Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology

Kathryn ROEDER, Kevin G. LYNCH, and Daniel S. NAGIN

Social scientists are commonly interested in relating a latent trait (e.g., criminal tendency) to measurable individual covariates (e.g., poor parenting) to understand what defines or perhaps causes the latent trait. In this article we develop an efficient and convenient method for answering such questions. The basic model presumes that two types of variables have been measured: response variables (possibly longitudinal) that partially determine the latent class membership, and covariates or risk factors that we wish to relate to these latent class variables. The model assumes that these observable variables are conditionally independent, given the latent class variable. We use a mixture model for the joint distribution of the observables. We apply this model to a longitudinal dataset assembled as part of the Cambridge Study of Delinquent Development to test a fundamental theory of criminal development. This theory holds that crime is committed by two distinct groups within the population: adolescent-limited offenders and life-course–persistent offenders. As these labels suggest, the two groups are distinguished by the longevity of their offending careers. The theory also predicts that life-course–persistent offenders are disproportionately comprised of individuals born with neurological deficits and reared by caregivers without the skills and resources to effectively socialize a difficult child.

KEY WORDS: Classification error; Latent class analysis; Mixture models.

1. INTRODUCTION

Latent class analysis, a technique widely used in the social sciences, is based on the theory that individuals differ in their behaviors due to some unobservable latent trait (Clogg 1995; Langeheine and Rost 1988; Muthen and Shedden 1999; Rost and Langeheine 1997). Social scientists often are interested in relating latent traits to some other variables, with the ultimate purpose of understanding what defines or perhaps causes the latent trait (Nagin, Farrington, and Moffitt 1995). If the latent traits could be observed, then it would be a simple matter to analyze the data using techniques such as contingency table analysis. In this article we develop a technique for handling uncertainty in latent class assignment by building a complex mixture model for the full dataset. This method is applied to a longitudinal study of youths to answer a key question in criminology.

The aim of life-course and developmental theories of crime and deviance (Farrington 1986; Hawkins, Lisher, Catalano, and Howard 1986; Huizinga, Esbensen, and Weiher 1991; Loeber and LeBlanc 1990; Moffitt 1993; Sampson and Laub 1991, 1993) is to document and explain the evolution of crime and deviance from childhood through adulthood. Because of its developmental emphasis, this literature aims to identify the causes for population differences in *trajectories* of offending. As a first step toward this end, Nagin and Land (1993) developed a semiparametric estimation procedure designed to identify distinctive groupings of offending trajectories. Fitting the model to a sample of British males who were tracked from ages 10–32, four distinctive age–crime trajectories were discovered.

In this article we investigate a theory proposed by Moffitt (1993) positing that the interaction of two key factors—poor neurological development and poor parenting—is highly predictive of criminal career development. To test this theory, we extend the mixture model approach of Land, McCall, and Nagin (1996) and Nagin and Land (1993) to incorporate time-stable covariates. We apply these methods to the longitudinal study of British males just described (Sect. 6). The methods that we develop go beyond those in the literature in two ways: We allow for the uncertainty of latent class membership; and we develop a model for multivariate analysis of risk factors (Sects. 3–5).

Although a joint probability model for offending patterns and risk factors is appealing, the resulting mixture model is complex. There are three levels of choices in the model structure. First, our semiparametric approach requires us to estimate the number of latent classes, K. Next, the form of each trajectory must be specified; for our application, even the marginal model for offending patterns has 3^{K+1} possible models. Finally, with the addition of covariates, the model space quickly escalates to unmanageable proportions. It is impractical to consider a complete search of the set of all possible models. Such an endeavor would be computationally intensive and tedious, as each model requires a careful choice of starting values to ensure covergence. In this article we develop approximations that allow fast, efficient comparison of a large number of competing models using standard software, after a preliminary exploration of the space of trajectories is completed (Sects. 4–5).

In contrast to a full mixture approach, a common practice is to do a two-stage analysis. In the first stage, response variables are used to categorize individuals by latent trait; then, in a second stage, standard methods of analysis are used to identify cross-group differences. As Clogg (1995) pointed out, there are inherent dangers in this *classify–analyze* paradigm, which ignores the uncertainty of latent

Kathryn Roeder is Professor, Department of Statistics (E-mail: *roeder@stat.cmu.edu*), and Daniel Nagin is Teresa and H. John Heinz III Professor, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213. Kevin G. Lynch is Research Instructor, Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213. The authors thank Bobby L. Jones for developing the SAS software required for the data analyses. The SAS procedure is available free of charge at *http://www.stat.cmu.edu/~bjones/traj.html*. This research was supported by National Science Foundation grant SBR-9511412 and the National Consortium for Violence Research.

Journal of the American Statistical Association
September 1999, Vol. 94, No. 447, Applications and Case Studies

trait assignment. In addition to examining the aforementioned hypotheses using our mixture models, we examine the importance of accounting for uncertainty in group membership by comparing our results to those obtained using the classify–analyze approach (Sect. 6).

Although the methods developed in this article are aimed at a model designed specifically for criminology data, the techniques for handling uncertainty in latent class analysis are quite generally applicable.

2. DATA

A central goal of developmental research on criminal behavior is to determine the risk factors that distinguish criminals from noncriminals and chronic offenders from infrequent offenders. To explore risk factors, we analyze a panel dataset of criminal involvement (Farrington and West 1990) comprising a prospective longitudinal survey of 403 males from London. Data collection began in 1961–1962, when the youths were 8 years old, and continued for 22 years. Their criminal involvement is measured by convictions for criminal offenses. For those convicted of at least one crime (36%), the average number of convictions over the observation period is 4.4.

According to Moffitt's (1993) theory, criminal and delinquent acts are committed by two groups: adolescent-limited offenders and life-course–persistent offenders. As the labels are intended to suggest, the two groups differ in the longevity of their offending careers. For the life-course–persistent offenders, criminality is part of an ongoing pattern of antisocial behavior that has its origins in childhood and continues throughout life. In childhood, life-course–persistent offenders are the most troublesome children, in adolescence they are delinquent, and as adults they commit crimes and abuse themselves and those around them. Moffitt argues that these offenders commonly suffer from neurological deficits that make them inattentive and very difficult infants and children. However, such deficits are not in themselves sufficient to result in a lifetime of antisocial behavioral pattern. Another necessary factor in the development of a propensity for chronic antisocial behavior is being reared by caretakers who lack the financial, psychological, and child-rearing skills necessary to effectively socialize a difficult child. By contrast, the brief tenure of offending of the adolescent-limited offenders does not reflect a fundamental failure in the person's capacity for socialization or the effectiveness of caretakers in providing such socialization. Indeed, Moffitt argues that some degree of delinquency is normal and simply represents a passing phase in the developmental process.

The Cambridge dataset is particularly rich in measurements of three categories of risk factors: (1) intelligence and attainments, which includes IQ and success in school; (2) antisocial family and parenting factors, which includes measurements taken during early adolescence of parental child-rearing practice and of antisocial behavior of parents and siblings; and (3) hyperactivity, impulsivity, and attention deficits, which includes reports by teachers, parents, and the individual himself of restlessness, impulsive and daring behavior, and an inability to concentrate. (For reviews of the association of these risk factors with antisocial behaviors, see Wilson and Herrnstein 1985 for the first category, Loeber and Stouthamer-Loeber 1986 for the second, and Moffitt 1993 for the third.)

