function for all $k = 1, \ldots, K$. K is the number of mixture components. The parameters $\pi_1, \ldots, \pi_K$ are the weights for each component and they indicate the probability of an underlying categorical latent variable C_i taking a value of k with $k = 1, 2, \ldots, K$. Thus, a latent class model on zero-inflated count responses includes two unobserved random variables. First, there is the latent categorical variable C_i, which follows a multinomial distribution: $C_i \sim \mathcal{M}ultinom(\pi_{i1}, \ldots, \pi_{iK})$. It divides a population into different subgroups. Within each subgroup, $B_{it} \sim \mathcal{B}ernoulli(p_{it})$, is a latent variable indicating the split between a structural zero process and a count process. For modeling longitudinal data, latent class variable C_i essentially summarizes different developmental trajectories over time, thus for each participant, their class memberships are constrained to be the same over time. However, over time an individual's response can change from a structural zero to a count or vice versa (e.g., a participant from class 1 can change from being a non-smoker at the beginning of the study to being a regular smoker at the follow-up).

To allow the probabilities of the latent class membership to be functionally related to individual characteristics, time-invariant covariates can be summarized and added to the model to affect the classification of underlying trajectory patterns. Hence, π_{ik} is related to a $r \times 1$ vector of covariates z_i via a logit link as follows:

$$(3) \qquad \pi_{ik} = \frac{e^{z_i^T \gamma_k}}{\sum_{h=1}^{K} e^{z_i^T \gamma_h}}, \quad \text{with } \gamma_1 = 0.$$

Conditioning on class membership, the regression models that predict the probability of being a structural zero (p_{itk}) and the mean of the count process (μ_{itk}) are given by:

$$(4) \qquad \text{logit}(p_{itk}) = \log\left(\frac{p_{itk}}{1-p_{itk}}\right) = x_{it}^T \alpha_k,$$

$$(5) \qquad \log(\mu_{itk}) = x_{it}^T \beta_k + b_{ki},$$

where x_{it} are $p \times 1$ vectors of fixed effect covariates; α_k and β_k are class specific fixed effect regression coefficients for class k; and $b_{ki} \sim \mathcal{N}(0, \sigma_k^2)$ is participant i's random effect for the count component with class specific variance σ_k^2.

When mixture models are fitted on large datasets, the number of classes tends to be overestimated by model selection criteria or other approaches. A mixed model can mitigate this problem and offer more parsimonious choices in terms of number of components, since individual heterogeneity can be incorporated within each class, thus obtaining a similar fit to fixed effect specifications with a larger number of classes. In this work, we fit latent class models with both fixed (LCGM) and mixed effect specifications (GMM); in the rest of the paper, for simplicity we will refer to them generally as LCM models, specifying if it is a fixed or a mixed effect specification, where necessary. In many longitudinal studies, the true trend over time for the underlying mean response is likely to happen in a relatively smooth pattern. Simple parametric curves such as linear and quadratic trends and semi-parametric curves such as piecewise linear trend can be used to describe how the mean response changes over time [9]. As a result, for modeling a quadratic trend, x_{it} includes an intercept, a linear time effect, and a quadratic time effect. Depending on different theoretical justifications, one might allow covariates that affect p and μ to be different. For illustrative purposes, we have the same set of predictors for the two components in this study.

3. MODEL ESTIMATION

3.1 Likelihood and prior specification

Let us consider an observed sample $(y_{11}, z_{11}, x_{11}), \ldots, (y_{NT}, z_{NT}, x_{NT})$ of $N \times T$ observations, where each response observed at time t for individual i is denoted by y_{it}. For the mixed effects ZIP-LCM model, the likelihood of obtaining the observed sample given the vector of parameters and the latent variable $P(Y|\Theta_{ZIP\,Mixed})$, where $\Theta_{ZIP\,Mixed} = \{\alpha_k, \beta_k, \gamma_k, C_i, b_{ki}\})$ has the following form:

$$
\begin{aligned}
&\prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i = k)\prod_{t=1}^{T}\Pr(Y_{it}|C_i = k) \\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\left\{\prod_{t:Y_{it}=0}\left[p_{itk} + (1 - p_{itk}\frac{1}{e^{\mu_{itk}}}\right]\right. \\
&\left. + \prod_{t:Y_{it}\neq 0}(1 - p_{itk})\frac{\mu_{itk}^{y_{it}}}{y_{it}!e^{\mu_{itk}}}\right\} \\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\frac{e^{z_i^T\gamma_k}}{\sum\limits_{h=1}^{K} e^{z_i^T\gamma_h}}\left\{\prod_{t:Y_{it}=0}\left[\frac{1}{e^{-(x_{it}^T\alpha_k)} + 1}\right.\right. \\
&\left. + \frac{1}{e^{e^{x_{it}^T\beta_k + b_{ki}}}(e^{x_{it}^T\alpha_k} + 1)}\right] \\
&\left. + \prod_{t:Y_{it}\neq 0}\frac{e^{(x_{it}^T\beta_k + b_{ki})y_{it}}}{y_{it}!e^{e^{x_{it}^T\beta_k + b_{ki}}}(e^{x_{it}^T\alpha_k} + 1)}\right\}.
\end{aligned}
$$

The likelihood of the mixed effects ZINB-LCM model can be found in Appendix A. We define with $b_{ki} \sim \mathcal{N}(0, \sigma_k^2)$ a Gaussian individual random effect. Prior distributions need to be specified for model parameters $\{\alpha_k, \beta_k, \gamma_k, b_{ki}\}$ and for the additional dispersion parameter ϕ_k in the case of the ZINB model. We assign multivariate normal priors for all class specific regression parameters, an inverse-gamma prior for the variance of the random effect, and a gamma prior for the dispersion parameter. That is, $\alpha_k \sim \mathcal{N}_p(\mu_\alpha, \sigma_\alpha^2 I_p), \beta_k \sim \mathcal{N}_p(\mu_\beta, \sigma_\beta^2 I_p), \gamma_k \sim \mathcal{N}_r(\mu_\gamma, \sigma_\gamma^2 I_r), \sigma_k^2 \sim \mathcal{IG}(a, b)$, and $\phi_k \sim \mathcal{G}(a, b)$. In our analyses, both for the simulation study and the real data application, we choose hyperparameters that characterize diffuse priors, so that the posterior estimates will be mostly determined by the data. When prior information on the parameter distributions is available, one may choose more strongly informative priors and also specify different priors for different classes.

