Biometrics 64, 538–545
June 2008

DOI: 10.1111/j.1541-0420.2007.00884.x

A General Class of Pattern Mixture Models for Nonignorable Dropout with Many Possible Dropout Times

Jason Roy

Center for Health Research, Geisinger Health System, Danville, Pennsylvania 17822, U.S.A.
email: jaroy@geisinger.edu

and

Michael J. Daniels

Departments of Epidemiology and Biostatistics and Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.
email: mdaniels@stat.ufl.edu

Summary. In this article we consider the problem of fitting pattern mixture models to longitudinal data when there are many unique dropout times. We propose a marginally specified latent class pattern mixture model. The marginal mean is assumed to follow a generalized linear model, whereas the mean conditional on the latent class and random effects is specified separately. Because the dimension of the parameter vector of interest (the marginal regression coefficients) does not depend on the assumed number of latent classes, we propose to treat the number of latent classes as a random variable. We specify a prior distribution for the number of classes, and calculate (approximate) posterior model probabilities. In order to avoid the complications with implementing a fully Bayesian model, we propose a simple approximation to these posterior probabilities. The ideas are illustrated using data from a longitudinal study of depression in HIV-infected women.

Key words: Bayesian model averaging; Incomplete data; Latent variable; Marginal model; Random effects.

1. Introduction

Dropout is a common occurrence in longitudinal studies. Missingness induced by dropout that depends only on the observed data is called missing at random (MAR) or random dropout. If missingness depends on the unobserved response at the time of dropout or at future times, even after conditioning on the observed data, then the missingness is called nonignorable or informative dropout (Little, 1995). There are many model-based approaches to deal with informative dropout that are characterized by how they factor the joint distribution of missingness and the response. We will focus on the pattern mixture approach. Pattern mixture models (PMM) are a flexible and transparent way to analyze incomplete longitudinal data where the missingness is nonignorable (Little, 1994; Hogan and Laird, 1997). The typical approach taken in PMM is to stratify on dropout time (i.e., the pattern) and assume that missing data within a pattern are MAR. Consider the case of T unique dropout times and define D_i to be the dropout time and Y_i to be the response vector for subject i. PMM account for nonignorable missingness by allowing the distribution of Y_i to differ by dropout time, that is, $f(y_i \mid D_i) \neq f(y_i)$. So, models are built for $[Y_i \mid D_i]$, but inferences are based on $f(y) = \sum_D f(y \mid D)p(D)$. One issue in this formulation, addressed in Fitzmaurice, Laird, and Shneyer (2001) and Wilkins and Fitzmaurice (2006), is that for nonlinear link functions connecting the means, $E[Y_i \mid D_i]$ to covariates, that is, $g(E[Y_{it} \mid D_i, X_{it}]) = X_{it} \times \beta(D_i)$, the marginal mean, $E[Y_{it}]$, is such that, in general, $g(E[Y_{it} \mid X_{it}]) \neq X_{it} \sum_D \beta(D)p(D)$. This is one issue we will address in our model.

The other issue we will address are situations where the number of unique dropout times T is large. In this setting stratification by dropout pattern may lead to sparse patterns, which will lead to unstable parameter estimates (or unidentified parameters) in those patterns. There are several ways to remedy this including allowing parameters to be shared across patterns (Hogan and Laird, 1997) or to group the dropout times into $m < T$ groups in an ad hoc fashion (Hogan, Roy and Korkontzelou, 2004). Roy (2003) proposed an automated way to do the latter using a latent variable approach within the context of normal models for continuous data. This approach assumes the existence of a discrete latent variable that explains the dependence between the response vector and the dropout time and allows incorporation of uncertainty about the groupings, conditional on a fixed number of groups. We will extend the approach of Roy (2003) by incorporating

uncertainty in the number of classes through (approximate) Bayesian model averaging.

A common way to account for the longitudinal correlation in the vector of responses for subject i, Y_i is to introduce random effects. However, for nonlinear link functions, similar to the above discussion, the link no longer holds for marginal covariate effects (Diggle et al., 2002). We will use the ideas in Heagerty (1999) within our model to directly model the marginal covariate effects. We briefly review Heagerty's approach below.

Let Y_{it} denote the response for the ith subject ($i = 1, \ldots, n$) at time t ($t = 1, \ldots, T$). Heagerty (1999) specifies marginalized logistic models in the following ways. First, the marginal mean of Y_{it} is specified as

$$\text{logit}\{P(Y_{it} = 1 \mid \beta)\} = X_{it}^T \beta. \tag{1}$$

Then the dependence among the Y_{it} is specified via a conditional model that is consistent with (1),

$$\text{logit}\{P(Y_{it} = 1 \mid b_i)\} = \Delta_{it} + b_i, \tag{2}$$

where $b_i \sim N(0, \theta)$. The quantity Δ_{it} is determined by the other parameters in the model and can be computed by solving the following convolution equation,

$$P(Y_{it} = 1) = \int P(Y_{it} = 1 \mid b_i) dF(b_i).$$

Note that Δ_{it} is a function of $X_{it}\beta$ and θ. The overall objective in our approach will be to propose a model that marginalizes over the random effects *and* the dropout distribution to directly model the marginal covariate effects of interest.