Our analysis focuses on the two risk factors identified by Moffitt: neurological deficiency and poor parenting. We recorded the risk factors into a binary format, with the transformed variable coded 1 if the individual scored high for the potential risk factor (i.e., generally the highest/lowest quartile of the measurement scale) and 0 otherwise.

Following Nagin et al. (1995), we create an 11-period panel, starting at age 10, in which each period is a 2-year interval. We observe the number of crimes committed by the ith subject in the jth time period, Y_{ij}, for $n = 403$ subjects at $J = 11$ time periods of length $b = 2$ years; $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ denotes the crime history for the ith individual. The time stable binary variables $\mathbf{Z}_i$, representing the two risk factors, are also given for each individual.

3. THE MODEL

Our goal is to identify childhood covariates that predict a tendency toward criminal behavior. A useful construct to facilitate understanding of the relationship between trajectories of crime and risk factors is a latent trait that explains differences in individual behavior. It is assumed that risk factors can influence (and be influenced by) latent class and that the latent class determines the likelihood of criminal behavior, but that criminal behavior and risk factors are conditionally independent, given latent class. That is, given the latent class, nothing more can be learned about the criminal activity from the risk factor, or vice versa. This conditional independence assumption greatly simplifies the resulting models, which allows for more flexible modeling of other components of the problem.

An unobservable discrete variable C_i indicates the latent class of the ith individual. This variable is assumed to take on K distinct values, each of which corresponds to a distinct expected crime trajectory.

At this stage, the model could be developed in two equivalent ways, either

$$f(\mathbf{y}, \mathbf{z}) = \sum_{k=1}^{K} \Pr(C = k)\Pr(\mathbf{z}|C = k)\Pr(\mathbf{y}|C = k), \quad (1)$$

where $\Pr(\mathbf{z}|C = k)$ is modeled as a multinomial (assuming that $\mathbf{Z}$ is a categorical variable) or

$$f(\mathbf{y}, \mathbf{z}) = \Pr(\mathbf{z}) \sum_{k=1}^{K} \Pr(C = k|\, \mathbf{z})\Pr(\mathbf{y}|C = k), \quad (2)$$

where $\Pr(C = k|\, \mathbf{z})$ is modeled as a K-outcome logit with covariate $\mathbf{Z}$. From the former representation, it is clear that the model follows a standard finite mixture distribution with K components (e.g., Lindsay 1995). We utilize the latter approach because we found it more convenient for estimation purposes. Moreover, this approach extends naturally to continuous covariates, $\mathbf{Z}$.

Others have either treated $\mathbf{Z}$ as a covariate in the model for $\mathbf{Y}$ (e.g., Land et al. 1996; Nagin and Land 1993) or relegated it to a post hoc analysis (e.g., Nagin et al. 1995). The former approach is not helpful in addressing the research question of why individuals are on different criminal trajectories. Rather, it attempts to determine how covariates modify trajectories. The latter approach is not rigorous statistically, because these post hoc analyses do not account for uncertainty in latent class assignment.

3.1 Modeling the Risk Factors

We use a polychotomous logistic regression model to relate the set of risk factors to the criminal career trajectories. Let

$$p_k(\mathbf{z}) = \Pr(C = k|\, \mathbf{Z} = \mathbf{z}) = \frac{\exp\{\theta_k + \gamma_k' \mathbf{z}\}}{\sum_{k=1}^{K} \exp\{\theta_k + \gamma_k' \mathbf{z}\}},$$

where $\mathbf{Z} = (Z_1, \ldots, Z_r)$ is a vector of random variables that are potential risk factors and interactions of these risk factors, θ_k is a scalar, and $\gamma_k = (\gamma_{1k}, \ldots, \gamma_{rk})$ is a vector of length r. Let $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_K)$ and $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_K)$; for identifiability, we take $\theta_1 = 0$ and $\gamma_1 = 0$. With this parameterization, level 1 is considered the baseline level, and the log odds of membership in level k versus 1 are linear in $\mathbf{Z}$: $\log\{p_k(\mathbf{z})/p_1(\mathbf{z})\} = \theta_k + \gamma_k' \mathbf{z}$. We wish to determine whether γ_k varies by k. For instance, if $\gamma_{h,2} = \cdots = \gamma_{h,K} = 0$, then Z_h is not related to criminal career trajectories.

3.2 Modeling the Trajectories

Although the focus of this article is modeling the relationship between $\mathbf{Z}$ and C, we use the longitudinal data $\mathbf{Y}$ to learn about the unobservable C. To model $\mathbf{Y}|C$, we adopt Nagin and Land's (1993) model, with minor modifications (see also Greene 1997). However, most of the methods that we present are applicable regardless of how the distribution of $\mathbf{Y}|C$ is modeled. For example, in Nagin and Temblay (1999), $\mathbf{Y}|C$ follows a censored normal distibution.

We call the model for $\mathbf{Y}|C$ a mixture of zero-inflated Poissons (MZIP), because it is a generalization of the zero-inflated Poisson (ZIP) model studied by Lambert (1992). These models were developed to handle situations where more 0's are present than expected if the data were Poisson distributed. In the criminology application, the zero inflation occurs because individuals seem to enter periods of dormancy during which the probability of crime is strictly 0. Nagin and Land's model makes the convenient assumption that periods of delinquent activity or dormancy occur in periods of fixed length b (the observation period). The individual is assumed to be active during all of b with probability $1 - \rho_{ij}$. During periods of activity, a Poisson (λ_{ijk}) model is used to describe the probability distribution of the number of crimes committed by individual i during time period j assuming that he is a member of latent class k. Notice that an individual may be active but still have no recorded offenses.

We assume that the observed crime histories $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ are independent and that, conditional on $C_i = k$, (1) a subject's counts are independent across time periods and (2) are distributed as ZIP's within a time period

$$Y_{ij}|C_i = k \sim \begin{cases} 0 & \text{with probability } \rho_{ij} \\ \text{Poisson}(\lambda_{ijk}) & \text{with probability } 1 - \rho_{ij}. \end{cases} \tag{3}$$

The parameters ρ_{ij} and λ_{ijk} are assumed to be linear in their canonical parameterizations with $\log \lambda_{ijk} = \mathbf{T}_{ij}\beta_k$ and $\text{logit}(\rho_{ij}) = \log[\rho_{ij}/(1 - \rho_{ij})] = \mathbf{X}_{ij}\alpha$, where $\mathbf{T}_{ij}$ and $\mathbf{X}_{ij}$ are vectors of covariates.

Among the best-documented facts about crime is the age–crime curve. On average, rates of offending rise rather rapidly during the early adolescence, reach a peak in the late teenage years, and then begin a gradual but steady decline (Farrington 1986; Hirschi and Gottfredson 1983). Consequently, for the crime trajectory, λ_{ijk}, we use a quadratic function of the age of the individual, $\mathbf{T}_{ij} = (1, t_{ij}, t_{ij}^2)$. By allowing $\beta_k = (\beta_{0k}, \beta_{1k}, \beta_{2k})$ to vary over latent classes, we obtain K age–crime trajectories. The propensity to commit crimes ρ_{ij}, which Nagin and Land call the *intermittency parameter*, may also change with age, so let $\mathbf{X}_{ij} = \mathbf{T}_{ij}$.