3.2 Posterior computation

Using Bayes' theorem, the joint posterior distribution is proportional to the product of the prior and the likelihood specified in Section 3.1. Since it is not feasible to analytically derive the joint posterior distribution, a Gibbs sampler is used to sample from the full conditional distribution of each parameter. For the fixed effects ZIP-LCM, the full conditional posterior distributions of each parameter and the latent class variable have the following forms:

$$
\begin{aligned}
&\gamma_k|\cdot \propto \prod_{i=1}^{N}[P(C_i = k|\gamma_k; z_i)]^{I(C_i=k)}\pi(\gamma_k), \\
&C_i|\cdot \sim \mathcal{M}ultinom(\rho_{ik}) \propto P(Y_i|C_i, \alpha_k, \beta_k; x_i)P(C_i|\gamma_k; z_i), \\
&\alpha_k|\cdot \propto P(D_k|C_i = k, \alpha_k, \beta_k; x_k)\pi(\alpha_k), \\
&\beta_k|\cdot \propto P(Y_k|C_i = k, \alpha_k, \beta_k; x_k)\pi(\beta_k).
\end{aligned}
$$

We introduce a variable D_{it} here, which is equal to 0 when $Y_{it} = 0$ and equal to 1 when $Y_{it} > 0$. We assume $D_{it} \sim \mathcal{B}inomial(\theta_{it})$ where θ_{it} is the probability of overall observed zeros which combines zeros from the zero-inflation process and the count process (e.g., for a ZIP model, $\theta_{it} = p_{it} + (1 - p_{it})e^{-\mu_{it}}$). For the fixed effects ZINB-LCM, we sample the dispersion parameter from its full conditional:

$$\phi_k|\cdot \propto P(Y_k|C_i = k, \alpha_k, \beta_k; x_k)\pi(\phi_k).$$

For mixed effects models, $\pi(\gamma_k|\cdot)$ has the same form as above, however, $\pi(C_i|\cdot)$, $\pi(\alpha_k|\cdot)$, and $\pi(\beta_k|\cdot)$ are also conditional on random effects b_k. The full conditionals for σ_k^2 and b_i are

$$
\begin{aligned}
&\sigma_k^2|\cdot \sim \mathcal{IG}\left(a + N_k, b + \frac{\sum\limits_{i=1}^{N_k} b_{ki}^2}{2}\right), \\
&b_i|\cdot \propto P(Y_i|C_i = k, \alpha_k, \beta_k, b_i)\pi(b_i),
\end{aligned}
$$

where N_k denotes the number of participants in class k. As for sampling C_i, ρ_{ik} is the posterior probability that individual i belongs to class k and it is given by

$$\rho_{ik} = \frac{\pi_{ik}(\gamma_k)\left[\prod_{t=1}^{T} f_{\text{ZIP}}(y_{itk}|p_{itk},\mu_{itk})\right]\mathcal{N}(b_i;0,\sigma_k^2)}{\sum_{h=1}^{K}\pi_{ih}(\gamma_h)\left[\prod_{t=1}^{T} f_{\text{ZIP}}(y_{itk}|p_{itk},\mu_{itk})\right]\mathcal{N}(b_i;0,\sigma_h^2)},$$

$$\rho_{ik} = \frac{\pi_{ik}(\gamma_k)\prod_{t=1}^{T} f_{\text{ZINB}}(y_{itk}|p_{itk},\mu_{itk},\phi_k)}{\sum_{h=1}^{K}\pi_{ih}(\gamma_h)\prod_{t=1}^{T} f_{\text{ZINB}}(y_{ith}|p_{ith},\mu_{ith},\phi_h)},$$

for a mixed effects ZIP-LCM and a fixed effects ZINB-LCM, respectively. Because no closed forms are available for the full conditional posterior distributions of α, β, γ, and b_i, we use a Metropolis algorithm to draw samples for these three parameters. As a result, for the mixed effects ZIP-LCM, the following algorithm can be used to generate samples from the above full conditional distributions:

1. Assign initial values to α_k, β_k, σ_k for $k = 1, \dots, K$, to γ_k for $k = 2, \dots, K$, to class membership indicator C_i, and to random intercepts b_i;
2. for $k = 2, \dots, K$, update γ using random walk Metropolis;
3. sample C_i from the multinomial distribution based on posterior probability ρ;
4. for $k = 1, \dots, K$, update α_k and β_k using a random walk Metropolis;
5. for $i = 1, \dots, N$, update b_i using a random walk Metropolis; and
6. sample σ_k^2 directly from the inverse-gamma distribution.

Similar steps can be used for the mixed effects ZINB-LCM except that for $k = 1, \dots, K$, we also update ϕ_k using random walk Metropolis-Hastings. For fixed effects models, we do not have to sample b_i and σ_k^2. The Metropolis algorithm proceeds by sampling a proposal value nearby the current value using a symmetric proposal (e.g., normal distribution), whereas the Metropolis-Hastings algorithm uses an asymmetric proposal distribution (e.g., log-normal distribution). While theoretically the proposal density can be arbitrary, in practice, only a distribution that is close to our target distribution will generate an efficient number of acceptances. The proposal density we use for the random walk Metropolis is a multivariate normal density centered at the previous value. As the posterior covariance for regression parameters are close to $\sigma_Y^2(X^TX)^{-1}$ and proportional to $(X^TX)^{-1}$ [14], to improve mixing, the proposal densities we use for updating α_k, β_k, and γ_k are $\mathcal{N}_p(\alpha_k^{old}, t_\alpha(X_k^TX_k)^{-1})$, $\mathcal{N}_p(\beta_k^{old}, t_\beta(X_k^TX_k)^{-1})$, and $\mathcal{N}_r(\gamma_k^{old}, t_\gamma(Z_k^TZ_k)^{-1})$, respectively. Since ϕ can only be positive values, we propose ϕ^{new} from a log-normal distribution, i.e., $\mathcal{LN}(\log(\phi^{old}), t_\phi\sigma_\phi^2)$. We indicate with $\{t_\alpha, t_\beta, t_\gamma, t_\phi\}$ the tuning parameters that can be altered in order to achieve a proper acceptance rate.

The performance of the MCMC algorithm is monitored by inspecting acceptance rates, trace and empirical autocorrelation function plots, and computing common diagnostics on simulated draws, including the effective sample size.

3.3 Model comparison

In the frequentist framework, standard model comparison criteria such as the Akaike information criterion (AIC) [1] and the Bayesian information criterion (BIC) [36] assume the number of parameters to be known, however, the number of parameters in hierarchical Bayesian models is not clear and cannot be determined directly. In the Bayesian framework, there are several approaches for model comparisons, such as Bayes factors and the Deviance Information Criterion (DIC). The former approach is computationally complex and sensitive to prior specifications. In this paper, we use the DIC as the default model selection criterion. Considering that sometimes the DIC can have unpleasant properties, e.g. a possibly negative number of effective parameters, we use as a secondary check for models with more than one class: the DIC_3, a modified version proposed by Celeux et al. [4] in the case of finite mixture models. Details of these models selection criteria calculation can be found in Appendix B. Generally, models with smaller DIC are preferred, but this cannot be an exclusive factor in choosing a model. While selection criteria can be very effective at identifying the data generating model in a synthetic data context, they can fall short with real datasets, when more than one reasonable model can describe the data comparably well.