This work is widely applicable, but was motivated by an HIV natural history study of depression. The HIV Epidemiology Research Study (HERS; Smith et al., 1997) was a longitudinal study of women with, or at high risk for, HIV infection. Data were collected from 1310 women at baseline. Investigators then attempted to collect data from each subject every 6 months for a total of 6 years. Thus, 12 total visits from each subject would be obtained if there were no missing data. Our interest is in studying the course of depression in the 849 women who had HIV infection at baseline. Depression was treated as a binary, yes/no, variable (Cook et al., 2004). A challenge with the analysis of these data is that less than half of these women remained in the study until the end. It is not hard to imagine a scenario where the course of depression over time might vary as a function of dropout time. Because there are many unique dropout times (12), some of which include very few subjects, we apply the latent class pattern mixture modeling approach to the analysis of these data.

In Section 2 we introduce the model. We provide computational details in Section 3. The example is analyzed in Section 4. A brief simulation study is given in Section 5. We conclude with a discussion in Section 6.

2. Model

Before we introduce the model, we first go through some additional notation needed for the latent class component. Define $S_i = (S_{i1}, \ldots, S_{iM})^T$ to be a vector of latent indicators, where S_{ij} is defined as an indicator for class j, $j = 1, \ldots, M$ ($M < T$; e.g., if subject i is in class j, then $S_{ij} = 1$ and $S_{ij'} = 0$ for all $j \neq j'$). The idea here will be to "group" the dropout times into the M classes as in Roy (2003).

All of the parameters in the following specification are a function of the number of latent classes, M; for example, $\beta^{(M)}$. However, we suppress the superscripts without loss of clarity in the following. First, we specify the marginal mean as

$$g\{\text{E}(Y_{it} \mid \beta)\} = X_{it}^T \beta. \tag{3}$$

By marginal, we mean marginalized over subject-specific random effects *and* over the latent class distribution (implicitly over the dropout distribution as well). If the number of classes M were known, then the parameters β would be of primary interest. We address the issue of M being unknown below.

In order to fully account for correlation due to repeated observations and informative censoring, we specify a conditional model in addition to the marginal model. Recall that we are taking a pattern mixture modeling approach to account for dropout. We assume that the relevant information in D is captured by the latent variable S. We therefore specify a mixture distribution over these latent classes, as opposed to over D itself. Before proceeding to describe the model, however, we first make two points. First, the parameters from the conditional model are not of scientific interest, and in fact are viewed as nuisance parameters; we are not interested in estimating subject-specific effects (i.e., effects conditional on the random effects) or class-specific covariate effects (i.e., effects of covariates on Y given a particular dropout class). Second, we must specify the conditional model in a way that is compatible with the marginal model (3). As we will see below, this leads to a somewhat complicated model. Specifying this conditional model is necessary, however, in order to account for the two types of dependencies (within-subject correlation and dependency between the outcome and dropout time).

We assume the data Y_{it}, conditional on random effects b_i and latent class S_i, are from an exponential family with distribution

$$f(Y_{it} \mid b_i, S_i) = \exp[\{Y_{it}\eta_{it} - \psi(\eta_{it})\}/(m_i\phi) + h(Y_{it}, \phi)],$$

where $\text{E}(Y_{it} \mid b_i, S_i) = g^{-1}(\eta_{it}) = \psi'(\eta_{it})$, η_{it} is the linear predictor, $\psi(\cdot)$ is a known function, ϕ is a scale parameter, and m_i is the prior weight. This family includes normal ($\psi(x) = x^2/2$), binomial ($\psi(x) = \log(1 + e^x)$), and Poisson ($\psi(x) = e^x$) distributions, among others. The conditional mean is specified as

$$g\{\text{E}(Y_{it} \mid b_i, S_i)\} = \Delta_{it} + b_i + \sum_{j=1}^{M} S_{ij} Z_{it}^T \alpha^{(j)}, \tag{4}$$

where, in the most general form of the model we allow the variance of b_i to depend on the latent class, that is, $[b_i \mid S_{ij} = 1] \sim N(0, \theta_j)$. For identifiability, we use a sum-to-zero constraint on the α's, namely, $\alpha^{(M)} = -\sum_{j=1}^{M-1} \alpha^{(j)}$. In this conditional model, each subject has its own intercept, and the effect of each covariate, Z_{itj} ($Z_{it} \subset X_{it}$), is allowed to differ by dropout class via the regression coefficients, $\alpha^{(j)}$.

The probabilities of the latent classes given the dropout time are specified as a proportional odds model,

$$\text{logit}\left\{P\left(\sum_{j=1}^{k} S_{ij} = 1 \,|\, D_i\right)\right\} = \lambda_{0k} + \lambda_1 D_i, \quad k = 1, \ldots, M-1, \quad (5)$$

where $\lambda_{01} \leq \lambda_{02} \leq \cdots \leq \lambda_{0,M-1}$ and λ_1 are unknown parameters. From this regression (5) it is clear that the class probabilities are a monotone function of dropout time (in fact, linear on the logit scale). Finally, the dropout times, D_i, follow a multinomial distribution with mass at each of the possible dropout times, parameterized by γ.