Given $C_i = k$, the probability of observing the ith individual's crime history is

$$q_{ik} = \prod_j \Pr(Y_{ij}|C_i = k), \tag{4}$$

where each term in the product is distributed as a ZIP,

$$\Pr(Y_{ij}|C_i = k) = \begin{cases} \rho_{ij} + (1 - \rho_{ij})e^{-\lambda_{ijk}} & Y_{ij} = 0 \\ (1 - \rho_{ij})\, \frac{e^{-\lambda_{ijk}} \lambda_{ijk}^{y_{ij}}}{y_{ij}!} & Y_{ij} > 0 \end{cases}. \tag{5}$$

The marginal likelihood based on $\mathbf{y}$ is

$$\mathcal{L}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \sum_k p_k\, q_{ik}, \tag{6}$$

where $p_k = \Pr(C = k)$. The joint likelihood based on $(\mathbf{y}, \mathbf{z})$ is

$$\mathcal{L}(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \Pr(\mathbf{Z}_i = \mathbf{z}_i) \sum_k p_k(\mathbf{z})\, q_{ik}. \tag{7}$$

Both likelihoods are identifiable under mild conditions, including those specified by Lambert (1992) for ZIP regression models and a restriction that no trajectory has mean 0.

Our model for $\mathbf{Y}|C$ differs from Nagin and Land's in that we use a logit rather than a probit link for ρ_{ij} and we model time-stable covariates with a logit link in the model for $C|\mathbf{Z}$. Nagin and Land include time-stable covariates as a composite variable in the covariate vectors, $\mathbf{T}$ and $\mathbf{X}$, and also include a lag variable for prior behavior in the intermittency parameter.

4. ESTIMATION

The parameters of this model can be estimated by a direct maximization procedure available in SAS (Jones, Nagin, and Roeder in press). But use of this procedure is dependent upon a careful choice of starting values. Here we present an EM algorithm (Dempster, Laird, and Rubin 1977) that leads to a factorization of the likelihood into independent components. Insights into the model gleaned from the EM representation motivate a highly economical algorithm for preliminary model screening.

4.1 Maximizing the Likelihood Using the EM Algorithm

Two latent variables underly the model for $f(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\beta})$: An individual belongs to a particular criminal career trajectory $(C = k)$, and an individual is in either a dormant $(D = 1)$ or an active $(D = 0)$ state. Treating the latent variables $(C_1, \ldots, C_n, D_{11}, \ldots, D_{Jn})$ as missing data leads naturally to the use of the EM algorithm. (See Lambert 1992 for similar results for the ZIP model.) The primary advantages of this algorithm are numerical stability and a factorization of the likelihood. It is easy to implement the EM algorithm in this setting using software designed for fitting generalized linear models (GLM's). The primary disadvantage is speed of convergence.

The EM algorithm requires the iterative expectation (E) and maximization (M) of the complete log-likelihood $l(\mathbf{y}, \mathbf{z}, c, d; \boldsymbol{\theta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, which equals

$$\begin{aligned}
&\log \prod_i \prod_k \Bigg\{ p_k(\mathbf{Z}_i) \\
&\quad \times \Bigg(\prod_j [\rho_{ij} \Pr(Y_{ij} | C_i = k, D_{ij} = 1)]^{I\{D_{ij}=1\}} \\
&\quad \times [(1 - \rho_{ij}) \Pr(Y_{ij} | C_i = k, D_{ij} = 0)]^{I\{D_{ij}=0\}} \Bigg) \Bigg\}^{I\{C_i=k\}} \\
&\quad \times \Pr(\mathbf{Z}_i). \qquad (8)
\end{aligned}$$

We fit the saturated model for the covariates (**Z**), and thus gain no information from these observations. Consequently, these terms may be dropped from the likelihood.

Let c_{ik} and d_{ij} be the estimates of $E[I\{C_i = k\} | \mathbf{Y}, \mathbf{Z}]$ and $E[I\{D_{ij} = 1\} | \mathbf{Y}, \mathbf{Z}]$ obtained in the E step:

$$c_{ik} = \Pr(C_i = k | \mathbf{Y}_i, \mathbf{Z}_i) = \frac{p_k(\mathbf{Z}_i)\, q_{ik}}{\sum_k p_k(\mathbf{Z}_i)\, q_{ik}} \qquad (9)$$

and

$$d_{ij} = \Pr(D_{ij} = 1 | \mathbf{Y}_i, \mathbf{Z}_i) = \frac{\rho_{ij} \Pr(Y_{ij} | \mathbf{Z}_i, D_{ij} = 1)}{\Pr(Y_{ij} | \mathbf{Z}_i)}, \qquad (10)$$

where

$$\Pr(Y_{ij} | \mathbf{Z}_i, D_{ij} = 0) = \sum_k p_k(\mathbf{Z}_i)\, \frac{e^{-\lambda_{ijk}} \lambda_{ijk}^{y_{ij}}}{y_{ij}!}. \qquad (11)$$

Note that d_{ij} simplifies to

$$d_{ij} = \begin{cases} 0 & Y_{ij} > 0 \\ \left(1 + \sum_k p_k(\mathbf{Z}_i) e^{-\mathbf{X}_{ij}\alpha - \exp\{\mathbf{T}_{ij}\beta\}}\right)^{-1} & Y_{ij} = 0. \end{cases} \qquad (12)$$

The conditional expectation of (8), given **Y** and Z, equals

$$\begin{aligned}
&\sum_i \sum_k c_{ik} \{\theta_k + \gamma_k' \mathbf{Z}_i\} - \log \left[\sum_l \exp\{\theta_l + \gamma_l' \mathbf{Z}_i\} \right] \\
&\quad + \sum_i \sum_j d_{ij} \mathbf{X}_{ij} \boldsymbol{\alpha} - \log(1 + e^{\mathbf{X}_{ij}\boldsymbol{\alpha}}) \\
&\quad + \sum_i \sum_k c_{ik} \sum_j (1 - d_{ij})(Y_{ij} \mathbf{T}_{ij} \beta_k - e^{\mathbf{T}_{ij}\beta_k}) \\
&= L(\boldsymbol{\theta}, \gamma) + L(\boldsymbol{\alpha}) + \sum_k L(\beta_k). \qquad (13)
\end{aligned}$$

This function is easy to maximize, because it splits into $K + 2$ independent terms. To estimate $(\boldsymbol{\theta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, iteratively maximize (13) and update d_{ij} and c_{ik} with current parameter estimates.

The maximization over $(\boldsymbol{\theta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ can be implemented using GLM functions with extensions of the methods described by Lambert (1992). For instance, when $K = 2$, a standard weighted logistic regression analysis can be performed to maximize $L(\boldsymbol{\theta}, \gamma)$. This can be achieved by essentially creating two copies of the data and then using the posterior probabilities as weights to obtain a likelihood identical to $L(\boldsymbol{\theta}, \gamma)$.

When $K > 2$, a series of separate simple logistic regression functions can be fit as a replacement for polychotomous logistic regression. Each level, $k = 2, \ldots, K$, is compared against the baseline category, $k = 1$. In a standard polychotomous regression setting, this approximation yields highly efficient estimators (Begg and Gray 1984).

For each simple logistic regression (group k vs. 1), we augment the data with n extra "data values." The first n copies have weights equal to $(c_{1k}, c_{2k}, \ldots, c_{nk})$, whereas the second set have weights $(c_{11}, c_{21}, \ldots, c_{n1})$. The augmented covariates are $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, and the augmented binary response vector is n 1s, followed by n 0s. With the augmented dataset, one can estimate (θ_k, γ_k) using weighted logistic regression.