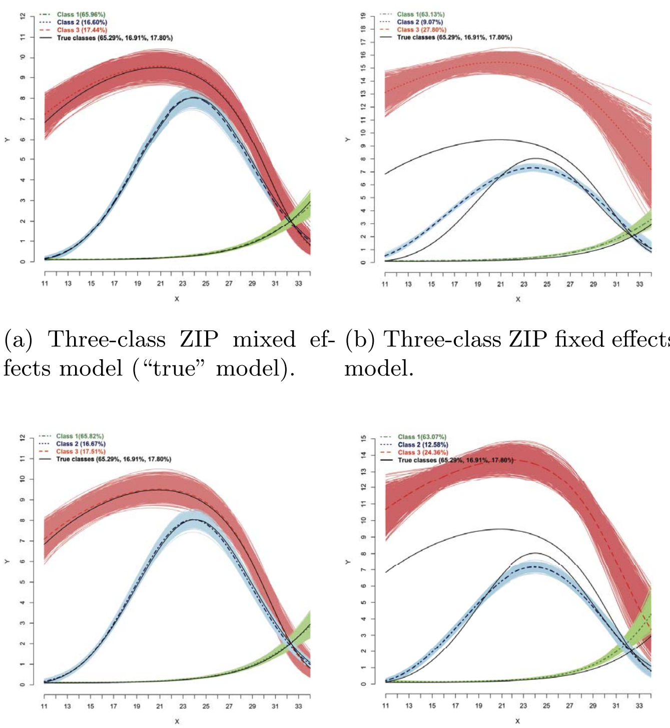
In order to further evaluate the performance of the different modeling specifications, for our real data application we apply some common posterior predictive checking [11] methods. If the model displays a good predictive performance, replicated data y^{rep} generated under the model should look similar to observed data y. Bayesian p-value, which represents the probability that the replicated statistics (T^{rep}) is more extreme than the observed statistics (T^{obs}), was used to offer a quantitative measurement of the discrepancy. A p-value closer to 0.5 indicates an adequate fit. We chose proportion of participants that never smoked (T_1), mean of positive counts (T_2), and standard deviation (T_3) as discrepancy statistics to highlight model performance in predicting proportion of non-smokers, average smoking level for those who smoked, and overall variation among the population. Both y^{rep} and T^{rep} can be obtained from the draws of model parameters generated from the MCMC output.

4. SIMULATION STUDY

To test the proposed models, a small simulation study was conducted. We generated Y as a mixture of three zero-

Table 1. Model selection statistics for the simulation study. With $*$ we indicate convergence/overfitting issues. With T we indicate the "true" model

Model	Classes	p_D	DIC	p_{D3}	DIC_3
ZIP Fixed	1	5.98	90711.50	–	–
	2	17.88	69464.93	33.73	69480.78
	3	29.83	62216.07	41.88	62228.13
	4	42.22	59569.64	44.13	59571.55
	5	53.60	58508.49	48.96	58503.85
	6	64.97	57918.29	48.32	57901.64
ZIP Mixed	1	3687.05	54457.16	–	–
	2	1429.15	54024.71	456.25	53051.81
	3 (T)	1651.32	53266.77	397.30	52012.75
	4 $*$	1757.37	53419.74	372.57	52034.94
	5 $*$	2205.18	53930.92	361.64	52087.38
	6 $*$	1737.88	53471.24	428.12	52161.48
ZINB Fixed	1	6.45	68092.95	–	–
	2	19.05	61501.29	86.18	61568.42
	3	32.40	58470.05	76.23	58513.88
	4	44.88	57696.74	77.27	57729.13
	5	56.53	57466.28	79.93	57489.68
	6	65.78	57368.19	76.03	57378.44
ZINB Mixed	1	3709.79	54971.85	–	–
	2	1584.41	54536.48	339.81	53291.88
	3	1715.62	53373.07	309.87	51967.33
	4 $*$	1715.29	53443.24	371.13	52099.07
	5 $*$	1728.22	53452.32	354.59	52078.69
	6 $*$	1648.83	53456.95	435.91	52244.04



(a) Three-class ZIP mixed effects model ("true" model). (b) Three-class ZIP fixed effects model.

(c) Three-class ZINB mixed effects model. (d) Three-class ZINB fixed effects model.

Figure 1. Posterior trajectories for the three-class models from the simulation study.

inflated Poisson distributions. The simulated dataset had a sample size of $N = 4500$, each with four observations over time. Both the binomial and the Poisson components contained class specific fixed effect intercepts (α_{k1} and β_{k1}), linear fixed time effects (α_{k2} and β_{k2}), quadratic fixed time effects (α_{k3} and β_{k3}), and random intercepts (b_i) for the Poisson component. One binary covariate and one 5-level categorical covariate were also generated to be associated with class membership probabilities. We dummy coded these two variables such that we had $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{k6})$ for $k = 2$ and 3.

We then fitted one to six class fixed effects ZIP/ZINB models and mixed effects ZIP/ZINB models to the simulated data. Table 1 presents model comparison statistics for the fitted models. The mixed effects ZIP/ZINB models with 4 or more classes were only able to identify three classes with the remaining ones having 0 or very few individuals, thus their DIC statistics need to be interpreted with care. This overfitting phenomenon is well known and expected. Models with a number of classes higher than 4 also tend to have some convergence issues. The DIC had the lowest value for the three-class mixed effects ZIP model, thus correctly identifying the "true" data generating model. A slightly overparametrized specification, the three-class mixed effects ZINB, showed a slightly higher DIC value and slightly smaller DIC_3 value compared with the three-class mixed effects ZIP model. However, the posterior means of the dispersion parameters (i.e. ϕ_1, ϕ_2, and ϕ_3) from the three-class mixed effects ZINB model were all approaching to zeros, indicating that the ZINB components were degenerating into ZIP distributions. Fixed effects ZINB models had lower DIC values than the fixed ZIP models but higher DIC values compared with all ZIP mixed effect models. By introducing random effects or additional dispersion parameters, less classes were needed to achieve an optimal fit. On the other hand, ignoring the individual variations would lead to incorrect classification and biased parameter estimates. We notice that DIC leads to very large values of the effective number of parameters p_D in mixed models because of its treatment of the random effect as a parameter. This feature is somewhat mitigated with the corresponding complexity parameter P_{D3}, as discussed in [4]. Figure 1 shows posterior and true trajectory patterns of y over x for the three-class models. The "true" model (with or without dispersion parameters) recovers very well the true latent trajectories, while their fixed counterparts appear to fit very different curves. The class proportions for class 1 to 3 were 65.96%, 16.60%, and 17.44%, respectively. These were almost identical with the true class proportions (65.29%, 16.91%, and 17.80%). True values of all parameters were contained in their 95% highest posterior density (HPD) intervals.

5. APPLICATION TO REAL DATA

To model the change of smoking behavior from early adolescence to adulthood and to identify latent subgroups from the population, we use data collected from the Add Health study. As described in the introduction, data from wave 1 to 4 will be combined to assess the full age range from early adolescence through the transition to adulthood. We used complete cases of the publicly available subsamples ($N = 2923$). To examine possible baseline risk factors for smoking patterns, we allow gender, race, peer smoking, and household smoking as covariates to influence class membership probabilities. Peer smoking was measured as the number of friends out of three best friends that were smokers and household smoking was a binary variable indicating whether or not there were smokers in the household. As a result, z_i in Equation (3) represented an 8×1 vector of covariates including an intercept and indicators for males, Asian descent, African descent, Hispanic, Native American and other, peer smoking, and household smoking. Female Caucasians, with no additional smokers in their household were set to be the reference group.