We point out that in the above formulation, Y_{it} is independent of D_i given S_i. This is a key assumption with this approach, which we will examine in Section 3.4.

The intercept Δ_{it} in (4) is determined by the relationship between (3) and (4), namely, the solution to

$$\text{E}(Y_{it} \,|\, \beta) = \sum_D \sum_S p(S_i \,|\, D_i) p(D_i) \int \text{E}(Y_{it} \,|\, b_i, S_i) p(b_i \,|\, S_i)\, db_i.$$

The main target of inference typically will be the covariate effects averaged over classes, that is, $\beta^{(M)}$ averaged over M. We denote this as $\beta^\star = \sum_m \beta^{(m)} p(m \,|\, y)$. We discuss computation of $p(m \,|\, y)$ in Section 3.3 and the corresponding computation of $\text{var}(\hat{\beta}^\star)$.

3. Computational Details

We provide details on computation of maximum likelihood (ML) estimates conditional on m, computation of the approximate posterior model probabilities, and model averaging.

3.1 *The Likelihood and ML Inference*

Denote the set of all parameters by $\omega = (\beta^T, \alpha^T, \theta^T, \phi, \lambda^T, \gamma^T)^T$. We partition the complete response data for subject i, Y_i^c, into observed and missing components. Denote by Y_i the observed part of the vector (i.e., values of Y^c prior to dropout) and by Y_i^m the response after dropout. In the following presentation, assume X_i and M are conditioned throughout.

The likelihood contribution for subject i corresponding to the models described in Section 2 is

$$L_i(Y_i, D_i; \omega) \propto \int \sum_{j=1}^{M} L_i(Y_i \,|\, S_{ij} = 1, b_i; \alpha^{(j)}, \phi) \times p(S_{ij} = 1 \,|\, D_i; \lambda) p(D_i \,|\, \gamma)\, dF(b_i \,|\, S_{ij}, \theta_j), \quad (6)$$

where

$$L_i(Y_i \,|\, S_{ij} = 1, b_i; \alpha^{(j)}, \phi) = \exp\left[\left\{Y_i^T \eta_i - \psi(\eta_i)\right\} \big/ (m_i \phi) + 1^T h(Y_i, \phi)\right],$$

with $\eta_i = \Delta_i + b_i 1 + \sum_{j=1}^{M} S_{ij} Z_i \alpha^{(j)}$, $p(S_{ij} = 1 \,|\, D_i; \lambda)$ is defined in (5), and $p(D_i \,|\, \gamma)$ is the distribution of D_i, which might depend on covariates, and is parameterized by γ. Proportionality in (6) holds because we assume that the missing and observed responses from subject i are independent, given S_i and b_i (i.e., $[Y_i^m \,|\, Y_i, b_i, S_i] = [Y_i^m \,|\, b_i, S_i]$).

Maximization of $\log\{\prod_{i=1}^{n} L_i(Y_i, D_i; \omega)\}$ with respect to the parameters ω is complicated by the possibly intractable integral in (6), and the need to calculate Δ_{it} at each iteration in the algorithm for every record in the data set. We provide details of the maximization algorithm in the Appendix.

3.2 *Posterior Model Probabilities*

The models introduced in Section 2 are indexed by the number of latent classes m ($m = 1, \ldots, M$, $M < T$). Given that our main interest is in the regression parameters β, it would be sensible to properly account for the uncertainty in the regression coefficients by averaging over the number of classes as opposed to conditioning the most likely number of classes. To do this, we need to first specify a prior distribution on the number of latent classes, m. We recommend specifying a prior to favor parsimony and/or to be consistent with subject matter considerations (if available). A convenient specification is a truncated Poisson distribution with rate parameter, μ, and truncated at an integer between 1 and T. Denote this prior as $p(m)$. The posterior probability of m classes is given by the expression,

$$p(m \,|\, y, x) = \frac{p(y \,|\, m, x) p(m)}{p(y \,|\, x)},$$

where $p(y \,|\, x) = \sum_m p(y \,|\, m, x) p(m)$ and $p(y \,|\, m,\ x)$ are the integrated likelihood, that is,

$$p(y \,|\, m, x) = \int p(y \,|\, m, x, \beta^{(m)}, \alpha^{(m)}, \lambda, \gamma, \theta) p(\lambda) \times p(\alpha^{(m)} \,|\, m) p(\beta^{(m)} \,|\, m) p(\gamma) \times p(\theta)\, d\beta^{(m)}\, d\alpha^{(m)}\, d\lambda\, d\gamma\, d\theta,$$

where $p(y \,|\, m, x, \beta^{(m)}, \alpha^{(m)}, \lambda, \gamma, \theta) = \sum_s \sum_D p(y \,|\, m, x, \beta^{(m)}, \alpha^{(m)}, \theta) p(S \,|\, m, x, D, \lambda) p(D \,|\, m, x, \gamma)$. Unfortunately, this integral is not available in closed form. We propose to use a Laplace approximation to evaluate this integral,

