2015

A BAYESIAN ZERO-INFLATED GENERALIZED GROWTH MIXTURE MODEL FOR ADOLESCENT HEALTH RISK BEHAVIORS

Si Yang
University of Rhode Island, yangsi06@gmail.com

A BAYESIAN ZERO-INFLATED GENERALIZED GROWTH MIXTURE MODEL FOR ADOLESCENT HEALTH RISK BEHAVIORS

BY

SI YANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE AND STATISTICS

UNIVERSITY OF RHODE ISLAND

2015

MASTER OF SCIENCE THESIS

OF

SI YANG

APPROVED:

Thesis Committee:

Major Professor Gavino Puggioni

Natallia V. Katenka

Robert Laforge

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2015

ABSTRACT

Using longitudinal data to model change patterns of behaviors is a major interest in the field of developmental psychology and behavioral science. As heterogeneity often exists in the population, researchers are more and more interested in a topological representation of the growth trajectories. That is, to identify distinct trajectories depending on their individual characteristics. Latent class models(LCM) are flexible methods of modeling unobserved heterogeneity in a population and it has been recently extended to analyzing longitudinal data. Latent class growth models(LCGM) assume individuals come from a finite number of latent classes and individuals share the same growth trajectory within each class. However, there is little literature on applying latent class models on zero-inflated outcomes. When the interest is to model rare events or behaviors that are less commonly endorsed, such as health risk behaviors (e.g., smoking, heroin use, suicide attempts, etc.), we often encounter a lot of zero responses causing the distribution of the outcome variable to exhibit a large spike at zero. This work focuses on developing latent class growth models for zero-inflated count response variables. Bayesian analysis, well known for its ability to incorporate prior information and greater flexibility to solve complex problems, was used in this paper. Specifically, appropriate prior distributions were specified for the model parameters, likelihood of the data was derived based on the zero-inflated latent class model, and joint posterior distribution was obtained by combining information from the prior and likelihood. Due to the fact that conditional posterior distributions of the model parameters are numerically intractable, simulation based approach Markov Chain Monte Carlo methods were used to approximate and summarize posterior quantities. A simulation study was first conducted to test the performance of the proposed model. As an illustration, data collected from the National Longitudinal

Study of Adolescent Health was then analyzed. This paper modeled the change of cigarettes smoking from early adolescence to adulthood and identified subgroups of trajectory patterns and risk factors contributing to the classification.

ACKNOWLEDGMENTS

First, I would like to thank my major professor, Dr. Gavino Puggioni, who has always been super enthusiastic about research and super supportive of the project. I truly appreciate the large amount of time he has spent on discussing the project with me and checking the R code. I learned so much from him. Without his persistent support and guidance, this thesis would not have been possible. I would also like to thank my committee, Dr. Lisa Harlow, Dr. Natalia Katenka, and Dr. Robert Laforge for their comments and input on the final work.

In addition, a thank you to Dr. Guan-Hua Huang from the Institute of Statistics, National Chiao Tung University, who advised me with the initial model development during my visit there.

And lastly, I would like to thank my parents and my husband for their support and tolerance of me being far away.

PREFACE

This thesis uses a manuscript format. The author is planning to submit the manuscript for publication in Biometrics.

TABLE OF CONTENTS

LIST OF FIGURES

MANUSCRIPT 1

A Bayesian Zero-inflated Latent Class Growth Model for Adolescent Health Risk Behaviors

Planning to submit the manuscript for publication in Biometrics.

Biometrics 000, 000–000 DOI: 000
000 0000

A Bayesian Zero-inflated Latent Class Growth Model for Adolescent Health Risk Behaviors

Si Yang and Gavino Puggioni*

Department of Computer Science and Statistics, University of Rhode Island, Kingston, Rhode Island, U.S.A.

**email:* puggioni@cs.uri.edu

email: si_yang@my.uri.edu

Summary: This paper focuses on developing latent class models for longitudinal data, in addition to having zero-inflated count response variables. The goal is to model discrete longitudinal patterns of change on counts of rare events, for instance, health-risky behavior, and to identify subject-specific covariates associated with latent class probabilities. Two kinds of discrete latent information exist in this type of model: first, a latent categorical variable classifies subgroups with distinct developmental trajectories and then a latent binary variable identifies whether an observation is from a zero-inflation process or a regular count process. Within each class, we fit a zero-inflated Poisson model or a zero-inflated negative binomial model to separately model the probability of structural zeros and the mean trajectories for those from a count process. We propose a joint estimation of the latent variables and regression parameters in the Bayesian framework. Our methods are illustrated through a simulation study and are applied to data on cigarette smoking from the National Longitudinal Study of Adolescent Health.

Key words: Bayesian analysis; Cigarette smoking; Finite mixture model; Latent class growth model; Zero-inflated negative binomial; Zero-inflated Poisson.

1. Introduction

Latent class models (LCMs), also known as finite mixture models, is a flexible method of modeling unobserved heterogeneity in a population. LCM assumes that a heterogeneous group can be reduced to several homogeneous subgroups through minimizing the association among responses across multiple variables. The goal is to categorize subjects into several groups, each one of which contains subjects who are similar to each other and different from subjects in other groups (Muthén and Muthén, 2000). A latent categorical variable is usually 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); and LCM has also been applied to identify phenotypes or genetic susceptibility for diseases based on clinical and biological data (Keel et al., 2004; Muthén and Muthén, 2000; Reboussin et al., 2006; Rindskopf and Rindskopf, 1986; Wenzel, 2012).

LCM has been recently extended to accommodate longitudinally observed data to identify distinct groups of change trajectories within a population. Growth mixture modeling is one of the extensions in a structural equation modeling framework. It is a combination of latent growth curve modeling and latent class modeling (McArdle and Epstein, 1987; Muthén and Shedden, 1999; Muthén and Muthén, 2000). In a latent growth curve model, initial status and slope of change for the outcome variables are considered as random continuous latent growth factors. Thus, a growth mixture model estimates a mean growth curve for each class and also allows individual variations within classes. A detailed description of latent growth curve models and growth mixture models was given by Muthén (2004). Using a semi-parametric strategy, Nagin (1999) developed a group based approach for estimating trajectories for longitudinal data with different types of outcomes. The developmental trajectories

are modeled by having parameters depend on time. In many practical applications, it is also 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) and was used in this paper. Using a frequentist approach, parameters from LCGM can be estimated through a SAS procedure TRAJ written by Jones et al. (2001). 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 (Nagin, 1999). 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 variant covariates can also be included to directly affect the observed outcome. When the observed outcome of interest is a count variable, it is usually characterized with a high incidence of zero counts. As an illustration, consider the data set in Table 1, which provides descriptive statistics of cigarettes smoking for a sample of 3249 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 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 as 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. Two main modeling approaches have been used to analyze count data with excessive zeros: zero-inflated models and zero-truncated models. The former approach

includes a zero-inflated Poisson (ZIP) model and a zero-inflated negative binomial (ZINB) model, and they are often referred to as "mixture models" (Greene, 1994; Lambert, 1992). The latter type is normally known as a "two-part model" in the literature, which includes a Poisson hurdle model and a negative binomial hurdle model (Mullahy, 1986; King, 1989). The main difference between these two modeling approaches 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 (Zuur et al., 2009). Zero-inflated models assume that there are two underlying processes generating zeros, one from the always zero (or structurally zero) process and one from the Poisson or negative binomial process. When observing a zero response in the data, we do not know which process it belongs to. A latent binary variable that follows a Bernoulli distribution could be specified to label structural zeros and non-structural zeros. Compared with zero-truncated models, zero-inflated models are particularly meaningful when there are theoretical justifications for modeling zeros in two separate processes. For instance, in public health and medical studies, zero-inflated models may be conceptualized as allowing zeros to arise from at-risk (susceptible) and not-at-risk (non-susceptible) populations (e.g. a zero count of cigarettes smoked could come from a non-smoker or a smoker who reported zero cigarette during the period of study). Therefore, in this paper, we use zero-inflated models to deal with the abundance of zeros.

In the Bayesian framework, both latent class models and zero-inflated regression models have been developed and applied separately by many researchers. Ghosh et al. (2006) first integrated a data augmentation method with Markov Chain Monte Carlo (MCMC) to generate posterior samples from zero-inflated models. Dagne (2004); Fu et al. (2014), andNeelon et al. (2010) proposed Bayesian analysis for correlated or clustered zero-inflated count data. Klein

et al. (2014) developed Bayesian generalized additive models for data with zero-inflation and over-dispersion. However, there is little literature on the Bayesian analysis of zero-inflated latent class models. The only known study is Neelon et al. (2011). They built a Bayesian 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 subjects who did not use any mental health service. In their paper, a binomial component was used to model the observed zeros and a lognormal component was used to model the right skewed nonzero values. Three classes of subjects 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.

Build on the previous work, this paper presents a Bayesian analysis of latent class growth modeling on zero-inflated count responses. The application of interest is to model trajectories of smoking behavior from adolescent to adulthood. Although a myriad of studies have been done 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 through 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. A few researchers have studied development trajectories of smoking behavior using longitudinal data. For instance, Colder et al. (2001) 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. (2002) 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. (2008) found six distinct trajectories of smoking: nonsmokers, triers, occasional users, early established, late established, and decliners. Chen and Jacobson (2012) 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 above described studies and some other studies (Evans-Polce et al., 2015; Mahalik et al., 2013; White et al., 2004) on smoking trajectories 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 ethnicity play a role in trajectories of cigarette use; and fourth, most of the study used a "two stage" approach that cigarettes 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 ignores the uncertainty in class membership. This paper proposes a joint estimation of the latent class membership and risk factors. We include gender, ethnicity, and some other smoking related risk factors in the model as covariates for smoking patterns. We also use latent class growth modeling with polynomial trends to reflect the curvilinear trends.

The joint posterior distribution of the parameters from the proposed model does not have a closed form, simulation based MCMC methods are used to obtain estimates of unknown parameters. MCMC methods are particularly powerful in dealing with high dimensional and complex data. This paper uses the open source program R to implement the MCMC

algorithm. The rest of this paper is organized as follows: Section 2 presents the proposed zero-inflated latent class growth models. In section 3, prior distributions are chosen for the model parameters, a MCMC algorithm is outlined for sampling from the posterior distribution of the model parameters and the latent class variable, and criteria of model comparisons are also discussed. Section 4 provides a small simulation study on a three-class mixture model. Section 5 illustrates the procedure with real life data. The final section 6 summarizes our findings and discusses directions for future research.

2. Zero-inflated Latent Class Growth Model

A zero-inflated model is a mixture model with a zero mass mixed with a Poisson distribution or a negative binomial distribution. 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:

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

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

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). p_{it} denotes the probability of being in a non-susceptible group and it can be estimated by information from covariates with

a logistic link. Conditioning on an individual is from the susceptible group, his or her count is a random variable from a Poisson distribution with mean μ_{it} or from a negative binomial distribution with mean μ_{it} and dispersion parameter of ϕ. The difference between a ZIP model and a ZINB model is that the ZINB model has an additional dispersion parameter that can also account for over-dispersion generated from positive values. ϕ only takes positive values and a bigger ϕ indicates a higher degree of dispersion. For negative binomial distributions, ϕ can only approach to zero but can never reach it(Hilbe, 2011). In practice, a ZINB model with a value of ϕ close to zero is statistically indistinguishable from a ZIP model.

Y is said to arise from a finite mixture of ZIP or ZINB distributions, if the probability mass function $p(y)$ takes the form of a mixture density for all $y \in Y$ as follows:

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

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 called the weights for each component and they also give the probability of an underlying categorical latent variable C_i taking a value of $k(k = 1, 2, \ldots, K)$. Thus, a latent class model on zero-inflated count responses actually consists of two kinds of unobserved information. First, there is the latent categorical variable C_i, which follows a multinomial distribution: $C_i \sim Multinom(\pi_{i1}, \ldots, \pi_{iK})$. It divides a population into different subgroups. Within each subgroup, $B_{it} \sim Bernoulli(p_{it})$, is another 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, thus for each subject, their class memberships are constrained to be the same over time. However, over time one's response can change from a structural zero process to a count process or vice versa (e.g., a

subject 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:

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

In this way, individuals from the same class share similar growth trajectories. 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:

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

$$\log(\mu_{itk}) = x_{it}^T \beta_k \tag{5}$$

where x_{it} is a $p \times 1$ vector of time varying covariates. In many longitudinal studies, the true trend over time for the underlying mean response is likely to happen in a relatively smooth and monotonically increasing or decreasing 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 (Fitzmaurice et al., 2012). 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. α_k and β_k are class specific regression coefficients for class k.