We ran a series of latent class models with the number of classes K ranging from two to five. Within each class, we fitted a fixed effects ZIP or ZINB model, a mixed effects ZIP or ZINB model as in Equations (1) and (2). As suggested in the literature, the developmental trajectories of smoking are not linear but curvilinear, thus for both the zero-inflation component and the count component, covariates vector x_{it} in Equations (4) and (5) comprised an intercept term, a linear age effect (age), and a quadratic age effect (age^2).

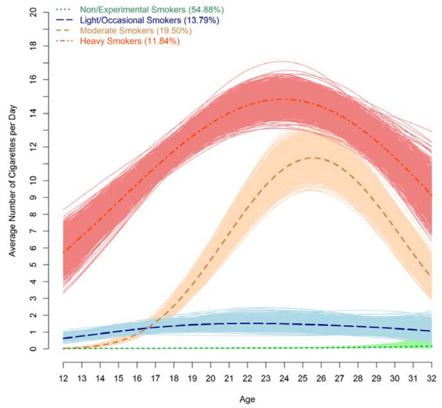
Models with different classes were fitted in R [30] using the MCMC algorithm as described in Section 3. The R code was developed by the authors and some parts were adapted from the code in Dr. Brian Neelon's website: http://people.musc.edu/~brn200/r/. Non-informative priors were specified for each parameter. Specifically, we had $\mu_\alpha = \mu_\beta = \mu_\gamma = \mu_{b_i} = 0$ and $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 100$ for regression parameters $\{\alpha_k, \beta_k, b_i, \gamma_k\}$, $a = 0.001, b = 0.001$ for dispersion parameter ϕ_k and variance of the random intercepts σ_k^2. For ZIP fixed effect models, we ran 400,000 iterations for each model, discarding the first 80,000 for burn-in. We then obtained 1 draw from every 100 iterations for thinning to reduce autocorrelation. As complexity increases for ZINB latent class models, we ran the same number of iterations for ZINB latent class models but allowed them to have a longer burn-in period of 240,000.

We examined trace plots and autocorrelation function plots for all parameters from all models. Specifications with more than 4 classes displayed issues of convergence for the ZIP mixed and ZINB (fixed and mixed), thus we did not pursue them further. When the number of classes is less than 4, all trace plots showed chains with good mixing properties, providing evidence of convergence to their stationary distribution. It is worth mentioning that one of the main challenges of Bayesian analysis of finite mixture models is "label switching". That is, due to the invariance of the likelihood under relabeling of the latent classes, the marginal posterior distributions for the parameters will be identical for each latent class, and therefore, during a MCMC run, the label of a certain class could switch to the label of another class. As a consequence of label switching, the class membership probabilities will be $1/K$ for every participant and the posterior distribution of the parameters will be highly symmetric and multimodal [41]. Thus, label switching results in misleading parameter estimates. Several online or post-hoc algorithms have been developed to relabel the latent classes [38, 41]. We carefully examined the MCMC output, however, and we found no evidence of label switching in our estimates. It is plausible that the inclusion of class membership covariates helped with the identifiability of the classification. We have also ran Poisson-LCM and NB-LCM (with no zero inflation) with and without random effects. As expected, the values of DIC were much larger compared with models with zero-inflation. Thus, the non-zero-inflated models were not considered for further comparison. Models with just one class performed rather poorly.

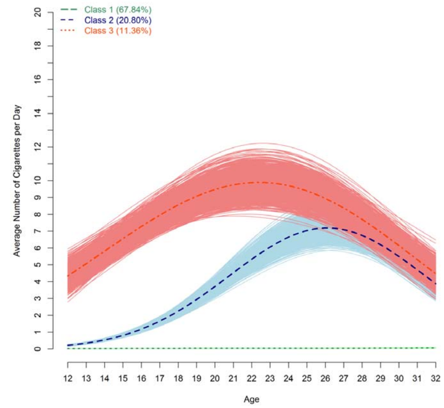
Table 2 presents DIC statistics for ZIP-LCM and ZINB-LCM with and without random effects. The four-class mixed effects ZIP model had the lowest DIC and the three-class mixed effects ZIP model had the lowest DIC_3 among all models. When comparing different models, regardless of the

Table 2. DIC statistics for the smoking study, with $*$ we indicate convergence/overfitting issues

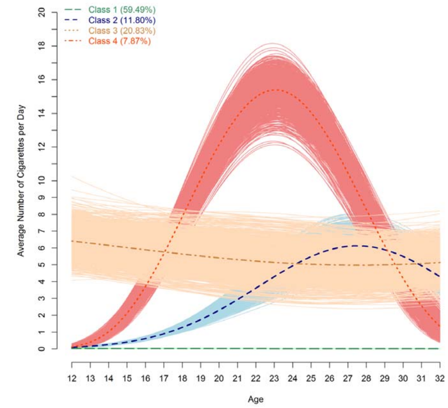
Model	Classes	p_D	DIC	p_{D3}	DIC_3
ZIP Fixed	1	5.85	49709.50	–	–
	2	19.91	38031.14	34.81	38046.05
	3	34.38	35757.29	40.58	35763.48
	4	46.20	34514.93	46.50	34520.93
	5	60.96	33761.10	54.07	33754.22
	6	73.65	33200.86	55.78	33182.99
ZIP Mixed	1	2378.20	32607.25	–	–
	2	1252.58	30361.35	315.40	29424.17
	3	775.80	29722.38	252.31	29198.89
	4	183.98	29507.50	299.89	29623.42
	5 $*$	1626.75	30413.51	232.59	29019.35
	6 $*$	1257.97	30081.44	308.76	29132.23
ZINB Fixed	1	6.41	36210.88	–	–
	2	20.84	32769.90	77.95	32827.01
	3	35.61	31875.94	88.37	31928.70
	4	50.61	31552.68	74.45	31576.51
	5 $*$	−92.46	31412.97	130.06	31635.50
	6 $*$	−991.21	30429.16	142.96	31563.34
ZINB Mixed	1	2242.15	30688.73	–	–
	2	174.56	30538.53	419.58	30783.55
	3	21.23	30324.10	410.13	30712.99
	4	190.10	30618.71	423.75	30852.36
	5 $*$	159.07	30531.59	331.47	30703.99
	6 $*$	423.46	30910.59	417.45	30904.58



(a) Four-class fixed effects ZINB-LCM.



(b) Three-class mixed effects ZIP-LCM.



(c) Four-class mixed effects ZIP-LCM.

Figure 2. Posterior class trajectories for 3 candidate models. Each color represents one smoking class and dashed lines are posterior mean trajectories.