When estimating $\boldsymbol{\alpha}$, suppose that n_0 is the number of observations with $Y_{ij} = 0$. Say $Y_{ij_1}, \ldots, Y_{ij_{n_0}}$ are 0. Augmenting the data with n_0 extra "data values" and using weighted logistic regression leads to a likelihood equivalent to the likelihood that we wish to maximize (Lambert 1992). For this model, the augmented binary response vector has $n \times J$ 0's followed by n_0 1's. The augmented covariates are the original $n \times J$ values, $(\mathbf{T}_{11}, \ldots, \mathbf{T}_{nJ})$, plus the covariates associated with $Y_{ij} = 0$, $(\mathbf{T}_{ij_1}, \ldots, \mathbf{T}_{ij_{n_0}})$. The appropriate weights are $(1 - d_{11}, \ldots, 1 - d_{nJ}, d_{ij_1}, \ldots, d_{ij_{n_0}})$.

For a ZIP model, β can be estimated using weighted log-linear Poisson regression with weights $(1 - d_{ij})$ (Lambert 1992). For an MZIP model, K independent runs, with weights $c_{ik} \times (1 - d_{ij})$, are required to estimate β_k, $k = 1, \ldots, K$.

4.2 An Approximation

Recall that we are primarily interested in making inferences about the relationship between C and **Z**, which is parameterized by $(\boldsymbol{\theta}, \gamma)$. However, the full likelihood depends on many parameters, namely $\Psi = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. A subset of these parameters, $\Phi = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, can be viewed as nuisance parameters. Here we develop an approximation that involves estimating Φ based on the marginal likelihood (6). The approximation allows models for $C|\mathbf{Z}$ to be fit using standard software for GLM's. We emphasize that exact maximization, as described in Section 4.1, is not difficult but is tedious if a very large number of models for $C|\mathbf{Z}$ are entertained in the model selection process. For instance, with the approximation described herein, hundreds of models can be compared in a manner of minutes with little user intervention. Alternatively, fitting a single model using a SAS procedure that maximizes the full likelihood takes 10–15 minutes for a dataset of size comparable to the Cambridge dataset. In addition, the full maximization requires the user to specify starting values for each model considered.

The marginal likelihood can be maximized using the techniques described in Section 4.1 by simply replacing $p_k(\mathbf{z})$ by p_k in the definition of c_{ik} and d_{ij}; let c_{ik}^m denote the resulting marginal posterior probability of membership, $\Pr(C_i = k|\mathbf{y}_i)$. As is usual for mixture problems, p_k is maximized by $1/n \sum_i c_{ik}^m$. Call this marginal estimator $\hat{\Phi}_m$. Next, we obtain an estimate of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ by plugging $\hat{\Phi}_m$ into (13) and using the EM algorithm; call this approximate maximum likelihood estimator (MLE) $(\hat{\boldsymbol{\theta}}_a, \hat{\boldsymbol{\gamma}}_a)$.

Because only the first term in (13) depends on $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, the approximation involves iteratively computing c_{ik} from (9) and maximizing $L(\boldsymbol{\theta}, \boldsymbol{\gamma})$ to update $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. In the first cycle, one could use c_{ik}^m for c_{ik}. This approximation falls short of a full maximization, because Φ is not updated as the EM iterates. In our analysis we have found that this approximation, even without iterations of the EM algorithm, yields an estimate of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ suitable for exploratory data analysis. Thus models for $C|\mathbf{Z}$ can be fit with no more effort than in a standard GLM setting.

From the marginal maximization, we obtain a root-n consistent estimator of Φ under the usual regularity conditions for mixture models (e.g., Redner and Walker 1984, thm. 3.1). In general this estimator is not as efficient as the full MLE. The approximate estimator of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is also not fully efficient, but for fixed K it is root-n consistent, under some regularity conditions. Proof of this theorem follows from the so-called plug-in theorem (e.g., van der Vaart 1998, thm. 5.31).

Because it is difficult to determine a priori when these approximations are sufficiently accurate for a particular purpose, we recommend using the approximation in the preliminary analyses when many models are being compared. Then, when this number has been winnowed down to a modest size, we recommend computing the full MLE's for the final analysis.

5. RISK FACTOR ANALYSIS

The objective of our analysis is to identify risk factors for the various criminal careers. This is statistically equivalent to determining whether γ differs across latent groups. If $\mathbf{Z}$ is categorical and D-dimensional, then the data can be imagined as generating an unobservable contingency table that is $(D+1)$ dimensional. We wish to determine which covariates are associated with particular latent classes. If latent classes were observable, then evaluating hypotheses of this sort would be simple even if $\mathbf{Z}$ included continuous covariates. The appropriate analysis would be either a contingency table analysis or a (polychotomous) logistic regression.

In this section we give a conservative approximation for testing hypotheses about the relationship between $\mathbf{Z}$ and C. The approximation can also be used as the basis of a model selection procedure such as the Bayesian information criterion (BIC; Schwarz 1978). The approximation is nearly as simple to compute as fitting a GLM, and the computations of the test statistic follow naturally from the parameter estimation phase using standard software. The approximation also illustrates the problems inherent in the classify–analyze approach.

To test a hypothesis with the classify-analyze approach, one classifies an individual to the group with the largest posterior probability, $\Pr(C_i = k|\mathbf{Y}_i)$, then performs the risk factor analysis as though C were known. The likelihood for this model is based on the conditional density of C given $\mathbf{Z}$, $\Pr(C_i|\mathbf{Z}_i) = \prod_k p_k(\mathbf{Z}_i)^{I\{C_i=k\}}$.

Assuming that C_i is known, the log-likelihood can be written as

$$l(\boldsymbol{\theta}, \gamma; c|z) = \sum_i \sum_k I\{C_i = k\}\,(\theta_k + \gamma_k'\mathbf{Z}_i) - \log\left(\sum_l \exp\{\theta_l + \gamma_l'\mathbf{Z}_i\}\right). \quad (14)$$

The problem with this analysis is that C is not actually observed. Treating it as if it were known can cause biases in both the estimate of $(\boldsymbol{\theta}, \gamma)$ and the test statistic.

To account for uncertainty in group membership, a likelihood approach based on the mixture model can be developed. The test of a nested hypothesis H_0: $\gamma \in \Omega_0$ versus H_1: $\gamma \in \Omega_1$ can be conducted using a likelihood ratio test:

$$\Lambda = 2[\max_{\boldsymbol{\theta}, \boldsymbol{\gamma} \in \Omega_1, \Phi} \log \mathcal{L}(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta}, \gamma, \Phi) - \max_{\boldsymbol{\theta}, \boldsymbol{\gamma} \in \Omega_0, \Phi} \log \mathcal{L}(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta}, \gamma, \Phi)]. \quad (15)$$

As the mixture likelihood is not regular, it is unclear whether Λ is approximately chi-squared distributed under the null hypothesis. When testing $K = p$ versus $p+1$, classical asymptotic results are known to fail for two reasons: The null hypothesis is on the boundary of the parameter space and, due to a certain lack of identifiability, it is not clear how to count degrees of freedom (Ghosh and Sen 1985). But for the problem at hand, the null hypothesis is not on the boundary of the parameter space, and the difference in dimension of competing models is clearly defined. Thus it seems likely that the traditional chi squared approximation will apply.

Often the objective is to choose among several competing models. In our application we use the BIC to select a model, regarding it as an approximation to the Bayes factors for the competing models. (Kass and Raftery 1995 is a good reference for Bayes factors. In particular, their section

3 provides definitions of, and suggestions on how to interpret Bayes factors, their section 4 discusses use of the BIC as an easily computed approximation, and their section 8 compares the use of Bayes factors with more standard non-Bayesian approaches to model selection.) Raftery (1995) has pointed out several practical difficulties commonly encontered when using the more standard p value–based approach to model selection, and showed how using the BIC provides a better approach.