$$\hat{p}(y \,|\, m, x) = (2\pi)^{d/2} \,|\, \hat{\Sigma} \,|\, ^{1/2} p(y \,|\, m, x, \hat{\beta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\lambda}, \hat{\gamma}, \hat{\theta}), \quad (7)$$

where $d = \dim(\beta, \alpha, \lambda, \gamma, \theta)$ and $(\hat{\beta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\lambda}^{(m)}, \hat{\gamma}^{(m)}, \hat{\theta}^{(m)})$ are the joint ML estimates of $(\beta^{(m)},\ \alpha^{(m)},\ \lambda^{(m)},\ \gamma^{(m)},\ \theta^{(m)})$ for the model with m classes, $p(y \,|\, m, x, \hat{\beta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\lambda}, \hat{\gamma}, \hat{\theta})$ is the value of the maximized integrated likelihood, and $\hat{\Sigma}$ is the inverse of the observed information matrix for $(\beta,\ \alpha,\ \lambda,\ \gamma,\ \theta)$ based on the integrated likelihood (6). These estimates are obtained using the algorithm described in the Appendix. It is clear that in (7) we have ignored the contribution of the prior, $p(\lambda) p(\alpha^{(m)} \,|\, m) p(\beta^{(m)} \,|\, m) p(\gamma) p(\theta)$, evaluated at the joint ML estimates. This is justified (asymptotically) because the maximized likelihood term, $p(y \,|\, m, x, \hat{\beta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\lambda}, \hat{\gamma}, \hat{\theta})$, is $O_p(n)$ whereas the prior is typically $O_p(1)$. Thus, the approximate posterior probabilities take the form,

$$\hat{p}(m \,|\, y, x) = \frac{\hat{p}(y \,|\, m, x) p(m)}{\sum_m \hat{p}(y \,|\, m, x) p(m)}. \quad (8)$$

3.3 *Model Averaging and Approximate Posterior Inference*

Once the posterior distribution $p(m \,|\, y)$ is estimated, we can then estimate the covariate effects averaged across class sizes. As described previously, we denote the average covariate effect over classes as β^*, which can be estimated as $\hat{\beta}^\star = \sum_m \hat{\beta}^{(m)} \hat{p}(m \,|\, y)$. The variance of $\hat{\beta}^*$ is

$$\text{var}(\hat{\beta}^*) = \text{E}[\text{var}(\hat{\beta}^* \mid M)] + \text{var}(\text{E}[\hat{\beta}^* \mid M])$$

$$= \sum_m \text{var}(\hat{\beta}^* \mid m)p(m \mid y) + \text{var}(\text{E}[\hat{\beta}^{(m)} \mid M]),$$

which can be estimated as

$$\widehat{\text{var}}(\hat{\beta}^\star) = \sum_m \text{var}(\hat{\beta}^{(m)} \mid m)\hat{p}(m \mid y) + \sum_m (\hat{\beta}^{(m)} - \hat{\beta}^*)^{\otimes 2}\hat{p}(m \mid y).$$

Note that if we conditioned the most likely value for the number of classes, m, the variance of the estimated regression coefficients would likely be too small due to ignoring the second term in the variance expression above.

3.4 *Model Checking*

Conditional independence between Y and D given S and X is a key assumption with this modeling approach. A simple method for checking the conditional independence assumption for a given class size M is as follows: this approach was originally proposed by Lin, McCulloch, and Rosenheck (2004), as a modification to the test proposed by Bandeen-Roche et al. (1997). The goal is to test the null hypothesis that model (4) holds versus the alternative that the true model is

$$g\{\text{E}(Y_{it} \mid b_i, S_i, D_i)\} = \Delta_{it} + b_i + \sum_{j=1}^{M} S_{ij} Z_{it}^T \alpha^{(j)} + \sum_{j=1}^{J} h_j(D_i)\phi_j, \quad (9)$$

where each $h_j(\cdot)$ is a known function and the ϕ's are parameters. The null hypothesis is that $\phi_1 = \cdots = \phi_J = 0$. A simple example with $J = 1$ is $h(D_i) = D_i$, which would assume a linear effect of D_i. If class membership S were known, then we could simply fit both the full model (9) and reduced model (3) using ML, and carry out a likelihood ratio test with J degrees of freedom. Because S is unknown, Lin et al. (2004) proposed the following approach.

First, fit the null model (3). We can then estimate the posterior probability of class membership for each subject as

$$\hat{P}(S_{ij} = 1 \mid D_i, Y_i, X_i; \hat{\omega}) = \frac{\int L_i(Y_{obs,i} \mid S_{ij} = 1, b_i; \hat{\alpha}^{(j)}, \hat{\phi})p(S_{ij} = 1 \mid D_i; \hat{\lambda})p(D_i \mid \hat{\gamma})dF(b_i \mid S_{ij}, \hat{\theta}_j)}{L_i(Y_i, D_i;\ \hat{\omega})},$$

where $L_i(Y_i,\ D_i;\ \omega)$ was defined in (6). The next step is to create M replicate pseudo data sets for each record, setting the latent class variable equal to j for the jth replicate of that record. In other words, the entire data set will be replicated M times, and the latent class variable will be set to j for every record in the jth replicate of the data set. Each record is then assigned a case weight based on the corresponding posterior probability of S. For example, a case weight of $\hat{P}(S_{ij} = 1 \mid D_i, Y_i, X_i; \hat{\omega})$ will be assigned to the jth replicate of subject i's data. We can then fit models (9) and (3) using the weighted likelihood, and carry out the likelihood ratio test.