Table 3. Posterior means and 95% credible intervals for the four-class fixed effects ZINB-LCM

Class (%)	Component	Parameter (Covariate)	Posterior Mean	95% Credible Interval
Non/Experimental	Binomial	α_{11} (Intercept)	-7.672	$(-15.449, 0.380)$
(54.88%)		α_{12} (Linear Age)	0.847	(0.151, 1.497)
		α_{13} (Quadratic Age)	-0.017	$(-0.030, -0.003)$
	NB	β_{11} (Intercept)	-9.851	$(-16.07, -2.211)$
		β_{12} (Linear Age)	0.698	(0.002, 1.275)
		β_{13} (Quadratic Age)	-0.012	$(-0.025, 0.004)$
		ϕ_1 (Dispersion Parameter)	0.044	(0.000, 0.492)
Light/Occasional	Binomial	α_{21} (Intercept)	-7.775	$(-12.157, -3.923)$
(13.79%)		α_{22} (Linear Age)	0.630	(0.292, 1.021)
		α_{23} (Quadratic Age)	-0.012	$(-0.021, -0.005)$
	NB	β_{21} (Intercept)	-4.896	$(-6.907, -2.989)$
		β_{22} (Linear Age)	0.502	(0.322, 0.694)
		β_{23} (Quadratic Age)	-0.010	$(-0.015, -0.006)$
		ϕ_2 (Dispersion Parameter)	0.306	(0.175, 0.460)
Moderate	Binomial	α_{31} (Intercept)	12.140	(9.396, 15.151)
(19.50%)		α_{32} (Linear Age)	-1.036	$(-1.317, -0.778)$
		α_{33} (Quadratic Age)	0.020	(0.014, 0.026)
	NB	β_{31} (Intercept)	-9.797	$(-11.483, -8.184)$
		β_{32} (Linear Age)	0.978	(0.839, 1.121)
		β_{33} (Quadratic Age)	-0.019	$(-0.022, -0.016)$
		ϕ_3 (Dispersion Parameter)	0.490	(0.407, 0.584)
Heavy	Binomial	α_{41} (Intercept)	1.069	$(-2.800, 4.728)$
(11.84%)		α_{42} (Linear Age)	-0.298	$(-0.634, 0.058)$
		α_{43} (Quadratic Age)	0.007	$(-0.001, 0.015)$
	NB	β_{41} (Intercept)	-0.563	$(-1.499, 0.260)$
		β_{42} (Linear Age)	0.278	(0.204, 0.363)
		β_{43} (Quadratic Age)	-0.006	$(-0.008, -0.004)$
		ϕ_4 (Dispersion Parameter)	0.336	(0.283, 0.389)

number of classes, fixed effects ZIP models had the highest DIC values, whereas mixed effects ZIP models had the lowest DIC values. Mixed effect ZINB models, being the most flexible and complicated models had higher DIC values than the mixed effects ZIP models, indicating that they over-fit the data. For fixed effects ZIP models, the DIC kept decreasing as the number of classes increased and it was the lowest for the six-class model. In the absence of random effects and parameters of over-dispersion, more classes were needed to explain the heterogeneity in the data. We excluded the fixed effects ZIP models and the mixed effects ZINB models for further analyses.

We kept the three and four-class mixed effects ZIP models and the four-class fixed effects ZINB model (the converging model without a random effect with the lowest DIC value). Figure 2 shows the cigarette smoking posterior trajectories for these three models, with each color representing a different smoking pattern. Posterior predictive checking was also performed on these three candidate models as described in Section 3.3. The posterior predictive p-values corresponds to proportion of four-time zeros (T_1), mean of positive counts (T_2), and standard deviation (T_3) were 0.33, 0.76, and 0.92 for the four-class fixed effects ZINB models, were 0.27, 1, and 1 for the three-class mixed effects ZIP model, and were 0.47, 0.98, and 1 for the four-class mixed effects ZIP model. Although, the mixed effects ZIP models had a lower DIC, posterior predictive checks show that it always tends to overestimate the positive mean and standard deviation of the smoking level. It is possible that the random effects model is too flexible for the data and assumes too much variability. The fixed effects ZINB model also displays a tendency to over-estimate positive mean and standard deviation compared with the observed data but Bayesian p-values were in an acceptable range. Based on the predictive checks we chose the four-class fixed effects ZINB model as our final model for the smoking study.

Posterior means and 95% credible intervals of αs and βs for the four-class fixed effects ZINB-LCM are presented in Table 3. Figure 3 presents posterior trajectories for probability of being a non-smoker (structural zero) and posterior trajectories of average number of cigarettes smoked given that the participant smokes (i.e. from the count process). These four smoking patterns differ by several aspects, such as level of smoking, initial time of smoking, turning point, and rate of change. The first class comprised 54.88% of the participants and the trajectory pattern was characterized by a very high probability of being a non-smoker and an average smoking level around 0 cigarettes. As shown in Figure 3, this class (the upper right plot) also included some participants who tried smoking occasionally at a very low level, especially after age 20. We labeled participants from this class as "non/experimental smokers". Class 2 included 13.79% of the participants and was characterized by a relatively high probability of being a non-smoker and a relatively low level of smoking. Participants from class 2 were termed as "light/occasional smokers".

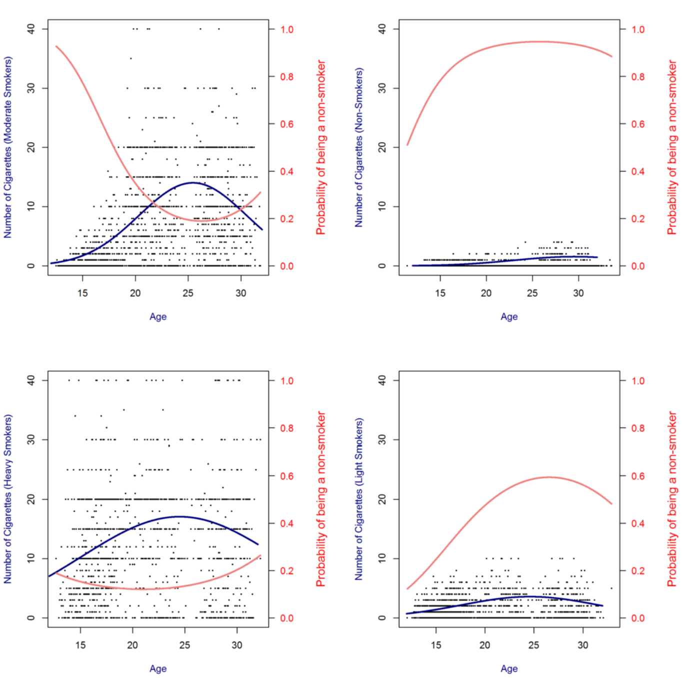


Figure 3. Posterior class trajectories for mean smoking level from the count process (blue) and probability of being a non-smoker (red) from the four-class fixed effects ZINB-LCM. Black dots are average number of cigarettes smoked per day.

Both the probability of being a non-smoker and the smoking level increased until around 25 and then decreased after. Class 3 had 11.84% of participants and we called this group "moderate smokers". Participants from this class had a high initial probability of being a non-smoker and then a rapid decreasing trend until the late 20s. The initial level of smoking was low in adolescence and then increased rapidly until the middle 20s. Only 11.84% of the participants were in class 4, which we described it as "heavy smokers". This class had a relatively stable and low probability of being a non-smoker from adolescence to adulthood (as shown in Table 3, both linear age (α_{42}) and quadratic age effects (α_{43}) were not significant for the heavy smokers class). The level of smoking was the highest among all classes.