One advantage of the BIC over traditional hypothesis testing is that it has good properties under weaker regularity conditions than the likelihood ratio test. For instance, Keribin (1998) demonstrated that under certain conditions, the BIC consistently determines the right number of components in the mixture model. Leroux (1992) and Roeder and Wasserman (1997) provided additional theoretical justification for the BIC in the mixture setting. Moreover, the BIC is consistent even when the models are not nested (Nishii 1988).

One important aspect of the work of Kass and Raftery (1995) and Raftery (1995) that we do not consider is the use of model averaging to account for model uncertainty. In our application we have two model selection problems; in Section 6.1 we choose a model for the number of latent classes, and in Section 6.2 we choose among many models describing the effects of two risk factors on class membership. In each case we find that a single model dominates all others, so that model uncertainty is not a problem for us.

5.1 Another Approximation

Some insights into (14) can be gained by the following representation. Recall that

$$\log\frac{f(\mathbf{y},\mathbf{z};\Psi_1)}{f(\mathbf{y},\mathbf{z};\Psi_0)} = \log\frac{f(\mathbf{y},\mathbf{z},c;\Psi_1)}{f(\mathbf{y},\mathbf{z},c;\Psi_0)} - \log\frac{f(c|\mathbf{y},\mathbf{z};\Psi_1)}{f(c|\mathbf{y},\mathbf{z};\Psi_0)}. \quad (16)$$

Taking the conditional expectation, with respect to the observables $(\mathbf{Y},\mathbf{Z})$ evaluated at Ψ_0, yields

$$R(\Psi_1,\Psi_0) = Q(\Psi_1,\Psi_0) - P(\Psi_1,\Psi_0), \quad (17)$$

where

$$R(\Psi_1,\Psi_0) = \log\frac{f(\mathbf{y},\mathbf{z};\Psi_1)}{f(\mathbf{y},\mathbf{z};\Psi_0)},$$

$$Q(\Psi_1,\Psi_0) = E\left[\log\frac{f(\mathbf{y},\mathbf{z},c;\Psi_1)}{f(\mathbf{y},\mathbf{z},c;\Psi_0)}\middle|\mathbf{y},\mathbf{z}\right],$$

and

$$P(\Psi_1,\Psi_0) = E\left[\log\frac{f(c|\mathbf{y},\mathbf{z};\Psi_1)}{f(c|\mathbf{y},\mathbf{z};\Psi_0}\middle|\mathbf{y},\mathbf{z}\right].$$

We define $\Psi_1 = \operatorname{argmax}_{\boldsymbol{\theta},\gamma\in\Omega_1,\Phi}\mathcal{L}(\mathbf{y},\mathbf{z};\boldsymbol{\theta},\gamma,\Phi)$ and $\Psi_0 = \operatorname{argmax}_{\boldsymbol{\theta},\gamma\in\Omega_0,\Phi}\mathcal{L}(\mathbf{y},\mathbf{z};\boldsymbol{\theta},\gamma,\Phi)$. With a standard likelihood ratio test, the simpler hypothesis $\gamma\in\Omega_0$ is rejected if $2R(\Psi_1,\Psi_0)$ is greater than a preselected chi-squared value with degrees of freedom equal to the difference in dimension of the hypotheses.

By Jensen's inequality, $P(\Psi_1,\Psi_0)\le 0$, so a conservative hypothesis test can be based on $2Q(\Psi_1,\Psi_0)$. The advantage of this approximation is that $Q(\Psi_1,\Psi_0)$ is computed as a byproduct of the EM algorithm; see (13). We conjecture that little is lost by approximating $P(\Psi_1,\Psi_0)$ by 0, provided that very little additional information about latent class membership is obtained from $\mathbf{Z}$, given that $\mathbf{Y}$ was already observed. This conjecture is motivated by the following argument. Let $\hat{\Psi}_m$ denote the MLE for Ψ when γ is constrained to be 0; this null model is exactly equivalent to the marginal model. Let $\hat{\Psi}_f$ be the MLE for the richest model for $C|\mathbf{Z}$ to be considered. Furthermore, define $c_{ik}^m = f(c=k|\mathbf{y},\mathbf{z};\hat{\Psi}_m)$ and $c_{ik}^f = f(c=k|\mathbf{y},\mathbf{z};\hat{\Psi}_f)$ accordingly. Presumably, $P(\Psi_1,\Psi_0)$ is bounded by

$$\sum_i\sum_k c_{ik}^m\log\{c_{ik}^f/c_{ik}^m\}. \quad (18)$$

This term is small relative to $Q(\Psi_1,\Psi_0)$ if the posterior probability of membership is essentially unchanged when we learn about $\mathbf{Z}$.

An even more convenient approximate hypothesis test emerges if the MLE for Φ is essentially equivalent whether computed using the full likelihood (7) or the marginal likelihood (6). This equivalence also occurs if (18) is small relative to $Q(\Psi_1,\Psi_0)$. Under this assumption, $\hat{\Phi}_1\approx\hat{\Phi}_0\approx\hat{\Phi}_m$. Hence the test statistic $2\,Q(\Psi_1,\Psi_0)$ is approximately equal to

$$2\{\max_{\boldsymbol{\theta},\gamma\in\Omega_1} L(\boldsymbol{\theta},\gamma) - \max_{\boldsymbol{\theta},\gamma\in\Omega_0} L(\boldsymbol{\theta},\gamma)\}, \quad (19)$$

where $L(\boldsymbol{\theta},\gamma)$ is defined in (13). But this is just the classify–analyze test [see (13)] with $I\{C_i=k\}$ replaced by the posterior probability of group membership. From this approximation, the nature of the error in the classify–analyze approach is made clear. The classify–analyze approach exaggerates the certainty of group membership, which tends to inflate the precision of the estimated risk factors, $\hat{\gamma}$ (see Sect. 6).

These approximations also apply in the model selection context. The BIC is defined as $\log\mathcal{L} - p/2\log n$, where p is the dimension of the model and n is the number of subjects. From the foregoing arguments, it follows that an approximate procedure can be based on replacing $\log\mathcal{L}$ with $L(\boldsymbol{\theta},\gamma)$ and picking the model with the largest approximate BIC.

To compute $L(\boldsymbol{\theta},\gamma)$, substitute parameter estimates obtained as described in Section 4.1, or the approximate estimates obtained as described in Section 4.2, into (13). We found that the approximate estimates were sufficient for the model screening phase (see Sect. 6).

6. APPLICATION

In this section we apply the methods just described to data from the Cambridge Study of Delinquent Development. Of particular interest is whether the life-course–persistent and adolescent-limited trajectories predicted by

Table 1. BIC Values for a Selection of MZIP Models

Model	K	Order	ΔBIC	Probability
1	2	2,2	9.80	*
2	3	0,2,2	35.80	.014
3	3	2,2,2	30.80	*
4	4	0,2,0,2	40.02	.966
5	4	2,2,1,2	36.14	.020
6	4	0,2,2,2	0	*

NOTE: Order indicates whether the trajectory was fit with a constant (0), linear (1), or quadratic (2) function. In each case, the trajectories are ordered from least to most number of crimes committed as an adult. Define ΔBIC as $\text{BIC}_j - \text{BIC}_6$. The intermittency parameter was fit with a constant function for all of the models given here; any other form resulted in an inferior BIC score.