4. Example

As briefly described in the Introduction, we were interested in analyzing data on the longitudinal course of depression of 850 HIV-infected women from the HERS. Depression was measured using the Center for Epidemiologic Studies Depression Scale (CES-D). The CES-D includes 20 questions related to mood, each of which can take a value from 0 (symptom rarely present) to 3 (symptom almost always present). Larger scores indicate the presence of more symptoms, and scores range from 0 to 60. A score of 16 or greater is frequently used as a depression cutoff (e.g., Cook et al., 2004). We therefore defined our outcome Y_{it} as the indicator of depression at visit t, meaning it took a value of 1 if subject i had a CES-D $\geq$16 at visit t, and took a value of 0 otherwise. Our goal was to describe changes in depression over time as a function of baseline characteristics, such as race/ethnicity, number of HIV-related symptoms, injection drug use (IDU), and number of recent adverse events (such as homelessness, violence, and death of a close person).

The observed proportion of depression decreased over time. However, the sample mean is only a valid estimate of the prevalence at each visit if the missing data were missing completely at random (MCAR); it would not be surprising if depression status was related to dropout. There was a substantial amount of dropout. By visit 12, less than half of the original sample remained in the study. We would like to account for the possibility that the prevalence of depression over time might be related to the dropout time.

4.1 *Models*

We first fitted a marginally specified logistic regression model under the MAR assumption. This could also be thought of as a special case of the proposed latent class model, but with $M = 1$ class. We assumed models (1) and (2) hold, where the covariate vector includes an intercept, indicator of black race (black), an indicator of Hispanic ethnicity (latina), an indicator of other race/ethnicity (other), number of HIV-related symptoms during the 6 months prior to the baseline visit (symptoms), an indicator that the subject has been an IDU, number of adverse events in 6 months prior to the baseline visit (adverse), and the HERS visit number (visit). Only `visit` was a time-varying covariate. White race was the reference category for the race/ethnicity variable.

We next fitted models (3–5), with M equal to classes 2, 3, and 4. The covariate vector X_{it} was the same as used in the previous model. We also set $Z_{it} = X_{it}$, meaning that every covariate was allowed to have an effect that varied by dropout class. In order to carry out the model averaging, we needed to estimate the posterior probability for the number of classes. We considered two prior distributions $p(m)$: a discrete uniform prior and a truncated Poisson prior distribution for $M - 1$, with mean equal to 0.5. The truncated Poisson prior placed more prior weight on smaller classes; specifically, the probabilities were 0.6076, 0.3038, 0.0759, 0.0127 for $M =$ 1, 2, 3, and 4, respectively. The posterior distribution of the number of classes for the uniform and truncated Poisson priors was estimated using equation (8). Once the posterior probabilities of the number of classes were calculated, we were able to estimate β^* as described in Section 2. All models were fitted using `R` 2.2.1 software (`http://www.r-project.org`). We

Table 1
The components of the Laplace approximation to the marginal likelihood and the corresponding approximate posterior model probabilities under two priors for the number of classes: a discrete uniform prior and a truncated Poisson prior

	Number of classes			
	1	2	3	4
Number of parameters	9	19	28	37
log likelihood	−3571.829	−3501.31	−3489.768	−3489.751
$(1/2)\log\lvert\hat{\Sigma}\rvert$	−24.51	−44.07	−55.84	−78.22
$d/2\log(2\pi)$	8.27	17.46	25.73	34.00
$P(m\,\vert\,y)$, uniform prior	0	0	1	0
$P(m\,\vert\,y)$, truncated Poisson prior	0	0	1	0

Table 2
Estimates and standard errors of marginal coefficients $\beta^{(m)}$. The estimated covariate effects averaged over classes, β^{*}, are also given in the column for $M = 3$.

	Number of classes							
	1		2		3		4	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.36	0.08	−1.11	0.32	−1.01	0.26	−1.01	0.28
Black	−0.79	0.28	0.67	0.35	0.56	0.25	0.56	0.25
Latina	0.39	0.30	1.37	0.33	1.26	0.27	1.26	0.30
Other	1.06	0.30	0.85	0.38	0.71	0.26	0.71	0.27
Idu	0.59	0.29	0.47	0.19	0.34	0.11	0.35	0.11
Symptom	0.20	0.13	0.45	0.06	0.48	0.05	0.48	0.05
Adverse	0.31	0.04	0.31	0.05	0.32	0.04	0.32	0.05
Visit	−0.04	0.007	−0.02	0.009	−0.03	0.011	−0.03	0.011

wrote functions to calculate each type of likelihood, and used the generic optimization function `optim` to maximize these likelihoods. More details are given in the Appendix.