While examining risk factors' influence on class membership probabilities, we found that gender, ethnicity, household smoking, and peer smoking all had significant effects on class assignment and smoking level. Figure 4 shows the weighted average cigarette smoking trajectories by different gender, ethnicity, and household smoking. Having household smoking clearly shifted up the cigarettes smoking level. Different ethnic groups also had different average levels of smoking throughout the whole age range. In particular, Caucasian and Native American/Other participants tend to smoke more cigarettes and African descent participants tend to smoke the least number of cigarettes.

Table 4 shows the predicted class membership probabilities for different covariate profiles. Compared with females, males had higher probabilities of being in the moderate and heavy smoking classes and lower probabilities of being in the non/experimental and light smoking classes. Ethnic disparities also exist in the probability of engaging in different smoking patterns. Native American/Other participants had the highest probability of belonging to the heavy smoking group and the lowest probability of belonging to the non-smoking group. Caucasian participants had the second highest probability of being in the heavy smoking group and the

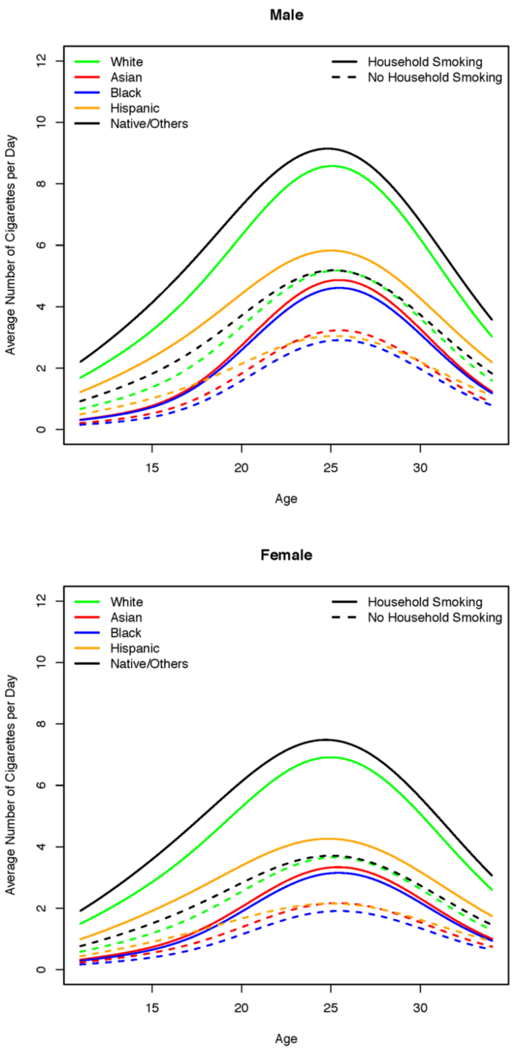


Figure 4. Weighted average cigarettes smoking trajectories by gender, ethnicity, and household smoking.

second lowest probability of being in the non-smoking group. Asian and African descent participants are more likely to be in the non-smoking class and less likely to be in the heavy smoking class. Hispanic participants had a high probability of being light smokers. These findings are consistent with White et al. [43] and Evans-Polce et al. [8]. Having smokers in the household increased participants' probability of smoking and level of smoking. In particular, those who reported having smokers in the household had a 0.109 decreased probability, a 0.136 increased probability, and a 0.055 increased probability of being non-smokers, moderate smokers, and heavy smokers, respectively. Participants who had peers who smoke were much less likely to be non-smokers and more likely to engage in light and occasional smoking.

6. DISCUSSION

We described latent class models for analyzing longitudinal count data that exhibit an excess of zeros. The modeling approach has several advantages over other commonly used methods. First, since the latent class variable can effectively summarize distinctive patterns of change in longitudinal data and the latent binary variable can identify whether an observation comes from a zero-inflation process or a regular count process at each time point, this modeling choice is very flexible to take into account both unobserved static and time varying heterogeneity. Second, it allows individual characteristic factors to be included in the model by influencing the latent class membership and time varying covariates, such as time and age, to be directly associated with the outcome. In addition, the joint estimation of the class membership and risk factors is more general than the traditional two-stage approach which does not take into account of the uncertainty in class membership. The method was applied to developmental trajectories of cigarette smoking behavior

Table 4. Predicted Class Membership Probabilities

Covariates	Non or Experimental	Light or Occasional	Moderate	Heavy
Gender				
Male	0.364	0.248	0.244	0.143
Female	0.399	0.338	0.149	0.114
Race				
Caucasian	0.309	0.247	0.239	0.204
Asian	0.419	0.333	0.228	0.021
African	0.560	0.199	0.216	0.026
Hispanic	0.352	0.400	0.127	0.121
Native/Other	0.268	0.288	0.173	0.271
Household Smoking				
Yes	0.327	0.252	0.224	0.197
No	0.436	0.335	0.169	0.061
Peer Smoking				
3 peers	0.064	0.575	0.054	0.307
None	0.593	0.206	0.191	0.010

from early adolescence to adulthood. We were able to identify four distinct groups of age-varying smoking trajectories: non/experimental smokers, light/occasional smokers, moderate smokers, and heavy smokers. Class specific smoking patterns differ not only by the probability of being a smoker and level of smoking but also by characteristics related to onset, escalation, and leveling off of smoking. Our results provide insights into gender and ethnic disparities on smoking patterns. In addition, we found that having smokers as peers and/or in the household was significantly correlated with higher levels of cigarette smoking, especially among heavy smokers. Although many educational and prevention programs exist focusing on smoking reduction, our findings suggest that more effective strategies may need to be age, gender, and ethnicity specific. Initiatives could also target the reduction of adolescents' exposure to smoking by encouraging household indoor smoking restrictions. Rodriguez et al. [35] suggested that indoor smoking restrictions, even when parents themselves smoke, could decrease exposure to peer smoking and decrease adolescents' smoking risk at a higher level.

This paper compared ZIP-LCM and ZINB-LCM with and without random effects. The flexibility of a mixed model is very appealing; however, such a complex model might occasionally lead to overfit of the data and offer less meaningful interpretation. As a result, for this dataset a four-class fixed effects ZINB model is the most useful, because of its parsimony in estimating the number of latent classes and ensuring enough flexibility to model the zero-inflation and over-dispersion within different smoking patterns. An alternative approach to model zero-inflated count data is using "two-part models", which include a Poisson hurdle model and a negative binomial hurdle model [22, 17]. The main difference between zero-inflated models and two-part models is how they deal with different types of zeros: while the count process of a two-part model is a zero-truncated Poisson or zero-truncated negative binomial model (i.e. the distribution of the response variable cannot have a value of zero), the count process of a mixture model can produce zeros [45]. Min and Agresti [21] suggested that zero-inflated models tend to be unstable when zero-deflation exists at some levels of the covariates. Although we did not encounter this problem from the simulation study, our proposed models can be easily adapted to the more robust hurdle latent class formulation. We intend to investigate hurdle models in future research. Another aspect worthy of consideration is a direct estimation of the number of classes, treating it as an unknown parameter. For simpler mixtures transdimensional algorithms like reversible jumps MCMC (see Richardson and Green [33]) or the birth and death processes MCMC by Stephens [40] have been proposed. These methods are usually very complex to implement, and the mixing can be very slow, especially with more structured likelihoods like the ones we considered in our work. Although mixtures with independent weights lose the connection to demographic variables, they allow for a certain level of simplification in estimating the number of classes, and in the case of zero inflated distributions are the subject of ongoing research.