3. Bayesian Analysis

3.1 *Likelihood and Prior Specification*

Now, 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}. Then the likelihood of obtaining the observed sample given the vector of parameters and the latent variable $\Theta = \{\alpha_k, \beta_k, \gamma_k, C_i\}$ has the following form:

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

Under a Bayesian framework, prior distributions are specified for regression parameters $\{\alpha_k, \beta_k, \gamma_k\}$ and an additional dispersion parameter ϕ_k for ZINB models. We assign multivariate normal priors to all class specific regression parameters and a gamma prior to the dispersion parameter. That is,

$$\begin{aligned}
\pi(\alpha_k) &= \mathcal{N}_p(\mu_\alpha, \sigma_\alpha^2 I_p), \\
\pi(\beta_k) &= \mathcal{N}_p(\mu_\beta, \sigma_\beta^2 I_p), \\
\pi(\gamma_k) &= \mathcal{N}_r(\mu_\gamma, \sigma_\gamma^2 I_r), \\
\text{and} \quad \pi(\phi_k) &= Gamma(a, b).
\end{aligned}$$

Diffuse priors are assigned for the simulation study and real data application so that the posterior estimates will be mostly determined by the data. It may be noted that when prior information on the parameter distributions is available, one may specify different priors for different classes and one can also replace the identity matrix with an informative prior variance-covariance matrix.

3.2 *Posterior Computation*

Assuming prior independence, the joint posterior distribution combined information from the prior and the data is proportional to the multiplication of the prior distribution and the likelihood specified and derived from section 3.1. Specifically, we have

$$P(\Theta|Y) = \frac{P(\Theta)P(Y|\Theta)}{P(Y)} \propto P(\Theta)P(Y|\Theta)$$

Since it is difficult to draw samples from the joint distribution, Gibbs sampler is used to sample from the full conditional distribution of each parameter. The full conditional posterior distributions of each parameter and the latent class variable have the following forms:

$$\pi(\gamma_k|\cdot) \propto \prod_{i=1}^{N}[P(C_i = k|\gamma_k; z_i)]^{I(C_i=k)}\pi(\gamma_k)$$

$$\pi(C_i|\cdot) = Multinom(\rho_{ik}) \propto P(Y_i|C_i, \alpha_k, \beta_k; x_i)P(C_i|\gamma_k; z_i)$$

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

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

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

We introduce another variable D_{it} here which indicates $Y_{it} = 0(D_{it} = 1)$ or $Y_{it} > 0(D_{it} = 0)$ and $D_{it} \sim Binomial(\theta_{it})$. 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, $\theta_{it} = p_{it} + (1 - p_{it})e^{-\mu_{it}}$). As for sampling C_i, ρ_{ik} is the posterior probability that individual i belongs to class k and it is given by

$$\rho_{ik} = \frac{\pi_{ik}(\gamma_k)\prod_{t=1}^{T} dzip(\mu_{itk}, p_{itk})}{\sum_{h=1}^{K}\pi_{ih}(\gamma_h)\prod_{t=1}^{T} dzip(\mu_{ith}, p_{ith})}$$

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

for ZIP models and ZINB models, respectively. Because no closed forms are available for the full conditional posterior distributions of α, β, and γ it is also difficult to draw samples directly from those distributions. We use a Metropolis algorithm to draw samples for these three parameters. As a result, for the ZIP LCGM, the following algorithm can be used to generate samples from the above full conditional distributions:

(1) Assign initial values to α_k and β_k for $k = 1, \ldots, K$, to γ_k for $k = 2, \ldots, K$, and to class membership indicator C_i;

(2) for $k = 2, \ldots, K$, update γ using random walk Metropolis;

(3) sample C_i from the multinomial distribution based on posterior probability ρ; and

(4) for $k = 1, \ldots, K$, update α_k and β_k using a random walk Metropolis.

Similar steps can be used for the ZINB LCGM except that for $k = 1, \ldots, K$, we also update ϕ_k using random walk Metropolis-Hastings. The Metropolis algorithm proceeds by sampling a proposal value nearby the current value using a symmetric proposal distribution (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 any kind of distribution, in practice, only a distribution that is close to our target distribution will generate a sufficient 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 $(X^TX)^{-1}$ (Hoff, 2009) to improve mixing, the proposal densities we use for updating α_k, β_k, and γ_k are $\mathcal{N}_p(\alpha_k^{old}, (X_k^TX_k)^{-1})$, $\mathcal{N}_p(\beta_k^{old}, (X_k^TX_k)^{-1})$, and $\mathcal{N}_r(\gamma_k^{old}, (Z_k^TZ_k)^{-1})$, respectively. Since ϕ can only be positive values, we propose ϕ^{new} from a log-normal distribution, i.e., $\ln\mathcal{N}(log(\phi^{old}), \sigma_\phi^2)$.

The performance of the MCMC algorithm is monitored by inspecting values of the acceptance rate, constructing graphs such as trace plot and autocorrelation function plot, and

computing diagnostic statistics on simulated draws. Effective sample size which informs the number of MCMC samples necessary to achieve a given level of precision for the approximation is also calculated to determine the extent of thinning. The **R** package **coda** was used for convergence diagnostics in this study.

3.3 *Model Comparison*

In the Bayesian framework, there are several approaches for model comparisons, such as Bayes factors and deviance information criterion (DIC). The former approach is computationally complex and sensitive to prior specifications. In this paper, we use a widely used criterion DIC for comparing models with different classes. DIC was introduced by Spiegelhalter et al. (2002) 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 $\tilde{\theta}$. $\overline{D(\theta)}$ is the posterior mean of the deviance and it offers summary information on how much discrepancy exists between the model and the data. In the frequentist framework, standard model comparison criteria such as Akaike information criterion (AIC) (Akaike, 1998) and Bayesian information criterion (BIC) (Schwarz et al., 1978) assume the number of parameters to be known, however, the number of parameters in hierarchical Bayesian models is not clear and can not be determined directly. 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, DIC assesses both a Bayesian measure of a model fit and the complexity of the model. Like AIC and BIC, a model with a smaller DIC is usually preferred.

Celeux et al. (2006) 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, $\theta^{(m)}$=($\mu^{(m)}$, $p^{(m)}$, and $\phi^{(m)}$). In particular,

$$\overline{D(\theta)} = -2\frac{1}{M}\sum_{m=1}^{M}\log\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}^{(m)}f(y_{ik}|\theta_{ik}^{(m)})$$

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

In the following real data application, we use both the traditional DIC and DIC_3 as criteria for model selection.

4. Simulation Study

To test the proposed model, we first conducted a small simulation study. In order to have a simulated dataset that is close to our real data, we fitted the smoking data from the Add Health study in **SAS Proc traj** (Jones et al., 2001) with 3 classes. We then used the parameter estimates from the SAS output to generate Y as a mixture of three zero-inflated Poisson distributions. The simulated dataset had a sample size of $n = 3000$, each with four repeated observations. The binomial and Poisson components contained class specific intercept (α_{k1} and β_{k1}), linear age (α_{k2} and β_{k2}), and quadratic age (α_{k3} and β_{k3}). Age at four time point was simulated as: $Age_0 \sim \mathcal{N}(15.6, 1.6)$; $Age_1 = Age_0 + \mathcal{N}(0.91, 0.14)$; and

$Age_2 = Age1 + \mathcal{N}(5.45, .22)$; and $Age_3 = Age_2 + \mathcal{N}(6.51, .30)$ to reflect baseline age and time intervals among these four occasions.

Two other covariates gender and ethnicity 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 a three-class model to the simulated data. Table 5 in the appendix presents the summary statistics for the model parameters. The zero-inflated latent class growth model was able to correctly identify 92.1% of the subjects' class membership. The class proportions for class 1 to 3 were 71%, 19%, and 10%, respectively. These were identical with the true class proportions. True values of all parameters were contained in their 95% highest posterior density (HPD) intervals.

5. Data Illustration

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. To examine possible risk factors for smoking patterns, we allow gender, ethnicity, 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, African, Hispanic, Native and other, peer smoking, and household smoking. Females, Caucasian, and no smokers in the household were set to be the reference groups. We ran a series of latent class models with the number of classes K ranging from two to six. Within each class, we fitted a ZIP model and a 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 (R Core Team, 2013) using the MCMC algorithm as described in Section 3. The R code was adapted from Dr. Brain Neelon's website: http://people.musc.edu/ brn200/r/. Non-informative priors were specified for each parameter. Specifically, we had $\mu_\alpha = \mu_\beta = \mu_\gamma = 0$ and $\sigma^2_\alpha = \sigma^2_\beta = \sigma^2_\gamma = 100$ for regression parameters $\{\alpha_k, \beta_k, \gamma_k\}$ and $a = 0.001, b = 0.001$ for dispersion parameter ϕ_k. For ZIP latent class models, we ran 480,000 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 the 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.

Both DIC and DIC_3 were used for model comparison. Table 2 presents DIC statistics for each of the five fitted ZIP latent class models and each of the three fitted ZINB latent class models. For ZIP models, the DIC values changed the most from $K = 2$ to $K = 3$ but there was little improvement after $K = 5$. Thus, the ZIP model with five classes was preferred. As for ZINB models, models with more than three classes were poorly identified and failed to converge. For the 3-class ZINB model, the second class comprised only 2.98% of the population and the dispersion parameter for this class (ϕ_2) dropped to zero, which implies that dispersion probably does not exist in this class and a zero-inflated Poisson distribution is more preferred. There was also little improvement from $K = 2$ to $K = 3$ as for DIC values. Therefore, the ZINB model with two classes was preferred. We discuss more about the identifiability of ZINB latent class models in Section 6.

Figures 1-6 from the Web appendix present trace plots and autocorrelation function plots for αs, βs, and γs. All plots show that the chain has a relatively good mixing and has

converged to its stationary distribution. It is worth mention 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 consequences of label switching, the class membership probabilities will be $1/K$ for every subject and the posterior distribution of the parameters will be highly symmetric and multimodal (Stephens, 2000). Thus, label switching results in misleading parameter estimates. Several online or post-hoc algorithms have been developed to relabel the latent classes (Sperrin et al., 2010; Stephens, 2000). We carefully examined the MCMC output, however, we found no evidence of label switching in our models. It is possible that the inclusion of class membership covariates helped with the identifiability of the classification. As shown in Figure 1, values for parameters from five classes generated after burn-in were quite separated and there was no sign of values jumping from one class to another.

Posterior means and 95% credible intervals of αs and βs for the five-class ZIP model are presented in Table 3. In figure 2, each color represents the overall average smoking level trajectory for each of the five classes. Figure 2 and 3 present posterior trajectories for probability of being a non-smoker (structural zero) and posterior trajectories of average number of cigarettes smoked given that they smoke (i.e. from a count process). These five 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 42.15% of the subjects and the trajectory pattern was characterized by a low initial probability of being a non-smoker and a slowly increasing trend until late 20s. We labelled subjects from the first class as "late smokers". Class 2 included 14.95% of the subjects and was characterized by a low initial probability of being a smoker, a low level of smoking until late 20s, and a rapid increasing

trend after 30. Subjects from class 2 were termed as "late escalators". Class 3 had the smallest proportion of subjects (7.73%) and we called this group "early light smokers" as it was characterized by a relatively high initial level of smoking and then a decreasing trend on both the probability of being a non-smoker and the level of smoking. Only 10.30% of the subjects were in class 4, which we described it as "early heavy smokers". This class had a relatively stable probability of being a non-smoker from adolescence to adulthood (as shown in Table 3, both linear age and quadratic age effects were not significant). For those who were smokers, they started smoking at a younger age and smoked more and more until 20s. Most of them quitted smoking after they reached their 20s. Class 5 comprised 24.87% of the subjects and it also had a stable probability of being a non-smoker (around 0.6). Subjects from this class were also characterized by having a turning point around middle 20s, i.e., there was an increasing trend before 25 and a decreasing trend after 25. Since subjects from this class also had a high average level of smoking, we labelled them as "late heavy smokers".

Posterior means and 95% credible intervals of αs and βs for the two-class ZINB model are presented in Table 4. Class 1 and class 2 both comprised about half of the subjects. As shown in Figure 6, compared with class 2, class 1 had an opposite trend as for probability of being a non-smoker over age. These two classes also differed by initiation time of smoking and time when peak level of smoking happened. Class 1 could be referred to as "early smokers" and class 2 could be referred to as "late smokers".

While examining risk factors' influence on class membership probabilities, we found that compared with the first class, there were less Asians and Hispanics in the fourth class, less males in the fifth class, and less peer smoking in the second and the fourth class, with more than 90% of the probability. Because the first class was served as the reference class in the model, comparisons were only made with the first class. We created a heatmap plot in order to compare the relative proportion of different gender, race, whether or not one had smokers

as best friends, and whether or not one had smokers in the household among each of the five classes. In Figure 8, the number on each cell represents the ratio of percentage in a certain class to average percentage in all five classes. A ratio smaller than 1 indicates a smaller probability of being in that class compared with the average probability and a ratio bigger than 1 suggests a bigger probability of being in that class. For instance, the ratio for males in class 3 was 1.26, which means compared with its average percentage, males were 1.26 times more likely to be in the third class.

6. Discussion

In the present paper, we described a latent class model for analyzing longitudinal count data that exhibit excess zeros. The modeling approach has several advantages. Frist, because the latent class variable (i.e., C_i) can effectively summarize distinctive patterns of change in longitudinal data and the latent binary variable (i.e., B_{it}) can distinguish whether it comes from a zero-inflation process or a regular count process for a certain observation in each time point. This model is very flexible for modeling both unobserved time stable and time varying heterogeneity. Second, it also 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 superior to the traditional two-stage approach which does not take into account of the uncertainty of the class membership.