4.2. *Results*

The results are given in Tables 1 and 2. In Table 1, we compared the four models based on the components of the Laplace approximation of the marginal distribution (7) and the corresponding approximate posterior distribution of the number of classes. First, we examined the maximized likelihood, $p(y \mid \hat{\omega})$. There was a substantial increase in the likelihood (relative to the increase in the number of parameters) by going from 1 to 2 classes. Similarly, there was a modest gain in the likelihood by going from 2 to 3 classes. The likelihood for the four-class model was almost identical to that in the three-class model. The four-class model provided essentially the same fit as the three-class model, but with nine extra parameters. Besides the maximized likelihood, the term $(d/2)\log(2\pi)$ always increases as the number of parameters (d) increases. However, the determinant of the estimated covariance matrix, $|\hat{\Sigma}|$ typically decreases as the number of parameters increases; this acts as a "penalty" term for adding parameters. In particular, consider the comparison between models 3 and 4. In model 4 we added nine new parameters. These parameters did little to improve the fit to the data, as the likelihood only increased by a small amount. These parameters were not well identified by the model, and tended to have large variances and high correlation with other parameters. This caused the determinant of the estimated covariance matrix to be considerably smaller than from the three-class model.

The posterior distribution of M was insensitive to the choice of the prior ($p(M = 3\,|\,y,\,x) = 0.9997$ with the uniform prior, and $p(M = 3\,|\,y,\,x) = 0.9987$ with the truncated Poisson prior). The three-class model was the clear "winner" based on the posterior model probabilities; no reasonable prior would change this conclusion. Due to the closeness of the posterior probability of the three-class model to 1, there was no need to carry out the model averaging. In particular, recall that $\hat{\beta}^{\star} = \sum_m \hat{\beta}^{(m)}\hat{p}(m\,|\,y)$. Because, from Table 1, $\hat{p}(M = 3\,|\,y) = 1$, then the estimated parameters from the three-class model, $\hat{\beta}^{(3)}$, were equivalent to the estimated parameters that were averaged over the number of classes $\hat{\beta}^{*}$.

The marginal regression coefficient estimates are presented in Table 2 for each model. The parameter estimates from the one-class model were quite different from the models with multiple classes. For example, based on the one-class model, we might conclude that the prevalence of depression was lower for blacks. However, once we account for dropout using the latent class model, we conclude the opposite.

Because the posterior probabilities overwhelmingly favored the three-class model, we will now focus on this model for our conclusions. Blacks, Latinas, and other non-white racial and ethnic groups were estimated to have a significantly higher

Table 3
Comparison of the estimated latent class probabilities as a function of the dropout time for the three-class model

	Dropout time (visit number of last observed value)											
Class	1	2	3	4	5	6	7	8	9	10	11	12
1	0.89	0.86	0.81	0.75	0.68	0.61	0.53	0.44	0.36	0.29	0.22	0.17
2	0.09	0.13	0.17	0.21	0.27	0.32	0.38	0.43	0.47	0.49	0.50	0.48
3	0.01	0.02	0.03	0.04	0.05	0.07	0.09	0.12	0.17	0.22	0.28	0.35

prevalence of depression as compared with whites. IDU, the number of adverse events, and HIV-related symptoms were associated with higher prevalence of depression. There was a significant, but somewhat gradual, decline in depression over time. We also considered interactions between race/ethnicity and visit number, but these interactions did not appear to be important in describing the data.

Table 3 displays estimated latent class probabilities as a function of dropout time, using the estimated values of λ, the ordinal regression parameters in (5). Individuals who dropped out early (after visit 1) were very likely to be in class 1. Individuals who remained in the study until the end were most likely to be in class 2. Class 3 consisted of a small subpopulation of the subjects who dropped out in the final few visits of the study.

4.3. *Checking the Conditional Independence Assumption*

We used the method described in Section 3.4 to test the null hypothesis of conditional independence. For each value of M (1–4), we fitted model (9), with $\sum_{j=1}^{J} h_j(D_i)\phi_j = D_i\phi$. The test statistic, which, under the null hypothesis follows an approximate χ^2_1 distribution, had values of 7.81, 2.64, 0.41, and 0.41 for $M = 1$ to $M = 4$, respectively. Thus, with respect to the specific alternative of a linear effect of dropout time, the conditional independence assumption appeared to be reasonable for $M = 3$.

5. Simulation Study

We carried out a brief simulation study, primarily to examine the effectiveness of the approximation to fully Bayesian inference. For covariates, we used variables from the HIV data described in the previous section. In particular, the X matrix included an intercept, the indicator of IDU, and visit number. The true values of the β parameters were -1.1, 0.45, and -0.02 for the intercept, IDU, and visit, respectively.

We first generated the response for the case where $M = 1$ (where the MAR assumption holds). The missing data pattern was just the observed pattern from the HIV data. The response was generated from models (1) and (2). We also generated data for the case where $M = 2$. The missing data pattern was the same, but now the response depended on class membership. The latent class variable was generated from model (5) with $\lambda_{01} = 4$ and $\lambda_1 = -0.7$. We then set $\alpha^{(1)} = (0.003, -0.16, 0.24)^T$ in (4) and generated the response. In each case, the variance of the random intercept was $\theta = 4$. These parameter values are equal to their estimated values from the two-class model fitted in the previous section.