A limiting aspect of the dataset we used is that observations were collected using a cohort sequential design. The baseline age ranged from 13 to 21 years and each participant only had four measurements with different time intervals. Though there was overlapping in age between different cohorts, each age cohort only contributes a different segment of the overall curve. It is possible that a trajectory for the whole age range is biased due to the small number of measurements. As for future analysis of the smoking data, the baseline age (i.e. the cohort effect) could be considered in the model by either affecting the class membership probability or as a random effect. On the applied side, our model can also be extended to accommodate multiple outcomes, with dual trajectory models linking the trajectory patterns of two behaviors [15]. As a more general consideration, zero-inflated latent class models can be used for a wide variety of applications when the interest is to model rare events or behaviors that are less commonly endorsed. In addition, there is a growing interest in studying multiple health risky behaviors and implementing specific interventions that target multiple co-occurrence of such behaviors. At the moment there is still surprisingly little understanding of the basic principles of multiple health behavior change, as discussed in Prochaska et al. [29].

APPENDIX A. MIXED EFFECTS ZINB-LCM MODEL LIKELIHOOD

We present the likelihood function for the mixed effects ZINB-LCM model as mentioned in Section 3, where $\Theta_{ZINB\,Mixed} = \{\alpha_k, \beta_k, \gamma_k, C_i, \phi_k, b_{ki}\}$.

$$
\begin{aligned}
&P(Y|\Theta_{ZINB\,Mixed})\\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i=k)\prod_{t=1}^{T}\Pr(Y_{it}|C_i=k)\\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\Bigg\{\prod_{t:Y_{it}=0}\left[p_{itk}+(1-p_{itk})\left(\frac{\phi}{\mu_{it}+\phi}\right)^{\phi}\right]\\
&\quad+\prod_{t:Y_{it}\neq 0}(1-p_{itk})\frac{\Gamma(\phi+y_{it})}{y_{it}!\Gamma(\phi)}\left(\frac{\mu_{it}}{\mu_{it}+\phi}\right)^{y_{it}}\left(\frac{\phi}{\mu_{it}+\phi}\right)^{\phi}\Bigg\}\\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\frac{e^{z_i^T\gamma_k}}{\sum_{h=1}^{K}e^{z_i^T\gamma_h}}\Bigg\{\prod_{t:Y_{it}=0}\Bigg[\frac{1}{e^{-(x_{it}^T\alpha_k)}+1}+\frac{1}{e^{x_{it}^T\alpha_k}+1}\\
&\quad\times\left(\frac{\phi}{e^{x_{it}^T\beta_k+b_{ki}}+\phi}\right)^{\phi}\Bigg]+\prod_{t:Y_{it}\neq 0}\frac{1}{e^{x_{it}^T\alpha_k}+1}\frac{\Gamma(\phi+y_{it})}{y_{it}!\Gamma(\phi)}\\
&\quad\times\left(\frac{e^{x_{it}^T\beta_k+b_{ki}}}{e^{x_{it}^T\beta_k+b_{ki}}+\phi}\right)^{y_{it}}\left(\frac{\phi}{e^{x_{it}^T\beta_k+b_{ki}}+\phi}\right)^{\phi}\Bigg\}.
\end{aligned}
$$

APPENDIX B. MODEL SELECTION DETAILS

The DIC was introduced by Spiegelhalter et al. [39] for comparing complex hierarchical models and it has the following form,

$$\begin{aligned} DIC &= \overline{D(\theta)} + p_D \\ &= E[D(\theta)|y] + (E[D(\theta)|y] - D(E[\theta|y])) \\ &= 2\overline{D(\theta)} - D(\tilde{\theta}) \\ &= -4E[\log f(y|\theta)|y] + 2\log f(y|\tilde{\theta}), \end{aligned}$$

where $\tilde{\theta}$ is an estimate of parameters depending on the distributional form of y. The posterior mean $\overline{\theta} = E[\theta|y]$ is often used for $\tilde{\theta}$. With $\overline{D(\theta)}$ we indicate the posterior mean of the deviance and it offers summary information on how much discrepancy exists between the model and the data. p_D measures the difference between the posterior mean of the deviance (i.e. $\overline{D(\theta)}$) and the deviance evaluated at the posterior mean of the parameters (i.e. $D(\tilde{\theta})$). It provides a way of assessing effective number of parameters. Thus, the DIC assesses both a Bayesian measure of a model fit and the complexity of the model. Similarly to AIC and BIC, a model with a smaller DIC is usually preferred.

Celeux et al. [4] provided an extension of DIC in the case of finite mixture models, which they referred to as DIC_3. DIC_3 has the same form as the traditional DIC except that it estimates $D(\tilde{\theta})$ by using the MCMC predictive density, which is a weighted average of the posterior mean of the marginal likelihood from all classes. We call this new deviance of the mean as $D(\tilde{\theta})_3$ and the new effective size of parameters as p_{D3}. Both $\overline{D(\theta)}$ and $D(\tilde{\theta})_3$ can be approximated using M simulated values $\theta^{(1)}, \ldots, \theta^{(M)}$ from MCMC chains. For ZIP latent class models, $\theta^{(m)} = (\mu^{(m)}, p^{(m)})$ and for ZINB latent class models, $\theta^{(m)} = (\mu^{(m)}, p^{(m)}, \text{and } \phi^{(m)})$. In particular,

$$\begin{aligned} \overline{D(\theta)} &= -2\frac{1}{M}\sum_{m=1}^{M}\log\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}^{(m)} f(y_{ik}|\theta_{ik}^{(m)}), \\ D(\tilde{\theta})_3 &= -2\log\frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}^{(m)} f(y_{ik}|\theta_{ik}^{(m)}). \end{aligned}$$

In the simulation study and real data application, we used both the original DIC and the DIC_3 as criteria for model selection.

ACKNOWLEDGEMENTS

We would like to thank Guan-Hua Huang, for help and suggestions during the early phases of this research, Brian Neelon for his prompt response to our requests of clarification, and Lisa Harlow for her editorial comments.

Received 15 November 2016

REFERENCES

[1] Akaike, H. (1998). "A Bayesian analysis of the minimum AIC procedure." In *Selected Papers of Hirotugu Akaike*, 275–280. Springer. MR1486823

[2] Bakk, Z., Oberski, D. L., and Vermunt, J. K. (2017). "Relating latent class assignments to external variables: standard errors for correct inference." *Political Analysis*, 22(4): 520–540.