* Probability less than 10^{-3}.

the Moffitt theory are present in the data, and whether the analysis supports Moffitt's predictions about the distinctive etiology of the former group. Specifically, we test for an interaction between symptoms of neurological deficits and poor child-rearing practice in heightening the probability of following a trajectory of chronic offending.

To address the research question, we used a model selection procedure. This involved estimating the number of latent classes, the order of the polynomial for each latent trajectory, and the covarariates to include in the model. We approached this question in two stages. First, using the marginal likelihoood, we determined K and the form of the trajectories. Then, conditional on this model, we considered the covariate models.

6.1 Estimating the Number of Latent Classes

In their analysis of the Cambridge data, Nagin and Land (1993) proposed a four-group model. They fitted each group with a trajectory quadratic in time, except for the group that had a very low constant probability of offending. For the intermittency parameter, they also used a quadratic function of time.

To determine the validity of this model, we conducted a model search of all possible models within the class for $K \leq 5$; for $K = 5$, the best model reduced to a four-group model. The order of the model was selected for each trajectory and for the intermittency parameter (0 = constant, 1 = linear, 2 = quadratic); this constitutes a total of 3^{K+1} models, for $K = 1, \ldots, 5$. This analysis was based on the marginal likelihood for the offending patterns. Table 1 gives the BIC values for the best-fitting models, including the best model for each of $K = 2, 3, 4$. Using the criterion of Jeffreys (1961), as presented by Kass and Raftery (1995), there is strong evidence for a four-group model. Furthermore, there is strong evidence that model 4 is the best four-group model considered. The posterior probability that a model is correct is approximated by $\exp\{\text{BIC}_j\}/\sum_j \exp\{\text{BIC}_j\}$ (Kass and Wasserman 1995; Schwarz 1978). Model 4 clearly dominates the others by this criterion, and hence we do not take a model averaging approach.

Our BIC analysis supports a four-group model much like the one fit by Nagin and Land (1993) (Fig. 1). The fitted model captures essentially the same features in the data as the model proposed by Nagin and Land, but the model

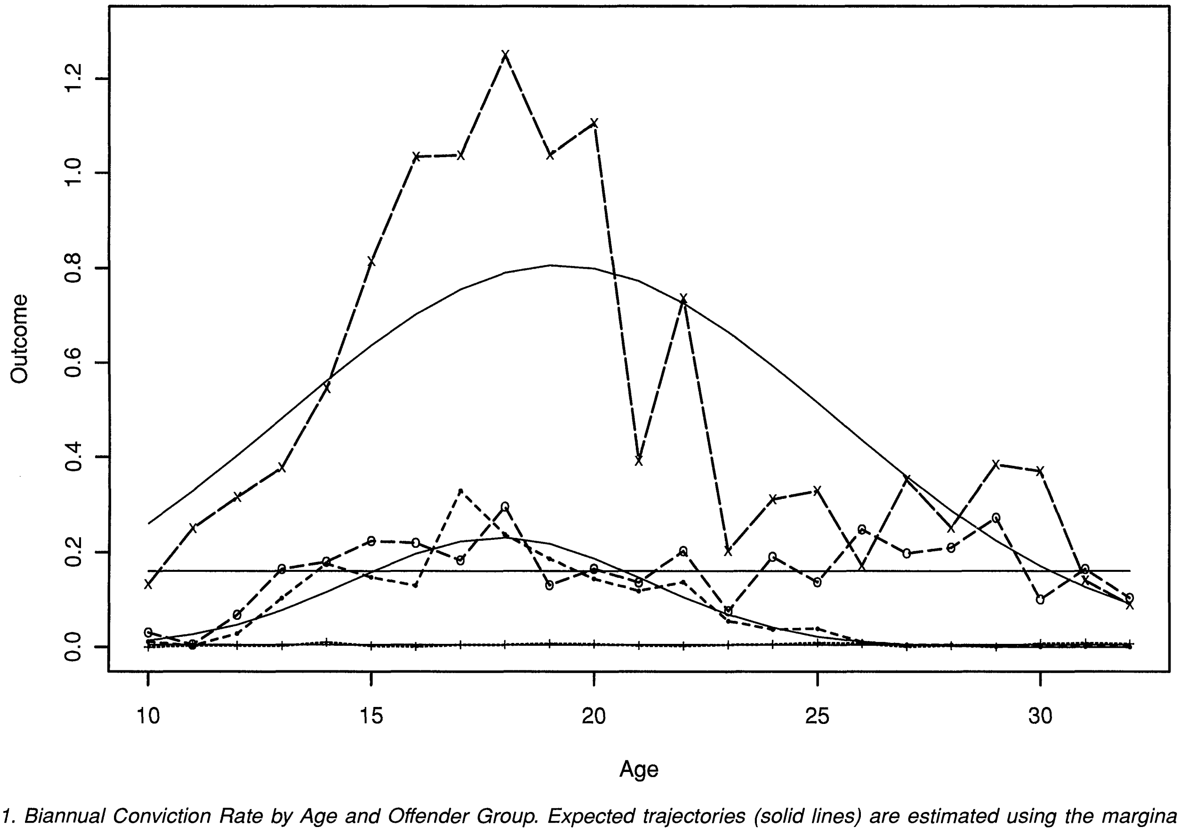

Figure 1. Biannual Conviction Rate by Age and Offender Group. Expected trajectories (solid lines) are estimated using the marginal likelihood mixture model. Observed trajectories are obtained as a weighted average over all of the observations (weights = c_{ik}). Trajectories are labeled as never-convicted (NC; ··· + ···), adolescent-limiteds (AL's; --- • ---), low-level chronics (LLC's; – – o – –), and high-level chronics (HLC's; — x —).

Table 2. Parameter Estimates for Model 4

Group	Parameter	Estimate	Error	Test	p-value
NC	β_0	−4.85	.32	−15.3	*
AL	θ_2	−1.28	.23	−5.52	*
	β_0	−15.52	3.32	−4.67	*
	β_1	16.24	3.80	4.27	*
	β_2	−4.52	1.07	−4.22	*
LLC	θ_3	−2.16	.30	−7.22	*
	β_0	−1.18	.20	−5.84	*
HLC	θ_4	−2.11	.24	−8.68	*
	β_0	−4.59	.82	−5.64	*
	β_1	5.19	.82	6.32	*
	β_2	−1.36	.20	−6.73	*
All	α_0	−.20	.16	−1.30	.193

NOTE: Trajectories are labeled as never convicted (NC), adolescent limiteds (AL's), low-level chronics (LLC's), and high-level chronics (HLC's).

* Probability less than 10^{-3}.

selected by the BIC is somewhat simpler. Contrary to Nagin and Land's fit, model 4 fits two trajectories and the intermittency parameter with a constant function rather than a quadratic. Nagin and Land labeled the four groups as 1 = never convicted (NC), 2 = adolescent limiteds (AL's), 3 = low-level chronics (LLC's), and 4 = high-level chronics (HLC's). The chronic groups are labeled as such because these individuals continue to commit a low level of crimes even as they enter their thirties. In constrast, the adolescent-limited group essentially terminates criminal activity as they enter their twenties.

Figure 1 clearly shows that the model provides a fairly good fit to the data. This four-group model is generally consistent with Moffitt's theory. It includes the two key groups in her taxonomy: adolescent limited and chronic offenders. Although Moffitt does not specifically predict a nonoffender group, our measure of criminal involvement (official statistics on conviction) is a highly filtered measure of criminality. However, the low-level chronic trajectory is clearly not anticipated by her theory. Thus our findings support Moffitt's main prediction about distinctive developmental trajectories but suggests that a richer taxonomy may be necessary.