For each generated data set, we fitted a marginally specified logistic regression model under the MAR assumption ($M = 1$). We also fitted the latent class model proposed in the manuscript. In that case, we fitted a one-, two-, and three-class model, and carried out model averaging assuming a discrete uniform prior over the three classes. One hundred simulated data sets were analyzed under each scenario. The percentage bias, average estimated standard error (SE), the estimated standard deviation of the estimates (ESD), as well as coverage probability were recorded. For model averaging, $\hat{\beta}^*$ was reported. The results are given in Table 4.

When the data were generated under the MAR assumption ($M = 1$), both modeling approaches worked reasonably well. The estimates had very little bias. The SEs tended to be slightly underestimated. Coverage was below the nominal

Table 4
Results from simulation study. Percentage bias, average of the estimated standard errors (SE), empirical standard deviation (ESD), and 95% coverage probabilities are reported for the estimated marginal regression coefficients.

	Fitted model: MAR				Fitted model: latent class			
Parameter	% bias	SE	ESD	coverage	% bias	SE	ESD	coverage
				True model: MAR				
Intercept	0.1	0.07	0.08	0.86	3.5	0.07	0.08	0.87
IDU	−1.1	0.12	0.14	0.91	−1.9	0.12	0.14	0.91
Visit	0.0	0.01	0.01	0.98	−3.2	0.01	0.01	0.93
				True model: latent class ($M = 2$)				
Intercept	7.6	0.07	0.07	0.74	8.0	0.07	0.07	0.98
IDU	−0.9	0.13	0.16	0.87	−1.6	0.12	0.12	0.94
Visit	−582	0.01	0.01	0.00	−26	0.02	0.02	0.91

for the intercept. We did not expect the coverage and ERs to be exact as we used large sample results for inference here.

When data were generated from the two-class model (MAR assumption violated), the model that relied on the MAR assumption ($M = 1$) no longer performed well. In general, coverage probabilities were too low. In particular, the estimated coefficient of visit number had a large negative bias (582%) and no coverage. The model averaging approach yielded better results. The coefficient of visit number had negative bias (26%) with coverage probability of 0.91. The bias comes from putting some weight on the incorrect model (MAR); the coefficient of visit number conditional on $M = 2$ had a bias of just 3%.

For data generated from the one-class model (MAR), the one-class model had the highest posterior probability in 44% of samples. Here, the two-class model was slightly favored, which is only an incorrect model in the sense that it has more parameters than necessary. For data generated from the two-class model, the two-class model had the highest posterior probability in 81% of samples. The one-class model (MAR) only had the highest probability in 2% of samples.

To confirm that the model probabilities would converge to the correct values as the sample size increased, we simulated data from the same model as described above, but with a sample size of 3400 (4 copies of the covariate data from 850 subjects were used). We fitted five simulated data sets from the one-class model (MAR) and from the two-class model. In each case, the posterior model probability for the correct M was greater than 0.99.

6. Discussion

We have proposed a new model for dealing with nonignorable missing data that parsimoniously addresses data sets with many possible dropout times (in an automated fashion) and directly models the marginal covariate effects of interest. Via approximate posterior model probabilities for the number of latent classes, this approach properly takes into account uncertainty in the unknown number of classes.

We fitted the model using approximate Bayesian methods. Reversible jump Markov chain Monte Carlo methods (Green, 1995) would be required to fit a fully Bayes model because the dimension of the parameter space changes with the number of latent classes.

For the model proposed here, we have assumed a simple within-class longitudinal dependence structure through the introduction of a random intercept. More flexible specifications of the dependence structure could be obtained by replacing the scalar random effect b_i with a set of correlated random effects $b_i = (b_{i1}, \ldots, b_{iT}$; though this will necessitate higher dimensional numerical integrations) or by allowing dependence through a Markov transition structure within class (Heagerty, 2002).

Alternative methods for specifying marginal effects for correlated binary data have been proposed. Caffo, An, and Rohde (2006) proposed a model for binary data with random effects, which uses mixtures of normals. Their approach is less computationally intensive than the Heagerty (1999) approach that we implemented here. However, extending their approach to also average over the discrete latent dropout distribution would likely prove challenging. In particular, the additional step of averaging over the latent dropout classes would make it difficult to preserve the marginal probit interpretation. Wang and Louis (2003) proposed a bridge distribution function for binary random intercept models. However, extending their approach to our setting would likely pose similar problems for mixture of normals approach of Caffo et al. (2006).

The model proposed here assumes conditional independence between the outcome and dropout processes, given the latent class and covariates. We tested this assumption against a very simple alternative hypothesis (linear effect of dropout time). A more complicated approach would be to leave the functional form of the dependence unspecified. Specifically, we could assume

$$g\{\mathrm{E}(Y_{it} \mid b_i, S_i, D_i)\} = \Delta_{it} + b_i + \sum_{j=1}^{M} S_{ij} Z_{it}^T \alpha^{(j)} + f(D_i),$$

where $f(\cdot)$ is a smooth, but otherwise unspecified function. The null hypothesis of conditional independence would be $f(D_i) = 0$. We plan to explore a score-type test similar to that proposed by Zhang and Lin (2003) and Lin, Zhang, and Davidian (2006) and examine its asymptotic distribution for the models proposed here.