We demonstrated the method in modeling developmental trajectories of cigarette smoking behavior from early adolescence to adulthood. By fitting a ZIP latent class model, we were able to identify five distinct groups of trajectories: late smokers, late escalators, early light smokers, early heavy smokers, and late heavy smokers. Two types of smokers were identified by fitting a ZINB latent class model. Different 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 on smoking. Compared with ZIP models, ZINB models can account for more variability in the data by having class-specific dispersion parameters and thus less number of classes were needed. Though the latter offer a better model fit, solutions from the ZINB models seem to be over-simplified for modeling smoking patterns, thus resulting in a less meaningful interpretation in this application. In addition, ZINB models with more than three classes had identifiability issues. This is probably due to the fact that more classes were not necessary. It is also likely that the model is too complicated in the sense that three types of mixture information exist in a ZINB latent class model. First, a negative binomial model is a mixture of Poisson and gamma distributions to account for over-dispersion when a Poisson model is not appropriate; then an additional mixture is added by allowing zero-inflation in the model; and a latent categorical variable is then added to account for possible typological variability in the change process. The flexibility of this modeling approach is very appealing, however, such a complicated model might overfit the data and offer less meaningful interpretation. As a result, we would prefer the five-class ZIP model for modeling the smoking patterns.

One feature of fitting zero-inflated latent class models is that the probability of being a smoker or non-smoker and level of smoking are estimated separately. This allows a separate examination of these two parameters of interest. However, the side effect of modeling the data in this way is that, each class comprised with both smokers and non-smokers. The model did not identify those who had always been non-smokers as a separate group. To overcome this problem, the model can be extended to have class-specific correlated random effects, thus allowing the zero-inflation component and the count component to be related at different levels across classes. Neelon et al. (2011) compared several latent class models with fixed effects, uncorrelated random effects, or correlated random effects and found that

the model with correlated random effects had the best fit and it requires fewer classes to capture the variability in the data.

The other limitation of our application was that the data were collected using a cohort sequential design. The baseline age ranged from 13-21 years and each subject only had four measurements with different time intervals. Though there was overlapping in ages 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.

Despite limitations of the models on this specific data application, 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 behavior and implementing interventions targeting on multiple health risk behaviors due to the fact that multiple unhealthy behaviors often co-occur. Prochaska et al. (2008) suggested that however, there exists surprisingly little understanding of the basic principles of multiple health behavior change. As for application interest, the model can also be extended to accommodate multiple outcomes, such as dual trajectory models linking the trajectory groups of two behaviors(Jones et al., 2001).

Supplementary Materials

Web Appendix A, referenced in Section 5, is available with this paper at the Biometrics website on Wiley Online Library.

Akaike, H. (1998). A bayesian analysis of the minimum aic procedure. In *Selected Papers of Hirotugu Akaike*, pages 275–280. Springer.

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,** 334–340.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1,** 651–673.

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,** 154–163.

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,** 127–135.

Dagne, G. A. (2004). Hierarchical bayesian analysis of correlated zero-inflated count data. *Biometrical Journal* **46,** 653–663.

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.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons.

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* pages 1–17.

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,** 1360–1375.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press.

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

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,** 374–393.

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,** 192–200.

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

Klein, N., Kneib, T., and Lang, S. (2014). Bayesian generalised additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association* pages 00–00.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14.

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,** 685.

McArdle, J. J. and Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development* pages 110–133.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal*

of Econometrics **33,** 341–365.

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.

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,** 882–891.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics* **55,** 463–469.

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

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,** 280–289.

Neelon, B. H., OMalley, 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,** 421–439.

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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,** 199–209.

Rindskopf, D. and Rindskopf, W. (1986). The value of latent class analysis in medical

diagnosis. *Statistics in Medicine* **5,** 21–27.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing* **20,** 357–366.

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,** 583–639.

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

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

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

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,** 167–178.

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.

Appendix Summary statistics for the simulation study

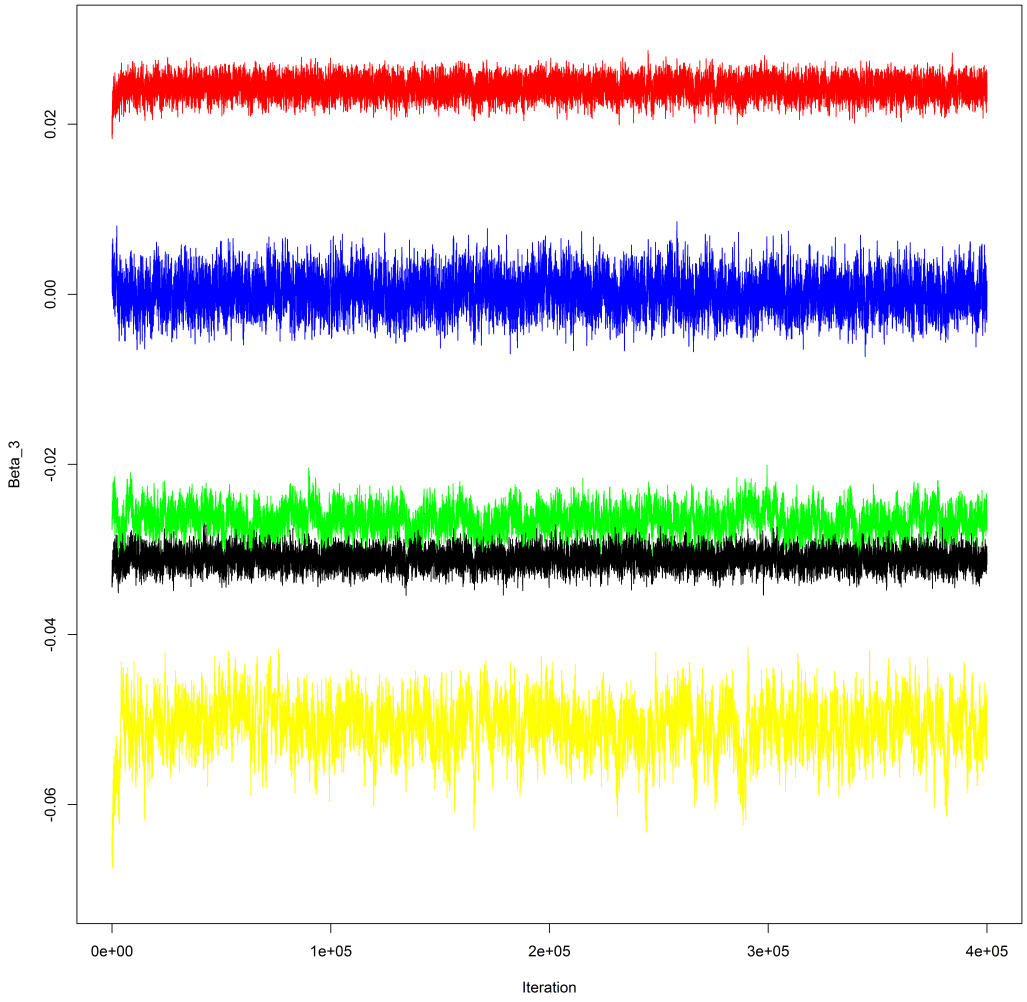

Figure 1: Post burn-in trace plots for β_{13} to β_{53}. In the figure, one color represents one β_{k3} from each of the five classes.

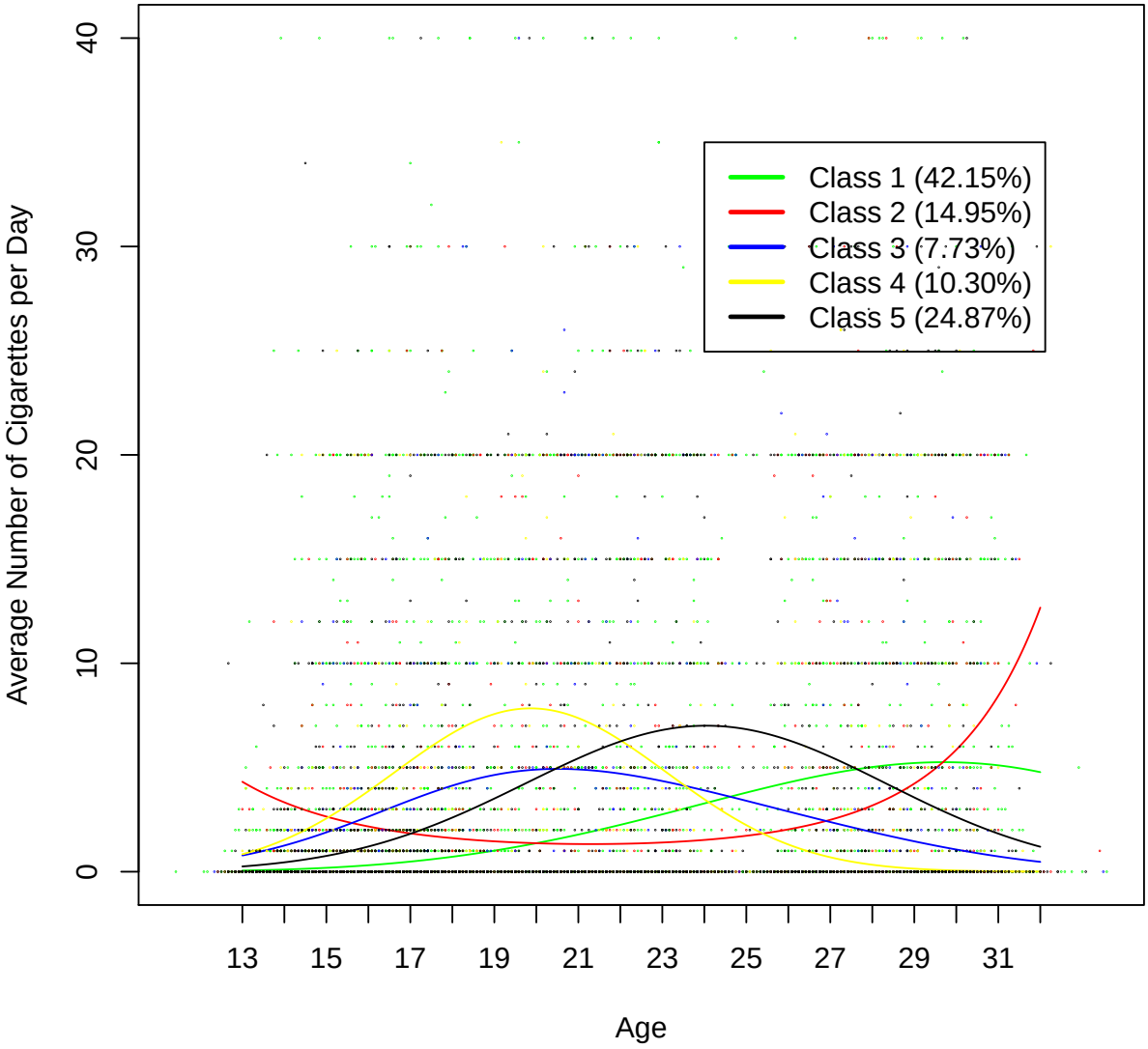

Figure 2: Overall posterior mean smoking trajectories for the five-class ZIP model

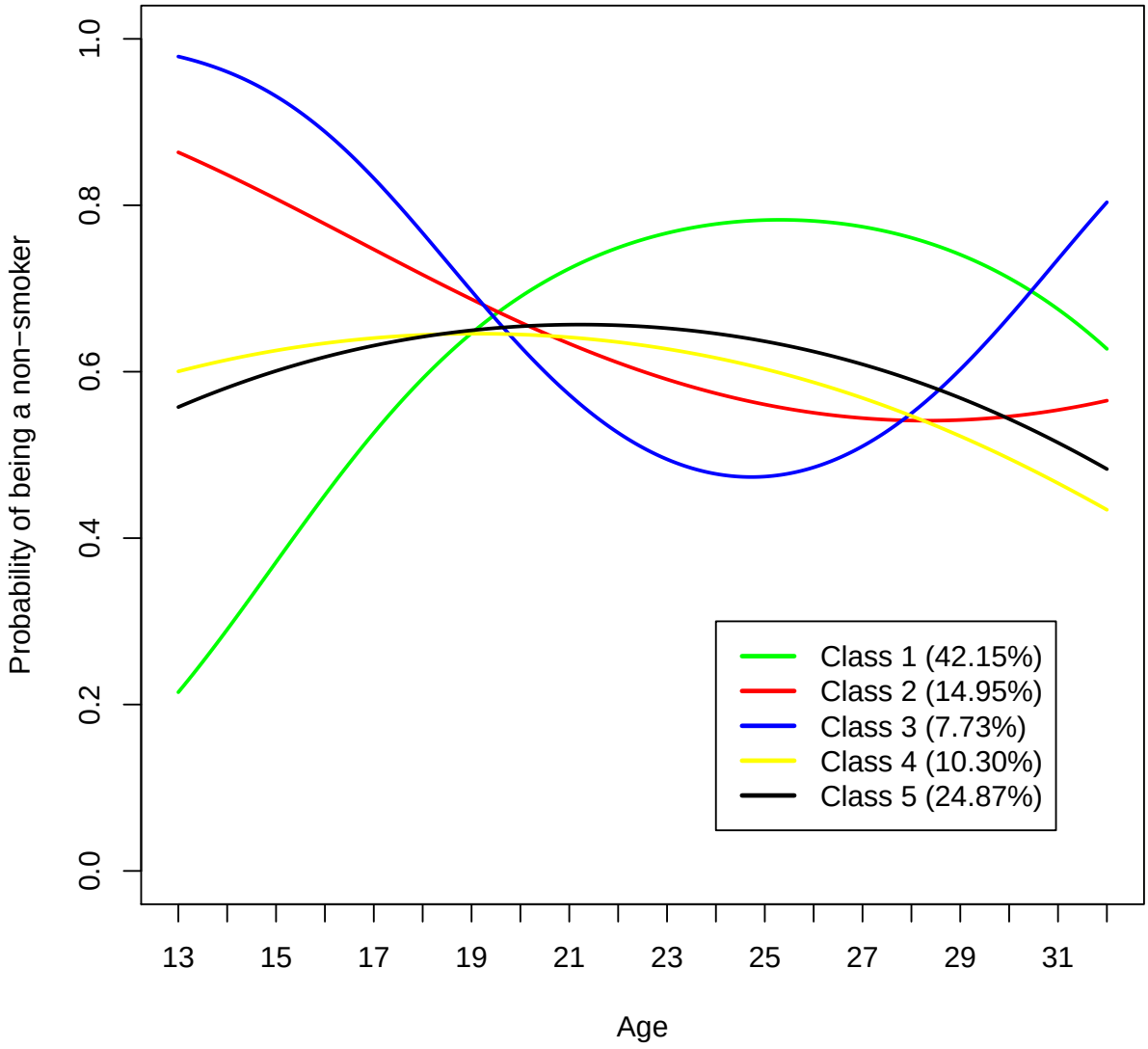

Figure 3: Posterior trajectories for probability of being a non-smoker for the five-class ZIP model