Table 2 gives the parameter estimates for the selected model. Notice that the intermittency parameter appears to be insignificant; however, dropping it from the model lowers the BIC value by 44 units, a very significant difference by Jeffreys's (1961) criterion. The NC group is by far the largest (66%), followed by the AL group (18%), but a substantial portion of the population (15%) is considered to be chronic offenders.

For each individual in the study, we computed the maximum posterior membership probabilities ($\operatorname{argmax}_k c_{ik}^m$). The median assignment probabilities for each group are high (Table 3), suggesting that a majority of individuals can be classified to a particular latent class with high probability. It is clear from the positions of the first quartiles of these modal probabilities that the model has little ambiguity when assigning to either the NC or the HLC classes, but somewhat more for the other two classes. Overall, this analysis provided additional support in favor of the four-group model.

Table 3. Distribution of the Maximum Posterior Membership Probabilities, in Percents

	1st Q	Median	3rd Q
NC	94	94	94
AL	67	75	92
LLC	59	77	93
HLC	91	100	100

6.2 Risk Factor Analysis

We turn now to question whether probability of membership in the HLC group is heightened by the interaction of symptoms of neurological deficits and poor child-rearing practice. Some of the hallmarks of the sort of neurological deficits emphasized by Moffitt are impulsivity, inattention, and a propensity to engage in risky behaviors. The Cambridge data includes such a measure, which its principal investigators call daring (Farrington and West 1990). The dataset also includes an index of poor parenting practice that measures such behaviors as lack of supervision, harsh and erratic punishment, and neglect. To test the impact of these covariates on trajectory group membership, we conducted a risk factor analysis.

The analysis investigated the effect of daring, of child-rearing, and of the interaction between these two factors. The models were allowed to differ for each level: baseline = NC versus 2 = AL, 3 = LLC, and 4 = HLC. For each level, we fit 7 models, for a total of 343 models: a = null model, b = daring, c = rearing, d = daring + rearing, e = daring + rearing + daring × rearing, f = daring + daring × rearing, and g = daring × rearing. Models f and g are somewhat unconventional, as they include an interaction without including all main effects. Model g was included because it most closely reflects the hypothesis of Moffitt.

Using the BIC, we selected the model g for each level. By Jeffreys's (1961) criterion, there is very strong support for this model. Table 4 presents the BIC scores for all seven models that have the same set of covariates for each level. None of the models excluded from the table garners appreciable probability using the BIC criterion. Because the chosen model attains a very high posterior probability, we do not use a model-averaging approach in our analyses. Table 5 presents the estimated parameter values. Comparing the parameter estimates obtained for the trajectories in Tables 2 and 5 clearly demonstrates introducing the covari-

Table 4. BIC for the Covariate Models

Model	ΔBIC	Probability	Approximate probability
a	0	*	*
b	13.2	.009	.018
c	3.0	*	*
d	8.5	*	*
e	4.9	*	*
f	13.2	.009	.005
g	17.9	.982	.977

NOTE: In each case the same model was fit for each level. The approximate probability was computed using the BIC obtained from the approximate fits. $\Delta\text{BIC} = \text{BIC}_j - \text{BIC}_1$.

* Probability less than 10^{-3}.

Table 5. Parameter Estimates, K = 4 and Covariate = Daring × Rearing

Group	Parameter	Estimate	Error	Test	p-value
NC	β_0	−4.83	.30	−16.17	*
AL	β_0	−15.58	3.29	−4.73	*
	β_1	16.46	3.76	4.37	*
	β_2	−4.61	1.06	−4.33	*
LLC	β_0	−1.24	.19	−6.59	*
HLC	β_0	−4.54	.82	−5.51	*
	β_1	5.12	.83	6.20	*
	β_2	−1.33	.20	−6.63	*
All	α_0	−.22	.16	−1.40	.162
AL	θ	−1.52	.25	−6.07	*
	γ	2.37	.77	3.08	.002
LLC	θ	−2.31	.32	−7.31	*
	γ	2.41	.88	2.75	.006
HLC	θ	−2.61	.30	−8.85	*
	γ	3.45	.70	4.95	*

NOTE: * Probability less than 10^{-3}.

ates does not change the estimated trajectories in a meaningful way.

The results generally conform to Moffitt's contention that neither neurological deficit nor poor parenting alone is sufficient to instigate the developmental process that results in chronic antisocial behavior. The highly significant interaction term of daring and poor parenting for the probability of the HLC group conforms exactly with her prediction. This interaction is also significant for the AL group, which does not conform to her theory. However, the magnitude of the impact rather than its statistical significance is the critical factor in judging her prediction.

Table 6 reports calculations of the group membership probabilities obtained from the model. The first row of the table reports the marginal probabilities without covariates for the model as reported in Table 2. The second row reports these probabilities for the model with covariates for the case in which at most one of the risk factors, poor parenting or daring, is present. The third row reports probabilities where both risk factors are present. Observe that when at most one risk factor is present, group membership probabilities conform closely to the marginal rates. However, when both risk factors are present, the probabilities change dramatically. The probability of the NC trajectory plunges from .72 to .15. The largest absolute increase is for the HLC group, which increases by a factor of seven, from .05 to .34. The next largest increase is for the LLC group, which increases by a factor of two, from .16 to .35. The probability of the AL group also increases by a factor of two, but in absolute terms the increase is smallest, from .07 to .16. Thus, judged in terms of magnitude of the increases, the results reasonably conform to Moffitt's predictions.

Table 6. Latent Class Membership Probabilities, in Percents, K = 4

	NC	AL	LLC	HLC
Marginal	66	18	8	8
Not Both	72 (70)	16 (17)	7 (7)	5 (6)
Both	15 (23)	35 (29)	16 (16)	34 (32)

NOTE: Marginal probabilities were obtained based on the no covariate model. The other probabilities were computed using the covariate model that included only the interaction term (daring × rearing). Not both and Both indicate the number of risk factors present. The probabilities in parentheses were obtained using the approximation.

Table 7. Parameter Estimates Obtained Using the Approximation, K = 4

Model	θ	γ
AL	−1.40	1.60
LLC	−2.33	1.93
HLC	−2.52	2.81

6.3 Analysis of Approximations

To search the large space of models, we implemented the approximate model-searching process described in Sections 4 and 5. Finally, for a subset of those models, we fit the full-likelihood model. From Table 4 we can also see that the approximation identifies the same top models as those found by the exact procedure.

Table 7 presents the estimated parameter estimates obtained via the approximation. These parameter estimates are attenuated relative to the MLE's. In fact, a slight attenuation was observed for all the models investigated. Nevertheless, the approximation proved to be an effective vehicle for identifying the best models in the class. Moreover, the latent class membership probabilities computed using the approximation do not vary substantially from those computed using the MLE's (Table 6).

Table 8 gives the parameter estimates and their associated standard errors obtained using the classify–analyze approach. Although the parameter estimates are only slightly attenuated relative to the MLE's, the estimated standard deviations are clearly underestimated even in this example, where the maximum posterior probabilities are quite high.