[3] Bernat, D. H., Erickson, D. J., Widome, R., Perry, C. L., and Forster, J. L. (2008). "Adolescent smoking trajectories: results from a population-based cohort study." *Journal of Adolescent Health*, 43(4): 334–340.

[4] Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1(4): 651–673. MR2282197

[5] Chen, P. and Jacobson, K. C. (2012). "Developmental trajectories of substance use from early adolescence to young adulthood: Gender and racial/ethnic differences." *Journal of Adolescent Health*, 50(2): 154–163.

[6] Colder, C. R., Mehta, P., Balanda, K., Campbell, R. T., Mayhew, K., Stanton, W. R., Pentz, M. A., and Flay, B. R. (2001). "Identifying trajectories of adolescent smoking: an application of latent growth mixture modeling." *Health Psychology*, 20(2): 127–135.

[7] Dagne, G. A. (2004). "Hierarchical Bayesian analysis of correlated zero-inflated count data." *Biometrical Journal*, 46(6): 653–663. MR2108609

[8] Evans-Polce, R. J., Vasilenko, S. A., and Lanza, S. T. (2015). "Changes in gender and racial/ethnic disparities in rates of cigarette use, regular heavy episodic drinking, and marijuana use: ages 14 to 32." *Addictive Behaviors*, 41: 218–222.

[9] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons. MR2830137

[10] Fu, Y. Z., Chu, P. X., and Lu, L. Y. (2014). "A Bayesian approach of joint models for clustered zero-inflated count data with skewness and measurement errors." *Journal of Applied Statistics*, (ahead-of-print): 1–17. MR3302803

[11] Gelman, A., Meng, X.-L., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6(4): 733–760. MR1422404

[12] Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. J. (2006). "Bayesian analysis of zero-inflated regression models." *Journal of Statistical Planning and Inference*, 136(4): 1360–1375. MR2253768

[13] Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press. MR2797563

[14] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods (pp. 180)*. Springer Science & Business Media. MR2648134

[15] Jones, B. L., Nagin, D. S., and Roeder, K. (2001). "A SAS procedure based on mixture models for estimating developmental trajectories." *Sociological Methods & Research*, 29(3): 374–393. MR1816410

[16] Keel, P. K., Fichter, M., Quadflieg, N., Bulik, C. M., Baxter, M. G., Thornton, L., Halmi, K. A., Kaplan, A. S., Strober, M., Woodside, D. B., et al. (2004). "Application of a latent class analysis to empirically define eating disorder phenotypes." *Archives of General Psychiatry*, 61(2): 192–200.

[17] King, G. (1989). "Event count models for international relations: generalizations and applications." *International Studies Quarterly*, 123–147.

[18] Klein, N., Kneib, T., and Lang, S. (2015). "Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data." *Journal of the American Statistical Association*, 110(509): 405–419. MR3338512

[19] Mahalik, J. R., Levine Coley, R., McPherran Lombardi, C., Doyle Lynch, A., Markowitz, A. J., and Jaffee, S. R. (2013). "Changes in health risk behaviors for males and females from early adolescence through early adulthood." *Health Psychology*, 32(6): 685.

[20] McArdle, J. J. and Epstein, D. (1987). "Latent growth curves

within developmental structural equation models." *Child Development*, 110–133.

[21] Min, Y. and Agresti, A. (2005). "Random effect models for repeated measures of zero-inflated count data." *Statistical Modelling*, 5(1): 1–19. MR2133525

[22] Mullahy, J. (1986). "Specification and testing of some modified count data models." *Journal of Econometrics*, 33(3): 341–365. MR0867980

[23] Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*, pp. 345–368 Newbury Park, CA: Sage Publications.

[24] Muthén, B. and Muthén, L. K. (2000). "Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes." *Alcoholism: Clinical and Experimental Research*, 24(6): 882–891.

[25] Muthén, B. and Shedden, K. (1999). "Finite mixture modeling with mixture outcomes using the EM algorithm." *Biometrics*, 55(2): 463–469.

[26] Nagin, D. S. (1999). "Analyzing developmental trajectories: a semiparametric, group-based approach." *Psychological Methods*, 4(2): 139.

[27] Neelon, B., O'Malley, A. J., and Normand, S.-L. T. (2011). "A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity." *Biometrics*, 67(1): 280–289. MR2898840

[28] Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). "A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use." *Statistical Modelling*, 10(4): 421–439. MR2797247

[29] Prochaska, J. J., Spring, B., and Nigg, C. R. (2008). "Multiple health behavior change research: an introduction and overview." *Preventive Medicine*, 46(3): 181–188.

[30] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/

[31] Reboussin, B. A., Song, E.-Y., Shrestha, A., Lohman, K. K., and Wolfson, M. (2006). "A latent class analysis of underage problem drinking: Evidence from a community sample of 16–20 year olds." *Drug and Alcohol Dependence*, 83(3): 199–209.

[32] Reinecke, J. and Seddig, D. (2011). "Growth mixture models in longitudinal research." *Advances in Statistical Analysis*, 95(4): 415–434. MR2862975

[33] Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 731–792. MR1483213

[34] Rindskopf, D. and Rindskopf, W. (1986). "The value of latent class analysis in medical diagnosis." *Statistics in Medicine*, 5(1): 21–27.

[35] Rodriguez, D., Tscherne, J., and Audrain-McGovern, J. (2007). "Contextual consistency and adolescent smoking: Testing the indirect effect of home indoor smoking restrictions on adolescent smoking through peer smoking." *Nicotine & Tobacco Research*, 9(11): 1155–1161.

[36] Schwarz, G. et al. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6(2): 461–464. MR0468014

[37] Shiyko, M., Li, Y., and Rindskopf, D. (2012). "Poisson growth mixture modeling of intensive longitudinal data: An application to smoking cessation behavior." *Structural Equation Modeling*, 19(1): 65–85. MR2880531

[38] Sperrin, M., Jaki, T., and Wit, E. (2010). "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models." *Statistics and Computing*, 20(3): 357–366. MR2725393

[39] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639. MR1979380

[40] Stephens, M. (2000a). "Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods." *Annals of Statistics*, 28(1): 40–47. MR1762903

[41] Stephens, M. (2000b). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. MR1796293

[42] Wenzel, S. E. (2012). "Asthma phenotypes: the evolution from clinical to molecular approaches." *Nature Medicine*, 18(5): 716–725.

[43] White, H. R., Nagin, D., Replogle, E., and Stouthamer-Loeber, M. (2004). "Racial differences in trajectories of cigarette use." *Drug and Alcohol Dependence*, 76(3): 219–227.

[44] White, H. R., Pandina, R. J., and Chen, P.-H. (2002). "Developmental trajectories of cigarette use from early adolescence into young adulthood." *Drug and Alcohol Dependence*, 65(2): 167–178.

[45] Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science & Business Media. MR2722501

Si Yang
Department of Computer Science and Statistics
University of Rhode Island
USA
E-mail address: si_yang@my.uri.edu

Gavino Puggioni
Department of Computer Science and Statistics
University of Rhode Island
246 Tyler Hall, 9 Greenhouse Rd.
02881 Kingston, RI
USA
E-mail address: gpuggioni@uri.edu