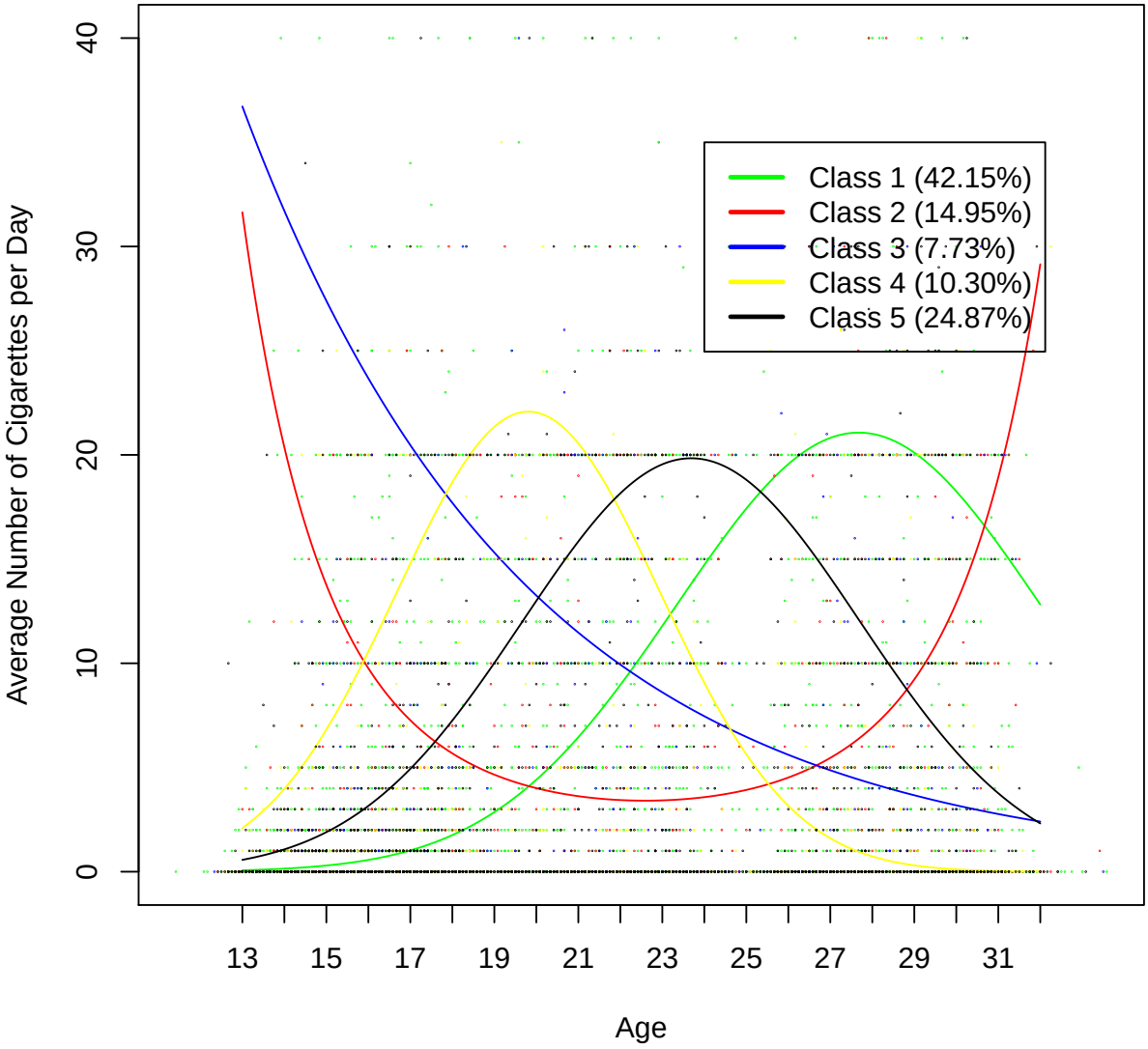

Figure 4: Posterior mean smoking trajectories for smokers for the five-class ZIP model

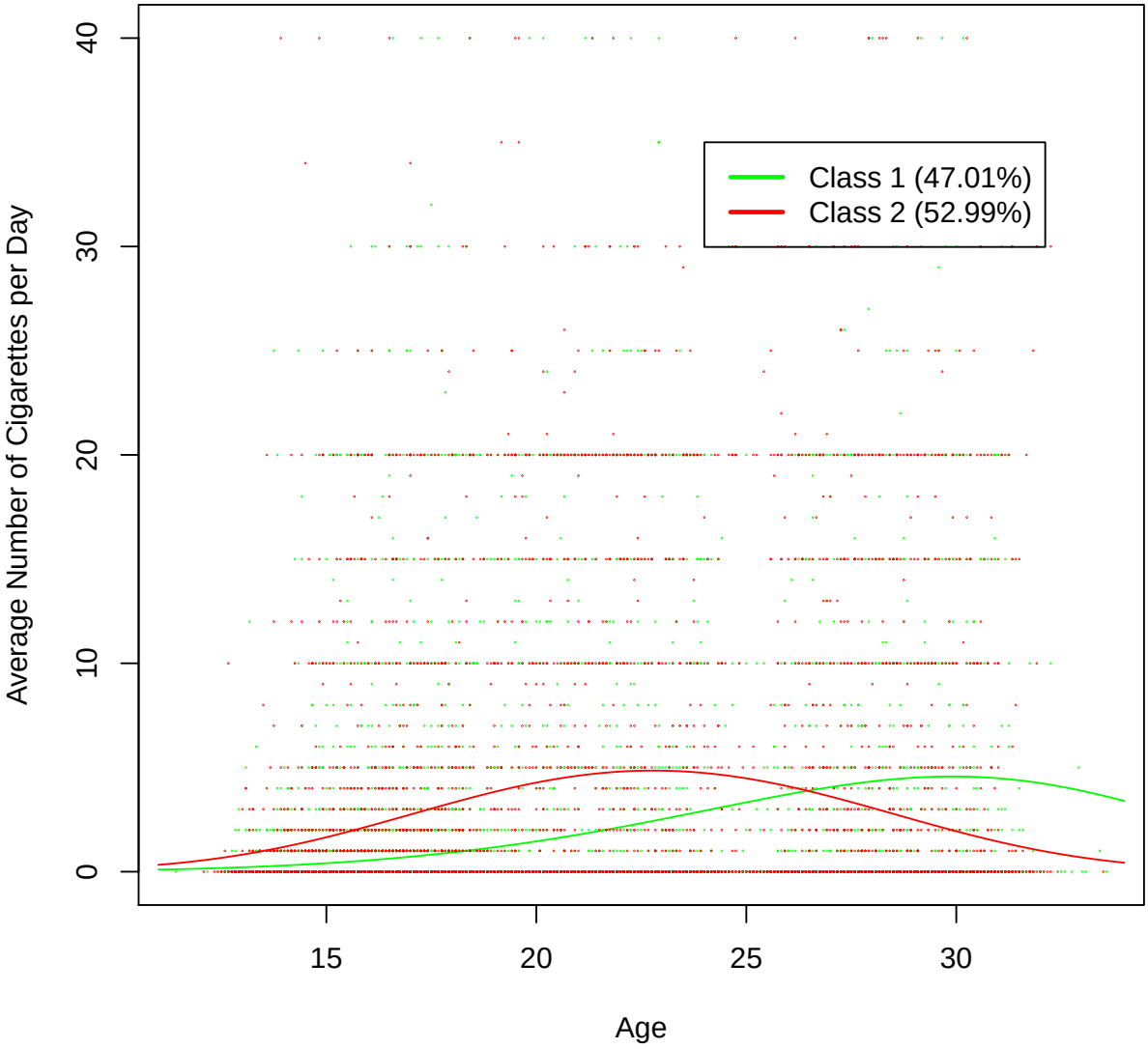

Figure 5: Overall posterior mean smoking trajectories for the two-class ZINB model

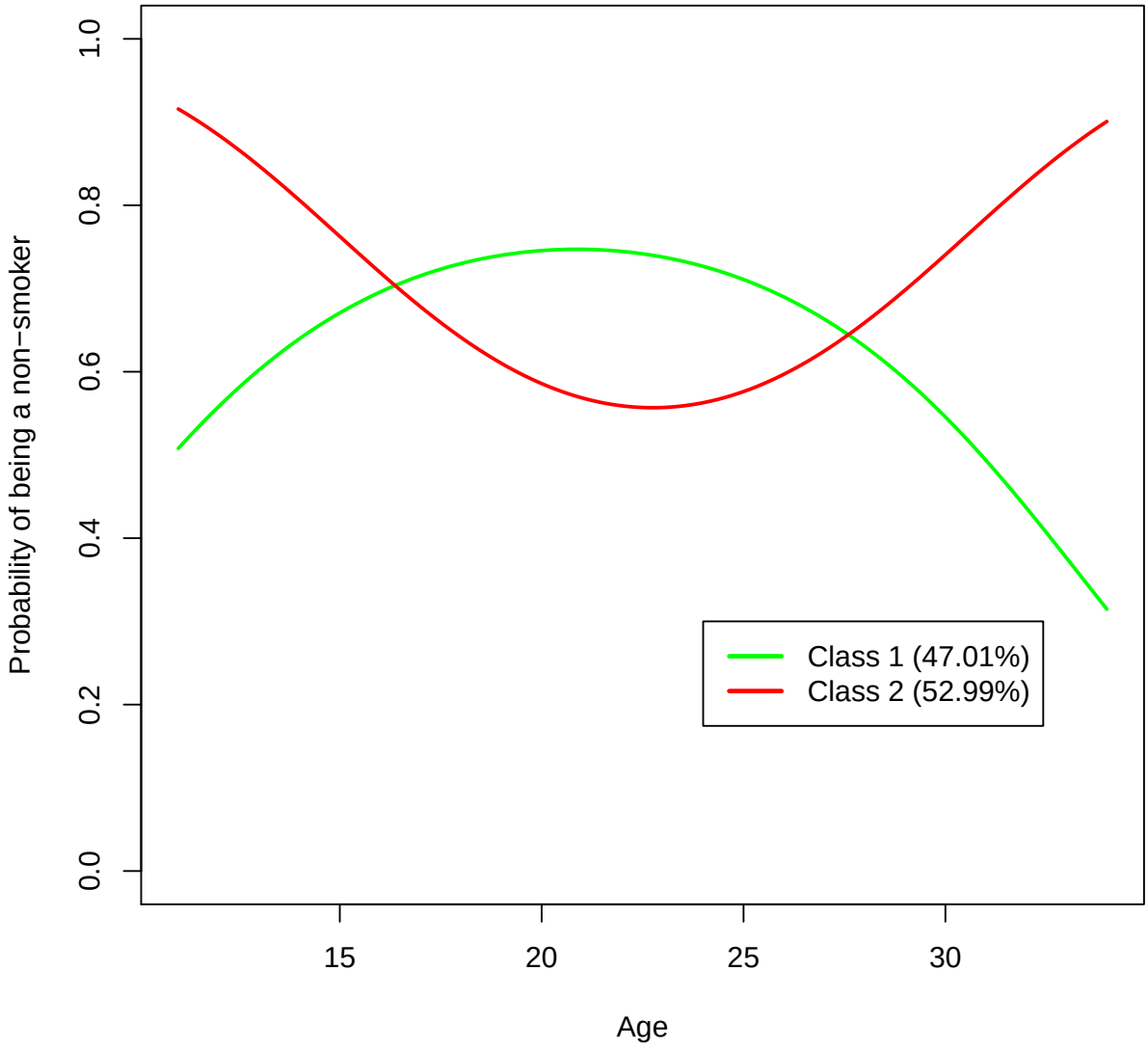

Figure 6: Posterior trajectories for probability of being a non-smoker for the two-class ZINB model