7. SUMMARY AND CONCLUSIONS

In this analysis we have tested a prominent theory in psychology about the development of adolescent-limited versus life-course–persistent antisocial behavior. Our findings generally conform to the predictions of the theory. Specifically, we found evidence of both types of trajectories of criminal offending anticipated by the theory. We also found that the probability of following the high-level chronic trajectory was dramatically increased for individuals who displayed evidence of neurological deficits and who were subject to

Table 8. Parameter Estimates Obtained Using the Classify–Analyze Procedure, K = 4

Group	Parameter	Estimate	Standard error
AL	θ	−1.66	.15
	γ	2.20	.50
LLC	θ	−2.44	.21
	γ	2.10	.62
HLC	θ	−2.78	.25
	γ	3.32	.54

poor child-rearing practice. However, the results were not fully consistent with Moffitt's theory. We found evidence of a fourth group not anticipated by her theory: the low-level chronics. We also found the probability of membership in the adolescent-limited group is significantly related to the interaction of poor parenting and daring.

Conventional practice in studies of latent traits is to use the classify–analyze paradigm—assign subjects to the latent category that is most likely and then treat this classification variable as though it were observed without error. This approach can cause errors in the statistical inferences. To avoid these errors, we have presented a simple method for accounting for the uncertainty inherent in classifying individuals to latent traits.

The model is designed for a situation where a response variable Y and a covariate or risk factor Z are measured. A latent trait variable C that takes on a finite number of states is assumed to "explain" the association between Y and Z; that is, Y and Z are assumed to be conditionally independent, given C. The likelihood of (Y, Z) is simple to express because of the conditional independence assumption. Given this structure, we illustrate that it is relatively easy to estimate the relationship between C and Z and to perform hypothesis tests about the parameters describing this relationship. We also discuss an approximation that facilitates implementation of the procedure using standard software. We contrast this with the classify–analyze method, which is based on the conditional distribution of C given Z. The classify–analyze approach produces incorrect inferences because it ignores the uncertainty about C. Although Y tells us a good deal about latent class membership, we observe Y, not C, and thus inference should account for residual uncertainty about C. With our approximation it is almost as easy to perform the correct analysis as it is to use the classify–analyze approach. The approximation may not always work, but it is superior to the classify–analyze approach.

Our application focuses on an example from criminology, but taxonomic theorems are commonplace in the social sciences (e.g., categories of organizations, types of personalities), so our method potentially has wide applicability. In contrast to the criminology application, for many such applications the response variable Y is not a count variable or even a longitudinal record. Many of our results still apply, provided that the response variable yields a sufficient amount of information about group membership to provide an estimate of the probability of latent class membership; that is, $\Pr(C = k|Y)$. Our analysis shows that unless there is very little uncertainty about latent class membership, it is preferable to work with the full-mixture likelihood probability than to classify the subject to the latent class with the highest membership probability.

[Received March 1997. Revised January 1999.]

REFERENCES

Begg, C. B., and Gray, R., (1984), "Calculation of Polytomous Logistic Regression Parameters Using Individualized Regressions," *Biometrika*, 71, 11–18.

Clogg, C. C. (1995), "Latent Class Models," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds. G. Arminger, C. C. Clogg, and M. E. Sobel, New York: Plenum.

Dempster, A. P., Laird N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser B, 39, 1–38.

Farrington, D. P. (1986), "Stepping Stones to Adult Criminal Careers," in *Development of Antisocial and Prosocial Behavior*, eds. D. Olweus, J. Block, and M. R. Yarrow, New York: Academic Press.

Farrington, D. P., and West, D. J. (1990), "The Cambridge Study in Delinquent Development: A Prospective Longitudinal Survey of 411 Males," in *Criminality: Personality, Behavior and Life History*, eds. H. Kerner and G. Kaiser, New York: Springer-Verlag.

Ghosh, J. K., and Sen, P. K. (1985), "On the Asymptotic Performance of the Log-Likelihood Ratio Statistic for the Mixture Model and Related Results," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. L. M. Le Cam and R. A. Olshen, Belmont, CA: Wadsworth, pp. 789–806.

Greene, W. H. (1997), *LIMDEP User's Manual*, Bellport, NY: Econometric Software, Inc.

Hawkins, J. D., Lisher, D. M., Catalano, R. F., and Howard, M.O. (1986), "Childhood Predictors of Adolescent Substance Abuse: Toward an Empirically Grounded Theory," *Journal of Children in Contemporary Society*, 8, 11–48.

Hirschi, T., and Gottfredson, M. (1983), "Age and Explanation of Crime," *American Journal of Sociology*, 89, 552–584.

Huizinga, D., Esbensen, F., and Weiher, A. W. (1991), "Are There Multiple Paths to Delinquency?," *Journal of Criminal Law and Criminology*, 82, 83–118.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), London: Oxford University Press.

Jones, R., Nagin, D. S., and Roeder, K. (in press), "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories," *Sociological Methods and Research.*

Kass, R., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Keribin, C. (1998), "Consistent Estimation of the Order of Mixture Models," Working paper 61, Université d'Evry-Val d'Essonne, Laboratorie Anlyse et Probabilité.

Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–13.

Land, K., McCall, P., and Nagin, D. S. (1996), "A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models With Empirical Application to Criminal Careers Data," *Sociological Methods and Research*, 24, 163–191.

Langeheine, R., and Rost, J. (eds.) (1988), *Latent Trait and Latent Class Models*, New York: Plenum.

Leroux, B. G. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360.

Lindsay B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, Hayward, CA: The Institute of Mathematical Statistics.

Loeber, R., and LeBlanc, M. (1990), "Toward a Developmental Criminology," in *Crime and Justice: An Annual Review of Research*, 12, eds. M. Tonry and N. Morris, Chicago, IL: University of Chicago Press.

Loeber, R., and Stouthamer-Loeber, M. (1986), "Family Factors as Correlates and Predictors of Juvenile Conduct Problems and Deliquency," in *Crime and Justice*, 7, eds. M. Tonry and N. Morris, Chicago, IL: University of Chicago Press.

Moffitt, T. E. (1993), "Adolescence-Limited and Life-Course–Persistent Antisocial Behavior: A Developmental Taxonomy," *Psychological Review*, 100, 674–701.

Muthen, B., and Shedden, K. (1999), "Finite Mixture Modeling With Mixture Outcomes Using the EM Algorithm," unpublished manuscript submitted to *Biometrics*.

Nagin, D. S., Farrington, D. P., and Moffitt, T. E. (1995), "Life-Course Trajectories of Different Types of Offenders," *Criminology*, 33, 111–139.

Nagin, D. S., and Land, K. C. (1993), "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model," *Criminology*, 31, 327-362.

Nagin, D. S., and Temblay, R. E. (1999), "Trajectories of Boys' Physical

Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency," unpublished manuscript submitted to *Child Development*.

Nishii, R. (1988), "Maximum Likelihood Principle and Model Selection When the True Model is Unspecified," *Journal of Multivariate Analysis*, 27, 392–403.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research" (with discussion), in *Sociological Methodology* ed. P. V. Marsden, Cambridge, MA: Blackwell.

Redner, R., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239 .

Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.

Rost, J., and Langeheine, R. (eds.) (1997), *Applications of Latent Trait and Latent Class Models in the Social Sciences*, New York: Waxmann.

Sampson, R. J., and Laub, J. H. (1991), "Crime and Deviance Over the Life Course: The Salience of Adult Social Bonds," *American Sociological Review*, 55, 608–627.

——— (1993), *Crime in Making: Pathways and Turning Points Through the Life Course*, Cambridge, MA: Harvard University Press.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambrige, U.K.: University Press.

Wilson, J. Q., and Herrnstein, R. (1985), *Crime and Human Nature*, New York: Simon and Shuster.