Acknowledgements

Dr Roy's research was supported by National Institutes of Health grant R01-HL-79457. Dr Daniels received National Institutes of Health grants R01-HL-79457 and R01-CA-85295. Data from the HER Study were collected under grant U64-CCU10675 from the U.S. Centers for Disease Control and Prevention. The authors thank the associate editor and a referee for their helpful comments and suggestions.

References

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92,** 1375–1386.

Caffo, B., An, M.-W., and Rohde, C. A. (2006). A flexible general class of marginal and conditional random intercept models for binary outcomes using mixtures of normals. Johns Hopkins University, Department of Biostatistics Working Papers. Working Paper 98. Available at: `http://www.bepress.com/jhubiostat/paper98`

Cook, J. A., Grey, D., Burke, J., Cohen, M. H., Gurtman, A. C., Richardson, J. L., Wilson, T. E., Young, M. A., and Hessol, N. A. (2004). Depressive symptoms and AIDS-related mortality among a multisite cohort of HIV-positive women. *American Journal of Public Health* **94,** 1133–1140.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *The Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

Fitzmaurice, G. M., Laird, N. M., and Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable dropouts. *Statistics in Medicine* **20,** 1009–1021.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55,** 688–698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58,** 342–351.

Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16,** 239–257.

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Handling dropout in longitudinal data. *Statistics in Medicine* **23,** 1455–1497.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* **60,** 295–305.

Lin, J., Zhang, D., and Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models *Biometrics* **62,** 803–812.

Little, R. J. A. (1994). A class of pattern mixture models for normal missing data. *Biometrika* **81,** 471–483.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90,** 1112–1121.

Roy, J. (2003). Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model. *Biometrics* **59,** 829–836.

Smith, D. K., Warren, D. L., Vlahov, D., Schuman, P., Stein, M. D., Greenberg, B. L., and Holmberg, S. D. (1997). Design and baseline participant characteristics of human immunodeficiency virus epidemiology research (HER) study: A prospective cohort study of human immunodeficiency virus infection in US women. *American Journal of Epidemiology* **146,** 459–469.

Wang, Z. and Louis, T. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* **90,** 765–775.

Wilkins, K. J. and Fitzmaurice, G. M. (2006). A hybrid model for nonignorable dropout in longitudinal binary responses. *Biometrics* **62,** 168–176.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

Received March 2006. *Revised April* 2007.
Accepted May 2007.

Appendix

Computational Details of ML

We propose the following approach to compute the ML estimates. First, we obtain initial values of the parameters. Initial values of β and θ could be obtained from ML estimates of a model that assumes an ignorable missing data mechanism. The parameters λ initially should be selected in a way that leads to marginal probabilities not too close to zero for any latent class. Initial values of α could be obtained by fitting a pattern mixture model with M groups of dropout times that have fixed boundaries. Given the data and parameters ω, we next calculate Δ_{it} for all i and t. We accomplish this using Newton Raphson with numerical differentiation and integration. Specifically, we solve $h(\Delta_{it}) - g^{-1}(X_{it}^T\beta) = 0$ for Δ_{it}, where

$$h(\Delta_{it}) = \sum_{d=1}^{T}\sum_{j=1}^{M}\left\{\int g^{-1}\left(\Delta_{it} + b_i + Z_{it}^T\alpha^{(j)}\right) p(b_i \mid S_{ij} = 1)\, db_i\right\} \times p(S_{ij} = 1 \mid D_i = d)p(D_i = d)$$

and $p(b_i \mid S_{ij} = 1)$ is $N(0, \theta_j)$, and $p(S_{ij} = 1 \mid D_i = d)$ can be found using equation (5). A 10-point Gauss–Hermite quadrature is used to integrate out the random effects b_i from the above equation. The derivative of $h(\Delta_{it})$ with respect to Δ_{it} is $h'(\Delta_{it})$, which is found using standard numerical techniques. We then find the value of Δ_{it} by repeatedly calculating $\Delta_{it}^{\text{new}} = \Delta_{it}^{\text{old}} - \{h(\Delta_{it}^{\text{old}}) - g^{-1}(X_{it}^T\beta)\}/h'(\Delta_{it}^{\text{old}})$ until convergence. Once we have values of Δ_{it} for the current set of parameters ω, we can then evaluate the likelihood (6), where again Gauss–Hermite quadrature is used to evaluate the integral. Many possible algorithms could then be used to find the ML estimates. For example, one could use a Newton Raphson approach, which would require calculating the likelihood at various points to get numerical estimates of the score and Hessian at each step. However, the log likelihood for many latent class models tends to be poorly behaved (e.g., more than one local maximum). Algorithms such as Newton Raphson or Fisher scoring may not perform well. Our recommendation is start with a more stable, robust algorithm, such as Nelder–Mead, and then switch to a faster algorithm such as Newton Raphson for the final steps.