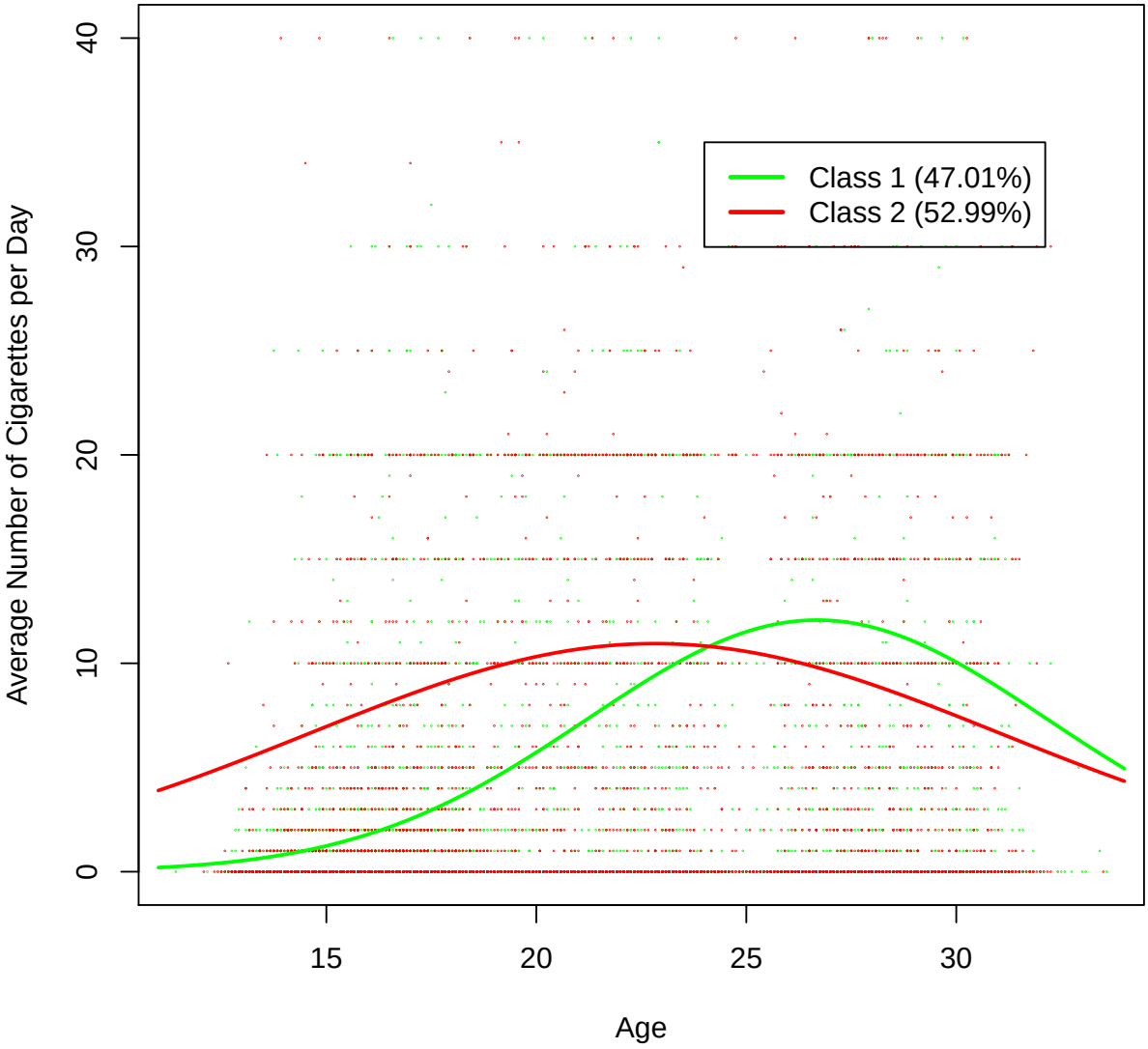

Figure 7: Posterior mean smoking trajectories for smokers for the two-class ZINB model

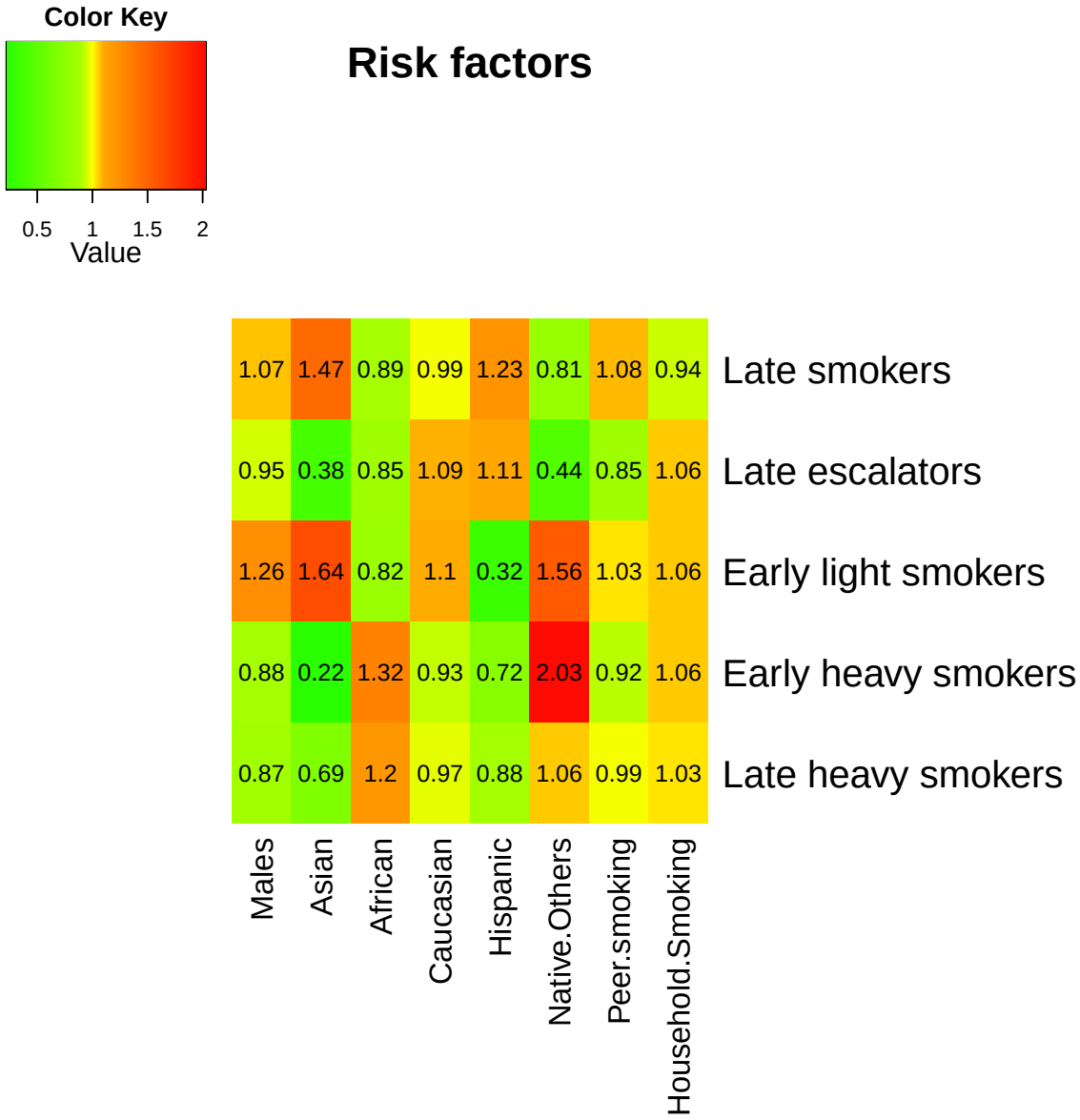

Figure 8: Risk factors on class membership. Green color represents smaller ratios (i.e., < 0.9) and red color represents bigger ratios (i.e., > 1.1), whereas yellow represents ratios between 0.9 to 1.1.

Table 1: Descriptive statistics for the smoking study (n=2923)

Year	Age (years) Mean(S.D.)	Cigarettes Mean(S.D.)	Observed zeros (%)
1994-1995	15.50(1.56)	1.42(4.26)	77.45
1996	16.41(1.57)	1.99(5.22)	68.94
2001-2002	21.86(1.58)	3.75(7.56)	66.75
2007-2008	28.36(1.59)	3.65(7.24)	64.15

Table 2: DIC statistics for ZIP and ZINB latent class models

Model	Number of classes	$\overline{D(\theta)}$	p_D	DIC	p_{D3}	DIC_3
ZIP	2	42719.08	19.80	42738.89	35.27	42754.35
	3	41577.21	34.21	41611.42	41.08	41618.29
	4	40203.67	47.88	40251.55	48.14	40251.80
	5	39733.77	62.11	39795.88	49.96	39783.73
	6	39555.96	67.26	39623.23	64.64	39620.60
ZINB	2	35975.61	21.37	35996.97	65.28	36040.89
	3	35888.70	32.95	35921.65	69.75	35958.45

Table 3: Posterior means and 95% credible intervals for the five-class ZIP latent class model

Class (%)	Model Component	Parameter (Covariate)	Posterior Mean	95% Credible Interval
1(42.15%)	Binomial	α_{11} (Intercept)	−9·607	(-13.915, -4.993)
		α_{12}(Linear Age)	0·860	(0.465, 1.231)
		α_{13}(Quadratic Age)	−0·017	(-0.025, -0.009)
	Poisson	β_{11} (Intercept)	−17·283	(-19.170, -15.525)
		β_{12}(Linear Age)	1·469	(1.327, 1.622)
		β_{13}(Quadratic Age)	−0·027	(-0.030, -0.024)
2(14.95%)	Binomial	α_{21} (Intercept)	5·912	(0.788, 8.185)
		α_{22}(Linear Age)	−0·406	(-0.615, 0.079)
		α_{23}(Quadratic Age)	0·007	(-0.004, 0.012)
	Poisson	β_{21} (Intercept)	13·592	(7.071, 14.530)
		β_{22}(Linear Age)	−1·095	(-1.182, -0.624)
		β_{23}(Quadratic Age)	0·024	(0.016, 0.026)
3(7.73%)	Binomial	α_{31} (Intercept)	17·405	(8.682, 21.548)
		α_{32}(Linear Age)	−1·416	(-1.798, -0.727)
		α_{33}(Quadratic Age)	0·029	(0.015, 0.037)
	Poisson	β_{31} (Intercept)	5·542	(3.784, 7.397)
		β_{32}(Linear Age)	−0·151	(-0.371, 0.014)
		β_{33}(Quadratic Age)	0·000	(-0.004, 0.006)
4(10.30%)	Binomial	α_{41} (Intercept)	−1·295	(-5.281, 4.336)
		α_{42}(Linear Age)	0·198	(-0.325, 0.604)
		α_{43}(Quadratic Age)	−0·005	(-0.015, 0.007)
	Poisson	β_{41} (Intercept)	−16·818	(-19.573, -14.903)
		β_{42}(Linear Age)	2·010	(1.810, 2.339)
		β_{43}(Quadratic Age)	−0·051	(-0.060, -0.046)
5(24.87%)	Binomial	α_{51} (Intercept)	−2·130	(-4.350, 0.495)
		α_{52}(Linear Age)	0·262	(0.014, 0.473)
		α_{53}(Quadratic Age)	−0·006	(-0.011, -0.001)
	Poisson	β_{51} (Intercept)	−14·458	(-15.491, -13.458)
		β_{52}(Linear Age)	1·473	(1.386, 1.566)
		β_{53}(Quadratic Age)	−0·031	(-0.033, -0.029)

Table 4: Posterior means and 95% credible intervals for the two-class ZINB latent class model

Class (%)	Model Component	Parameter (Covariate)	Posterior Mean	95% Credible Interval
1(46.90%)	Binomial	α_{11} (Intercept)	−3·576	(-6.424, -0.954)
		α_{12}(Linear Age)	0·447	(0.206, 0.709)
		α_{13}(Quadratic Age)	−0·011	(-0.016, -0.005)
	Negative Binomial	β_{11} (Intercept)	−9·332	(-11.209, -7.593)
		β_{12}(Linear Age)	0·886	(0.722, 1.062)
		β_{13}(Quadratic Age)	−0·017	(-0.020, -0.013)
2(53.10%)	Binomial	α_{21} (Intercept)	8·352	(6.691,10.052)
		α_{22}(Linear Age)	−0·716	(-0.885, -0.551)
		α_{23}(Quadratic Age)	0·016	(0.012, 0.020)
	Negative Binomial	β_{21} (Intercept)	−1·470	(-2.980, 0.058)
		β_{22}(Linear Age)	0·339	(0.193, 0.486)
		β_{23}(Quadratic Age)	−0·007	(-0.011, -0.004)

Table 5: Summary statistics for the simulation study

Class (%)	Model Component	Parameter (Covariate)	True Mean	Posterior Mean	95% HPD Interval
1(71%)	Binomial	α_{11} (Intercept)	−·008	·407	(-1.159, 2.023)
		α_{12}(Linear Age)	·140	·114	(-0.030, 0.265)
		α_{13}(Quadratic Age)	−·003	−·002	(-0.006, 0.001)
	Poisson	β_{11} (Intercept)	−5·373	−5·250	(-6.288, -4.168)
		β_{12}(Linear Age)	·512	·494	(0.397, 0.588)
		β_{13}(Quadratic Age)	−·010	−·010	(-0.012, -0.007)
2(19%)	Binomial	α_{21} (Intercept)	8·416	9·480	(6.171, 12.726)
		α_{22}(Linear Age)	−·810	−·923	(-1.246, -0.619)
		α_{23}(Quadratic Age)	·016	·019	(0.012, 0.026)
	Poisson	β_{21} (Intercept)	−5·268	−5·402	(-6.292, -4.407)
		β_{22}(Linear Age)	·512	0·526	(0.437, 0.601)
		β_{23}(Quadratic Age)	−·010	−·010	(-0.012, -0.008)
3(10%)	Binomial	α_{31} (Intercept)	11·012	12·595	(9.804, 15.157)
		α_{32}(Linear Age)	−1·052	−1·204	(-1.469, -0.946)
		α_{33}(Quadratic Age)	·021	·024	(0.019, 0.031)
	Poisson	β_{31} (Intercept)	1·059	·994	(0.676, 1.301)
		β_{32}(Linear Age)	·158	·164	(0.134, 0.192)
		β_{33}(Quadratic Age)	−·003	−·003	(-0.004, -0.002)
	Class Membership	γ_{21}	−1·134	−1·013	(-1.228, -0.800)
		γ_{22}	·327	·224	(0.006, 0.459)
		γ_{23}	−·387	−·621	(-1.373, 0.038)
		γ_{24}	−1·252	−1·152	(-1.504, -0.748)
		γ_{25}	−1·020	−·979	(-1.432, -0.535)
		γ_{26}	·084	−·297	(-0.931, 0.345)
		γ_{31}	−1·739	−1·825	(-2.015, -1.638)
		γ_{32}	·518	·569	(0.345, 0.809)
		γ_{33}	−2·453	−3·246	(-5.699, -1.287)
		γ_{34}	−1·870	−2·004	(-2.545, -1.512)
		γ_{35}	−1·032	−1·197	(-1.702, -0.700)
		γ_{36}	·255	·218	(-0.325, 0.793)

LIST OF REFERENCES

APPENDIX A

Convergence Diagnosis Plots

This appendix presents trace plots and autocorrelation function plots for MCMC convergence diagnosis. Trace plots of MCMC samples versus the simulation index is useful in diagnosis of convergence and mixing of a chain. All trace plots show relatively constant means and variances which indicate the chain has converged to its target distribution (i.e., the posterior distribution). Lag-t autocorrelation function estimates the correlation between elements of the sequence that are t steps apart (Hoff, 2009). Thus, plotting the autocorrelation functions tell how much correlation exist in the MCMC samples. High correlation indicates poor mixing of the chain and a larger sample size is required to achieve a given level of precision of the approximation. Figures A2, A4, A6, and A8 show that most of the autocorrelations are close to zero and within 0.1 after lag-10.

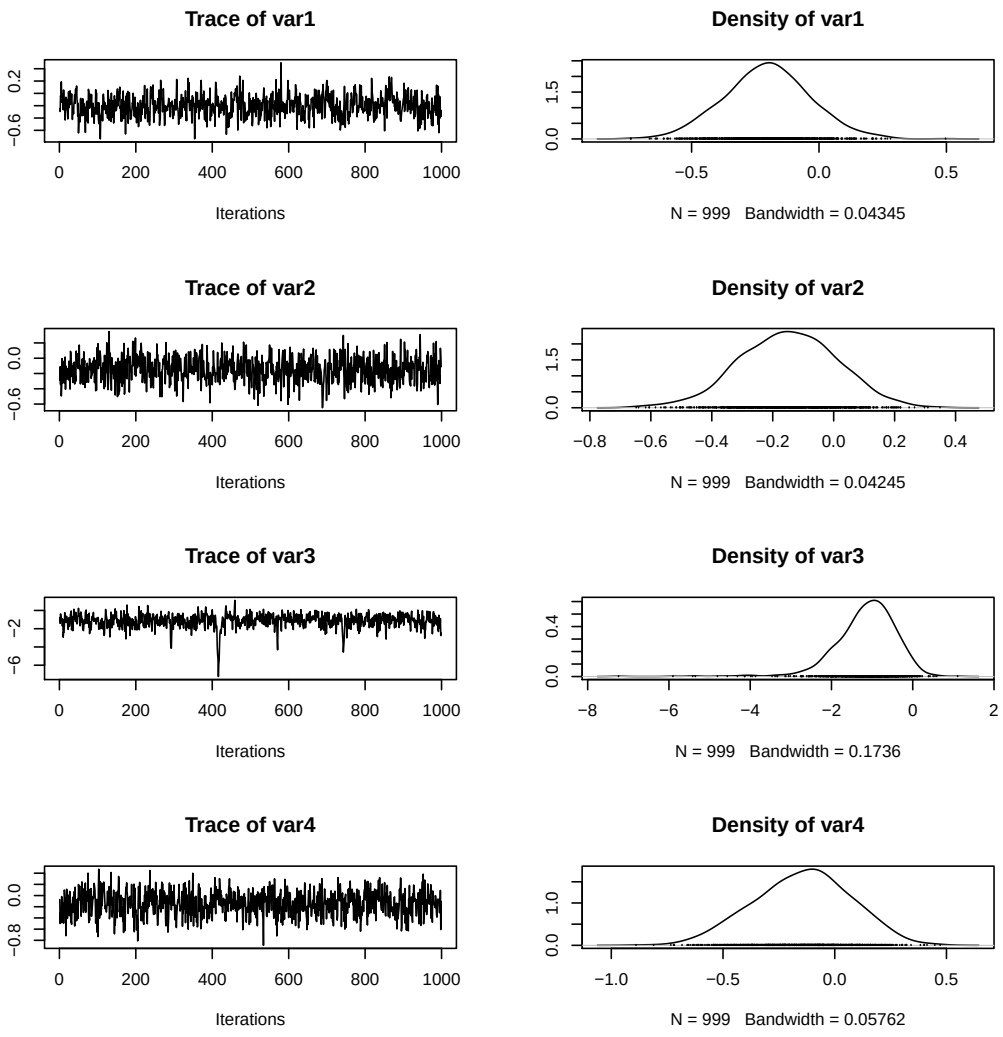

Figure A.1: Trace plot for γs for the five-class ZIP model

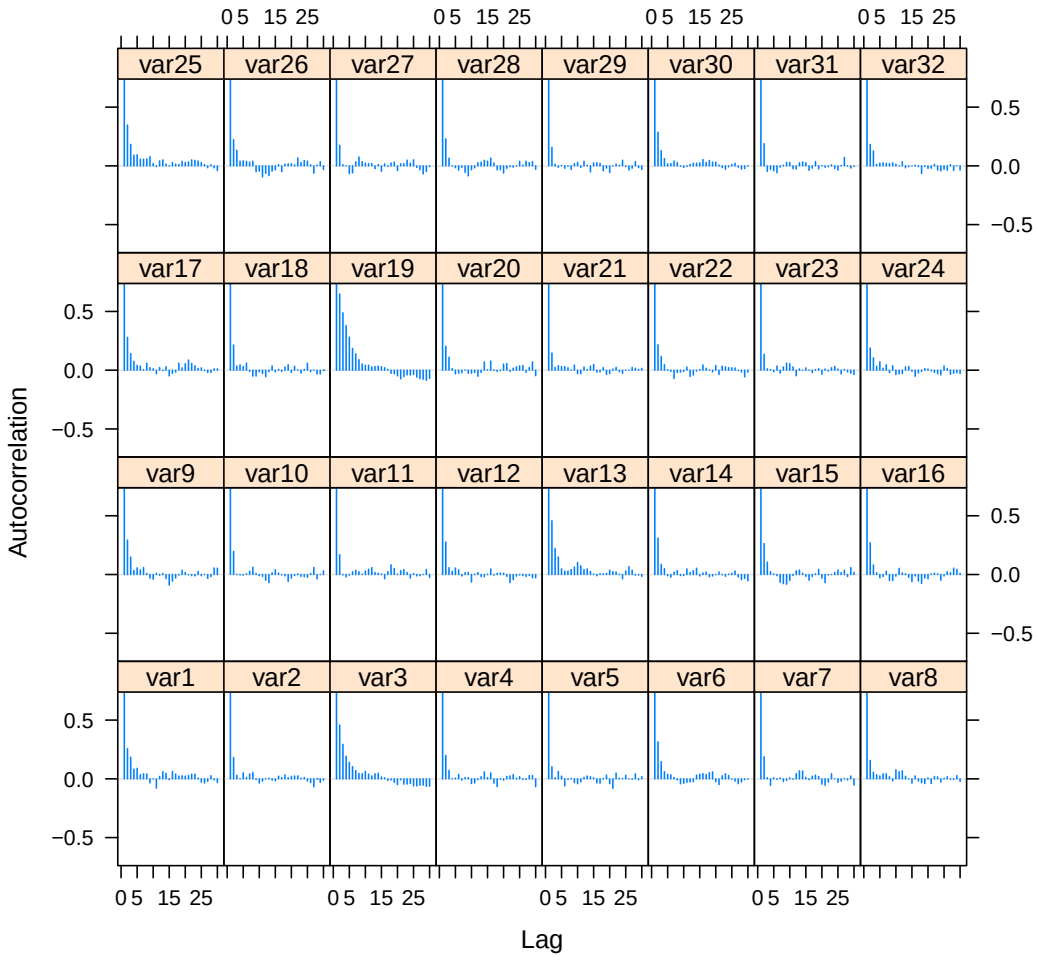

Figure A.2: Autocorrelation function plot for γs for the five-class ZIP model

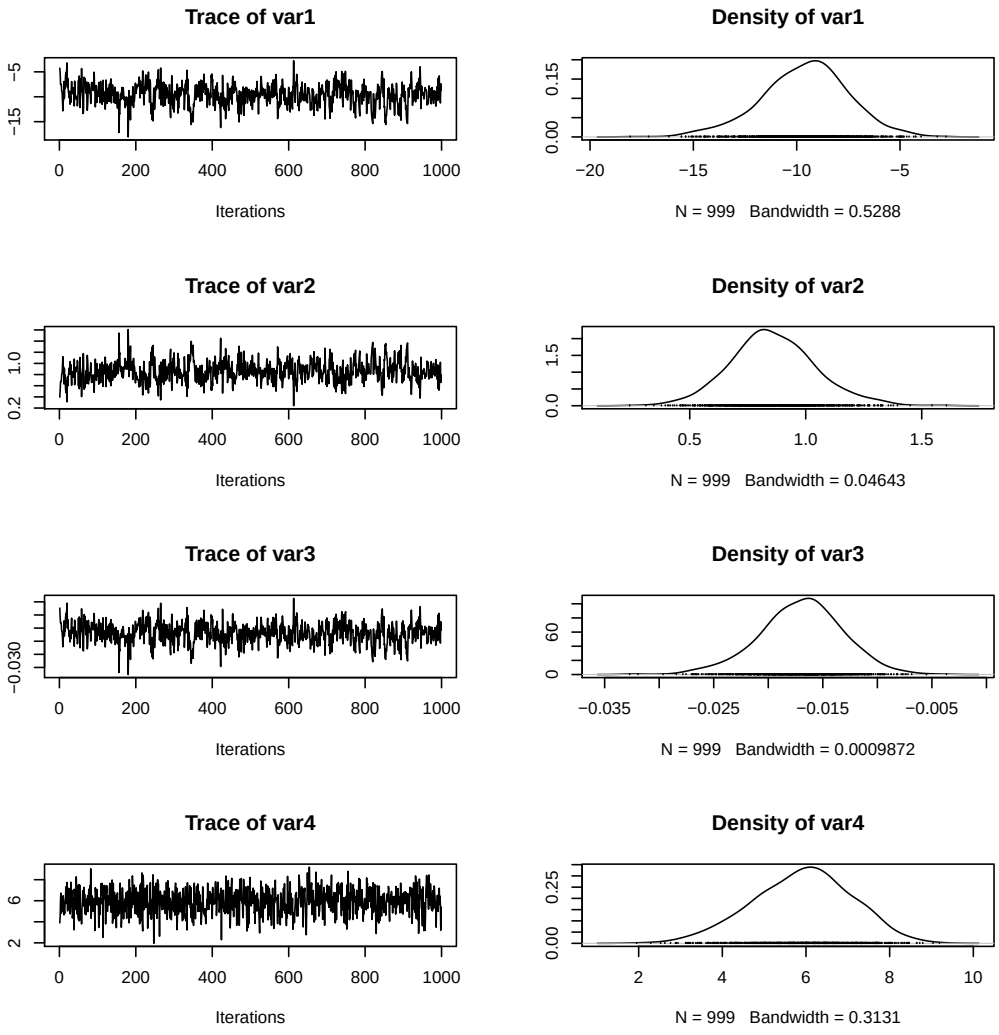

Figure A.3: Trace plot for αs for the five-class ZIP model

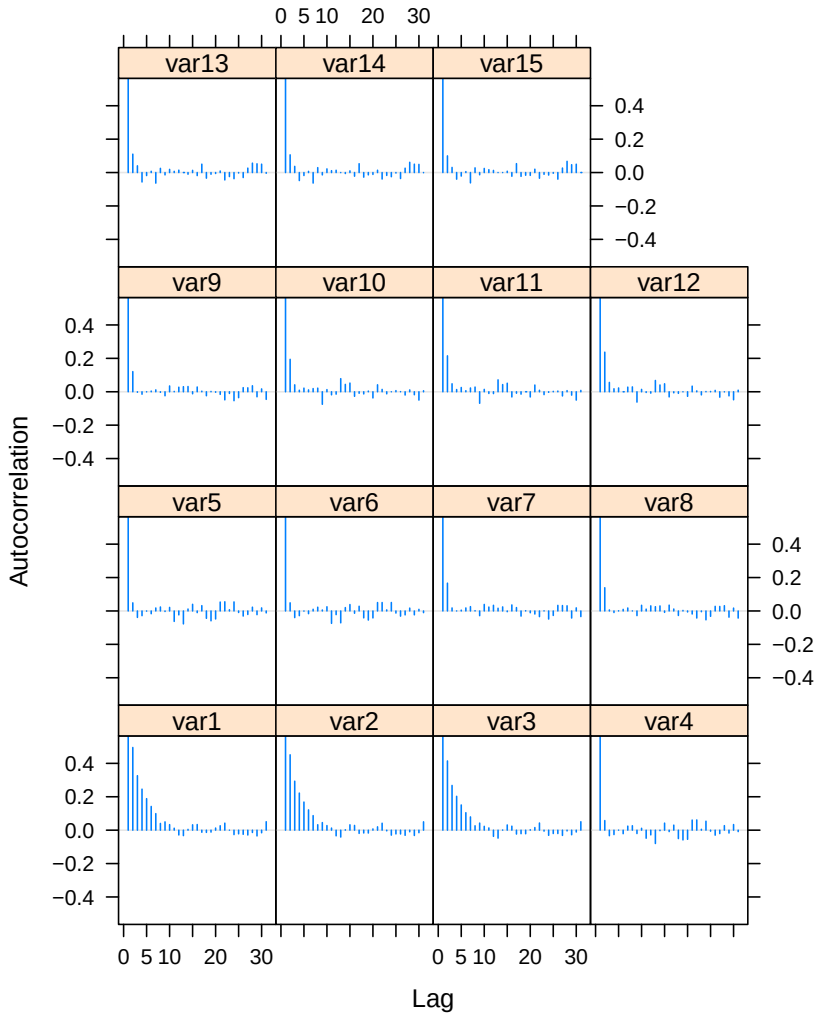

Figure A.4: Autocorrelation function plot for αs for the five-class ZIP model

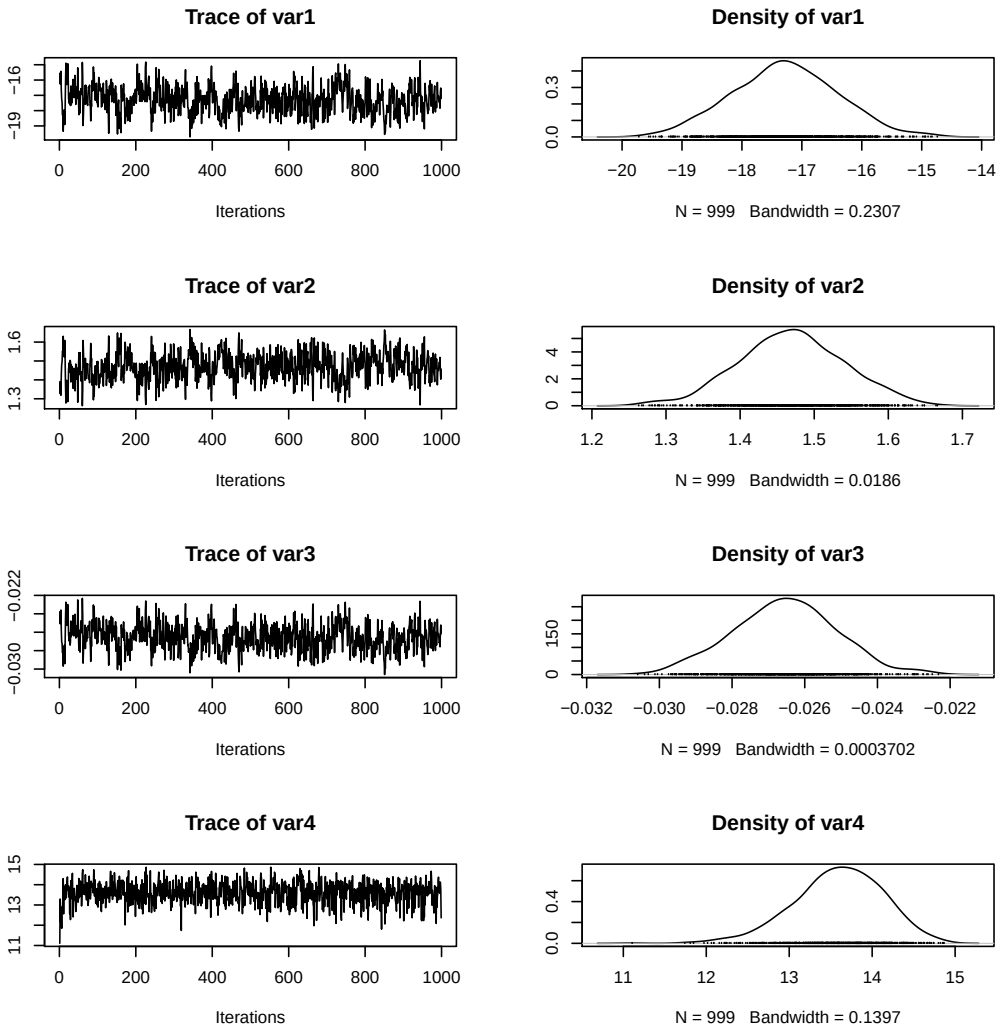

Figure A.5: Trace plot for βs for the five-class ZIP model

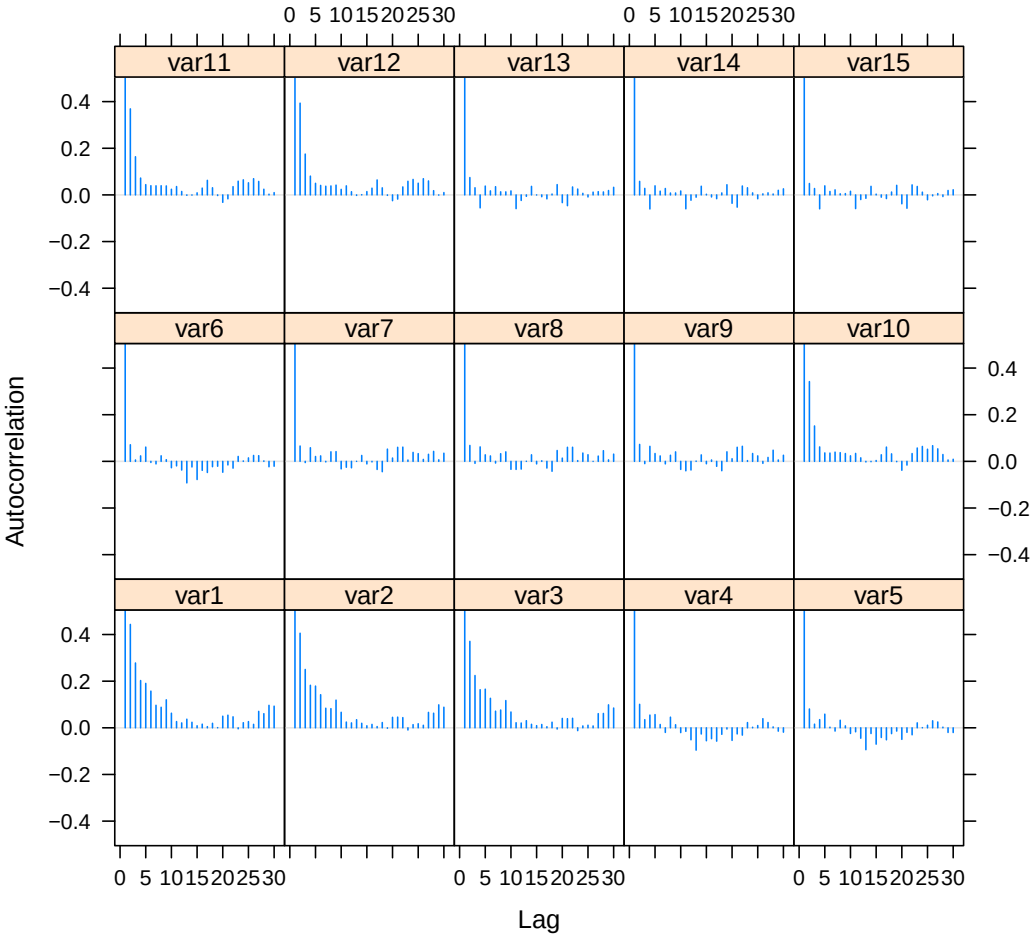

Figure A.6: Autocorrelation function plot for βs for the five-class ZIP model

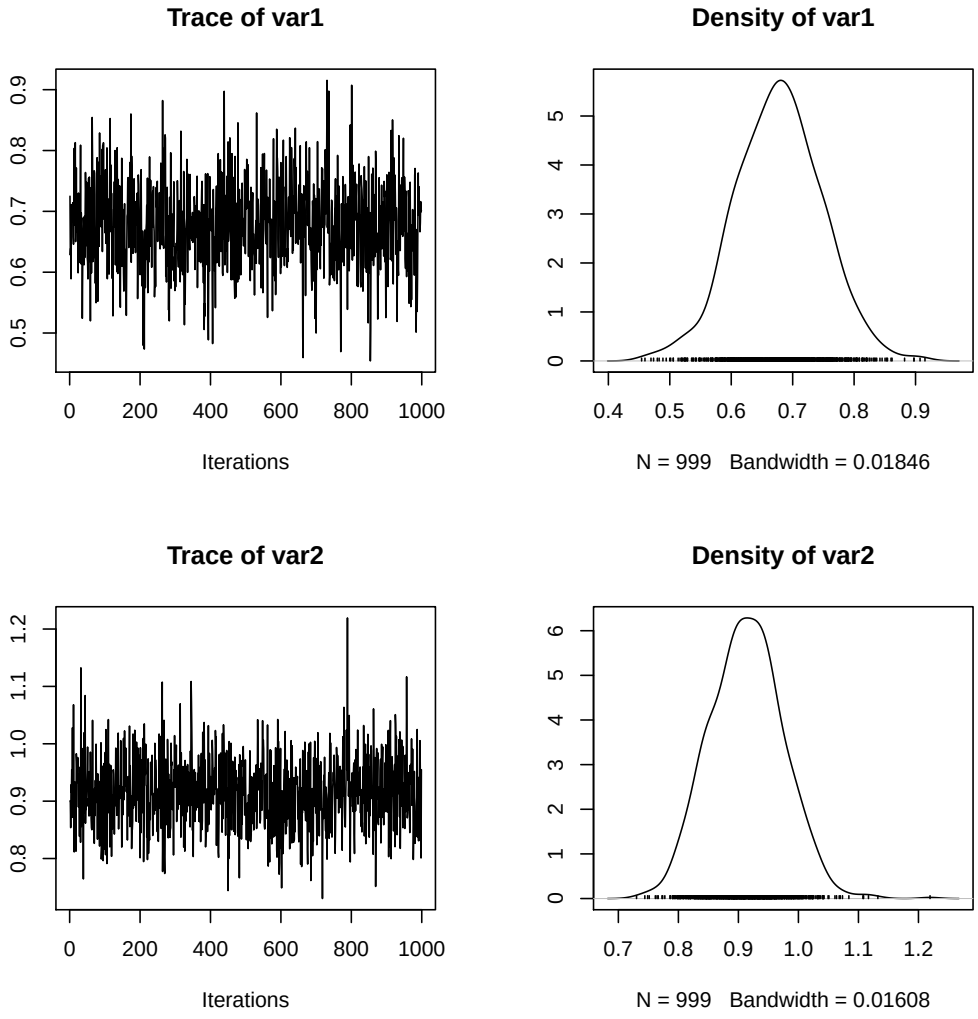

Figure A.7: Trace plot for ϕs for the two-class ZINB model

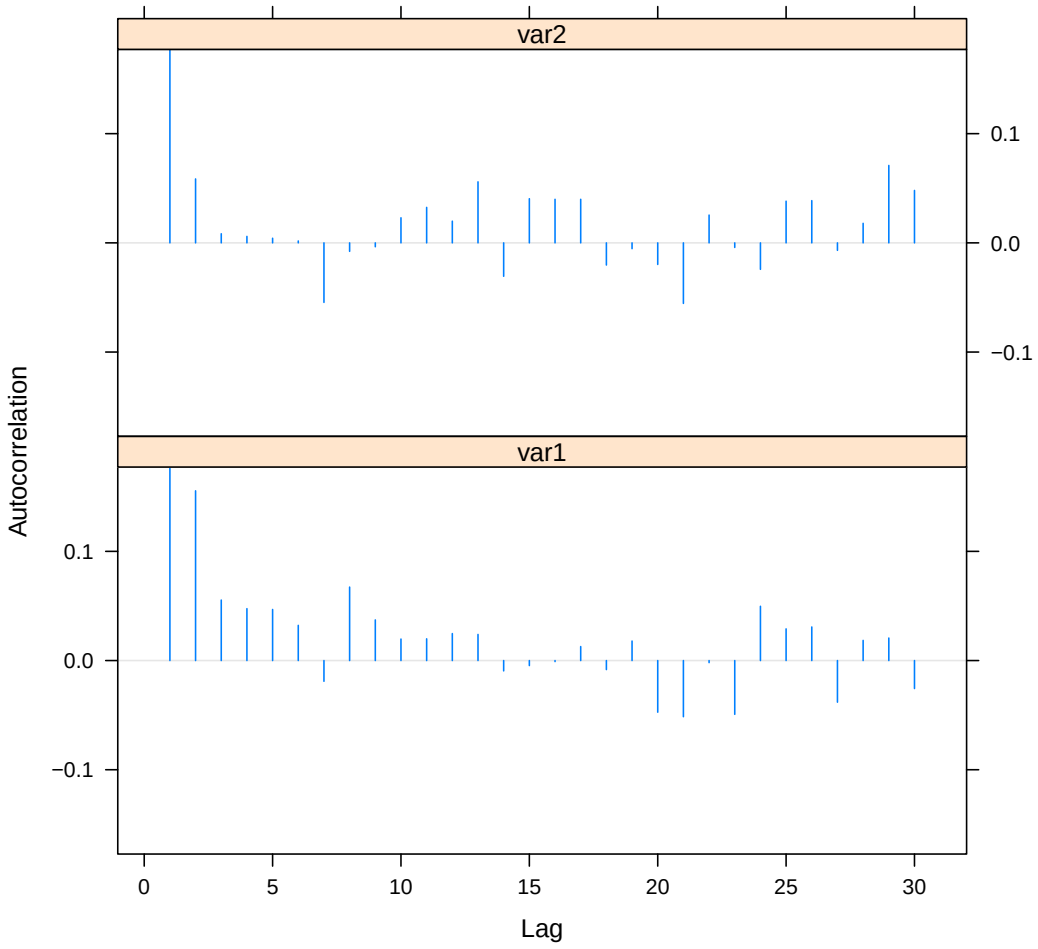

Figure A.8: Autocorrelation function plot for ϕs for the two-class ZINB model

APPENDIX B

Model Specification for Latent Class Zero-inflated Poisson Model

This appendix describes distributional forms of the random quantities in the proposed model. It also presents derivations of getting the complete data likelihood and full conditional distributions for parameters of interest in detail.

1. **Random quantities**

$$Y = (Y_1, \ldots, Y_N), Y_i = (Y_{i1}, \ldots, Y_{iT}), Y_{it}|C_i = k \sim \text{ZIP}(\mu_{itk}, p_{itk});$$

$$B = (B_1, \ldots, B_N), B_i = (B_{i1}, \ldots, B_{iT}), B_{it}|C_i = k \sim \text{Bernoulli}(p_{itk});$$

$$\mu_{itk} = e^{x_{it}\beta_k};$$

$$p_{itk} = \frac{1}{e^{-(x_{it}\alpha_k)} + 1};$$

$$C = (C_1, \ldots, C_N), C_i \sim \text{Cat}(\pi_{i1}, \ldots, \pi_{iK});$$

$$\pi_{ik} = \frac{e^{z_i\gamma_k}}{\sum_{h=1}^{K} e^{z_i\gamma_h}}, \text{with} \quad \gamma_1 = 0;$$

$$\alpha = (\alpha_1, \ldots, \alpha_K), \alpha_k \sim N_p(\mu_\alpha, \sigma_\alpha I_p);$$

$$\beta = (\beta_1, \ldots, \beta_K), \beta_k \sim N_p(\mu_\beta, \sigma_\beta I_p);$$

$$\gamma = (\gamma_2, \ldots, \gamma_K), \gamma_k \sim N_r(\mu_\gamma, \sigma_\gamma I_r).$$

2. **Getting joint distribution**

$$\begin{aligned}
&P(Y, C, \alpha, \beta, \gamma; x, z) \\
&= P(Y|C, \alpha, \beta, \gamma; x, z) \times P(C|\alpha, \beta, \gamma; x, z) \times P(\alpha, \beta, \gamma; x, z) \\
&= P(Y|C, \alpha, \beta; x) \times P(C|\gamma; z) \times P(\alpha) \times P(\beta) \times P(\gamma)
\end{aligned}$$

$$P(Y|C, \alpha, \beta; x) = \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{k=1}^{K} P(Y_{it}|C_i = k, \alpha_{\text{k}}, \beta_{\text{k}}; x_{it})^{I(C_i=k)}$$

$$P(C|\gamma; z) = \prod_{i=1}^{N}\prod_{k=1}^{K} P(C_i = k|\gamma_k; z_i)^{I(C_i=k)}$$

$$
\begin{aligned}
&P(Y, C, \alpha,\ \beta, \gamma; x, z) \\
&= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{N} \left\{ P(C_i = k|\gamma_k; z_i) \prod_{t=1}^{T} P(Y_{it}|C_i = k, \alpha_{\mathrm{k}}, \beta_{\mathrm{k}}; x_{it}) \right\}^{I(C_i=k)} P(\alpha_{\mathrm{k}}) P(\beta_k) P(\gamma_k) \right\} \\
&= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{N} \left\{ \pi_{ik} \prod_{t=1}^{T} dzip(\mu_{itk}, p_{itk}) \right\}^{I(C_i=k)} P(\alpha_{\mathrm{k}}) P(\beta_k) P(\gamma_k) \right\} \\
&= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{N} \left\{ \frac{e^{z_i'\gamma_k}}{\sum_{h=1}^{K} e^{z_i'\gamma_h}} \prod_{t=1}^{T} \left[\left[p_{itk} + (1 - p_{itk}) \frac{1}{e^{\mu_{itk}}} \right]^{I(Y_{it}=0)} \left[(1 - p_{itk}) \frac{{\mu_{itk}}^{y_{it}}}{y_{it}! e^{\mu_{itk}}} \right]^{I(Y_{it}>0)} \right] \right\}^{I(C_i=k)} P(\alpha_{\mathrm{k}}) P(\beta_k) P(\gamma_k) \right\} \\
&= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{N} \left\{ \frac{e^{z_i'\gamma_k}}{\sum_{h=1}^{K} e^{z_i'\gamma_h}} \prod_{t=1}^{T} \left[\left[\frac{1}{e^{-({x'}_{it}\alpha_k)} + 1} + \frac{1}{e^{e^{{x'}_{it}\beta_k}} (e^{{x'}_{it}\alpha_k} + 1)} \right]^{I(Y_{it}=0)} \times \left[\frac{e^{({x'}_{it}\beta_k) y_{it}}}{y_{it}! e^{e^{{x'}_{it}\beta_k}} (e^{{x'}_{it}\alpha_k} + 1)} \right]^{I(Y_{it}>0)} \right] \right\}^{I(C_i=k)} P(\alpha_{\mathrm{k}}) P(\beta_k) P(\gamma_k) \right\}
\end{aligned}
$$

3. **Getting full conditional distributions**

$$\pi(\gamma_k|\cdot) \propto \prod_{i=1}^{N} [P(C_i = k|\gamma_k; z_i)]^{I(C_i=k)} \times P(\gamma_k) = \prod_{i:C_i=k} (\frac{e^{z_i'\gamma_k}}{\sum_{h=1}^{K} e^{z_i'\gamma_h}}) \mathrm{N}_r(\mu_\gamma, \sigma_\gamma I_r)$$

$$\pi(C_i|\cdot) = Cat(\rho_{ik}) \propto P(Y_i|C_i, \alpha, \beta; x_i) \times P(C_i|\gamma; z_i)$$

$$\rho_{ik} = \frac{\pi_{ik}(\gamma_k) \prod_{t=1}^{T} [dzip(\mu_{itk}, p_{itk})]}{\sum_{h=1}^{K} \pi_{ih}(\gamma_h) \prod_{t=1}^{T} [dzip(\mu_{ith}, p_{ith})]}$$

$$= \frac{e^{z_i'\gamma_k} \prod_{t=1}^{T} \left\{ \left[p_{itk} + (1-p_{itk}) \frac{1}{e^{\mu_{itk}}}\right]^{I(Y_{it}=0)} \left[(1-p_{itk}) \frac{{\mu_{itk}}^{y_{it}}}{y_{it}! e^{\mu_{itk}}}\right]^{I(Y_{it}>0)} \right\}}{\sum_{h=1}^{K} e^{z_i'\gamma_h} \prod_{t=1}^{T} \left\{ \left[p_{ith} + (1-p_{itk}) \frac{1}{e^{\mu_{itk}}}\right]^{I(Y_{it}=0)} \left[(1-p_{itk}) \frac{{\mu_{itk}}^{y_{it}}}{y_{it}! e^{\mu_{itk}}}\right]^{I(Y_{it}>0)} \right\}}$$

$$\pi(\alpha_k, \beta_k|\cdot) \propto P(Y|C, \alpha_k, \beta_k; x) \times P(\alpha_k) \times P(\beta_k)$$

$$= \prod_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{1}{e^{-(x'_{it}\alpha_k)} + 1} + \frac{1}{e^{e^{x'_{it}\beta_k}} (e^{x'_{it}\alpha_k} + 1)}\right]^{I(Y_{it}=0)} \left[\frac{e^{(x'_{it}\beta_k) y_{it}}}{y_{it}! e^{e^{x'_{it}\beta_k}} (e^{x'_{it}\alpha_k} + 1)}\right]^{I(Y_{it}>0)} \mathrm{N}_p(\mu_\alpha, \sigma_\alpha I_p) \mathrm{N}_p(\mu_\beta, \sigma_\beta I_p)$$

4. **Complete Data Likelihood**

$$
\begin{aligned}
L=&\prod_{i=1}^{N}\Pr(Y_{i1},...,Y_{iT})=\prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i=k)\Pr(Y_{i1},...,Y_{iT}|C_i=k)\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i=k)\prod_{t=1}^{T}\Pr(Y_{it}|C_i=k)\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i=k)\prod_{t=1}^{T}[\Pr(B_{it}=1|C_i=k)\Pr(Y_{it}|C_i=k,B_{it}=1)+\Pr(B_{it}=0|C_i=k)\Pr(Y_{it}|C_i=k,B_{it}=0)]\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\Pr(C_i=k)\prod_{t=1}^{T}\{\Pr(B_{it}=1|C_i=k)\times[I(Y_{it}=0)\Pr(Y_{it}|C_i=k,B_{it}=1)+I(Y_{it}\neq 0)\Pr(Y_{it}|C_i=k,B_{it}=1)]\\
&+\Pr(B_{it}=0|C_i=k)\times[I(Y_{it}=0)\Pr(Y_{it}|C_i=k,B_{it}=0)+I(Y_{it}\neq 0)\Pr(Y_{it}|C_i=k,B_{it}=0)]\}\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\prod_{t=1}^{T}\{p_{itk}\times I(Y_{it}=0)+(1-p_{itk})\times[I(Y_{it}=0)\frac{1}{e^{\mu_{itk}}}+I(Y_{it}\neq 0)\frac{{\mu_{itk}}^{y_{it}}}{y_{it}!e^{\mu_{itk}}}]\}(if Y_{it}\sim \text{ZIP})\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\prod_{t=1}^{T}\{I(Y_{it}=0)\times[p_{itk}+(1-p_{itk})\frac{1}{e^{\mu_{itk}}}]+I(Y_{it}\neq 0)\times(1-p_{itk})\frac{{\mu_{itk}}^{y_{it}}}{y_{it}!e^{\mu_{itk}}}\}\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_{ik}\left\{\prod_{t:Y_{it}=0}[p_{itk}+(1-p_{itk})\frac{1}{e^{\mu_{itk}}}]+\prod_{t:Y_{it}\neq 0}[(1-p_{itk})\frac{{\mu_{itk}}^{y_{it}}}{y_{it}!e^{\mu_{itk}}}\right\}\\
=&\prod_{i=1}^{N}\sum_{k=1}^{K}\frac{e^{z_i'\gamma_k}}{\sum_{h=1}^{K}e^{z_i'\gamma_h}}\left\{\prod_{t:Y_{it}=0}\left[\frac{1}{e^{-(x'_{it}\alpha_k)}+1}+\frac{1}{e^{e^{x'_{it}\beta_k}}(e^{x'_{it}\alpha_k}+1)}\right]+\prod_{t:Y_{it}\neq 0}\left[\frac{e^{(x'_{it}\beta_k)y_{it}}}{y_{it}!e^{e^{x'_{it}\beta_k}}(e^{x'_{it}\alpha_k}+1)}\right]\right\}
\end{aligned}
$$