Published in final edited form as:

Psychometrika. 2008 December ; 73(4): 625–646. doi:10.1007/s11336-008-9077-9.

Regularized finite mixture models for probability trajectories

Kerby Shedden

Department of Statistics University of Michigan

Robert A. Zucker

Departments of Psychiatry and Psychology University of Michigan

Abstract

Finite mixture models are widely used in the analysis of growth trajectory data to discover subgroups of individuals exhibiting similar patterns of behavior over time. In practice, trajectories are usually modeled as polynomials, which may fail to capture important features of the longitudinal pattern. Focusing on dichotomous response measures, we propose a likelihood penalization approach for parameter estimation that is able to capture a variety of nonlinear class mean trajectory shapes with higher precision than maximum likelihood estimates. We show how parameter estimation and inference for whether trajectories are time-invariant, linear time-varying, or nonlinear time-varying can be carried out for such models. To illustrate the method, we use simulation studies and data from a long-term longitudinal study of children at high risk for substance abuse.

Keywords

Finite mixture models; growth mixture models; latent class analysis; functional data analysis; penalized likelihood; longitudinal data analysis; trajectory analysis

Introduction

In finite mixture modeling, or Latent Class Analysis (LCA) (e.g. Titterton et al. 1985, Yung 1997), individuals are modeled as belonging to disjoint groups, where an individual's unobserved group membership determines the statistical properties of his or her observable traits. For finite mixture analysis with trajectories (e.g. Verbeke & Lesaffre 1996, Muthén & Shedden 1998, Nagin 1999, 2001, 2005), the traits are observed longitudinally, and the pattern of an individual's trajectory determines his or her class membership. In trajectory analysis that does not use covariates, finite mixture models provide a convenient method for moving beyond a description of the overall population to something approaching a more individualized analysis.

A key challenge in carrying out LCA with trajectory data is parameterizing the time-varying response structure within each latent class. Convenient alternatives include polynomials, splines, and change-point parameterizations. However class mean trajectories could plausibly show a wide range of other patterns, including patterns with two or more peaks, skewed convex or concave patterns, monotone growth or decay approaching an asymptote, multiple steps, or response limited to a narrow time interval. No convenient and parsimonious parametric family encompasses all of these patterns.

Corresponding author, to whom requests for reprints should be sent: Kerby Shedden, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107. e-mail: kshedden@umich.edu, Phone: 734-764-0438, FAX: 734-763-4676.. Contact information for other author: Robert A. Zucker. Departments of Psychiatry and Psychology. University of Michigan, Ann Arbor, MI 48109-0740. e-mail: zuckerra@umich.edu, Phone: 734-232-0280, FAX: 734-998-7992..

One way to allow flexible modeling of class means trajectories is to use a large number of parameters to describe the population structure. In longitudinal behavioral studies, measurements can be viewed as occurring on a fine grid of time points (e.g. monthly or yearly). If the number of study subjects is large, growth mixture models may be parameterized by allotting a free parameter to each possible measurement time point within each latent class (a “saturated” parameterization). However in most studies, each individual is generally measured at a subset of the possible measurement points, and the number of study subjects may be small or moderate. If a saturated parameterization is used, sampling fluctuations are likely to dominate the trajectory estimates due to the limited information available for estimating each parameter. Such fluctuating trajectories are usually implausible as growth trajectories and are difficult to interpret. Furthermore, imprecision in the trajectory parameter estimates may substantially degrade the power of model selection and other inference procedures.

Here we investigate an approach for increasing the precision of growth mixture model trajectory estimates, while maintaining flexibility in the range of possible growth trajectory shapes. The method begins with a saturated model. Sampling variation in the parameter estimates is reduced through a likelihood penalty that favors class mean trajectories varying smoothly in time. This reduces the variability of parameter estimates and ensures that the fitted trajectories are interpretable. The degree of penalization is determined in a data-driven fashion that maximizes the precision of the estimated class mean trajectories relative to the population values.

We evaluate the proposed strategy using simulation studies, and illustrate its performance using data from a long term longitudinal behavioral study with a relatively sparse three-year sampling frequency. Our major focus for inference will be the temporal pattern of mean responses within each class. Specifically, we aim to reliably distinguish time-varying from constant trajectories, and nonlinear trajectories from linear trajectories. Estimation of the number of latent classes will also be considered. We will use information-based criteria similar to the Akaike Information Criteria (AIC, Akaike 1974) for carrying out these inferences. A mathematical extension of the AIC to the regularized estimation framework is provided.

Model formulation and estimation

We focus here on trajectories of dichotomous behavioral indicators measured at a sequence of points in time. However much of the methodology would be equally applicable to Gaussian or other parametric models for continuous responses. In the data used to illustrate the method, individuals are assessed longitudinally on a particular behavioral item, where at assessment time t the item is rated as “not true” or “sometimes/always true.” Denote by Y_{it} the response for individual i at measurement time point $t \in \{1, \dots, T\}$, with $Y_{it} = 1$ representing an affirmative response and $Y_{it} = 0$ representing a negative response. Our method and software also allow values to be missing at random, as discussed further below, but for simplicity we do not notate that here.

For a given individual or group of individuals, the responses at time t can be viewed as arising from a distribution defined by “symptom endorsement probabilities” (SEP’s) p_t , where p_t is the probability that the statement is rated “sometimes/always true” and $1 - p_t$ is the probability that the statement is rated “not true.” Formulated as a latent class model, the probabilities p_t are viewed as class-specific, with all individuals belonging to class k sharing a common probability trajectory denoted p_{tk} .

Let Q_k denote the marginal probability of membership in class k , where there are K classes in all. We will work with the model having the log-likelihood

$$\sum_i \log \left(\sum_k Q_k \prod_t p_{ik}^{Y_{it}} (1 - p_{ik})^{(1-Y_{it})} \right), \quad (1)$$

which has been widely used for finite mixture modeling of multivariate binary responses (e.g. Fergusson et al. 2000). We note that the model assumes conditional independence of an individual's measurements at different time points given the class membership. Conditioned on the class membership, the responses have binomial distributions with success probabilities depending on both class and assessment time. If an individual is only observed at a subset of the possible measurement points, the product over t in (1) can be restricted to the observed data. This gives meaningful parameter estimates if the data are "missing at random" (Little and Rubin, 1987), which is appropriate, for example, if the missing data are due to the assessment schedules specified by the study design. In principle, every individual can be assessed at completely unique times, and the regularization to be discussed below will control for the very large number of model parameters that would result. In practice, it is usually sufficient to round assessment dates to a grid of time points (e.g. monthly or yearly).

Sampling variance in the estimates of p_{tk} is reduced through a likelihood penalty that draws the probability trajectory estimates toward smooth functions of t for each k . Such penalties have been widely used in regression analysis involving curves and surfaces (Rice and Silverman 1992, Ramsay and Silverman 2005) to reduce the variance of parameter estimates without requiring pre-specification of a parametric form such as a quadratic or linear temporal pattern. Although the resulting parameter estimates are biased, they generally have substantially lower mean squared error (MSE) than maximum likelihood estimates.

The quadratic smoothness penalty

$$\lambda \sum_k \sum_{t=1}^{T-2} (p_{t+2,k} - 2p_{t+1,k} + p_{t,k})^2 \quad (2)$$

is the key component of the modeling approach described in this paper. It is a commonly-used smoothing penalty for serially-observed data on an equally-spaced grid (a simple generalization allows the observations to be unequally spaced). One widely-cited application of (2) is in estimating the covariance structures of curves (Rice and Silverman, 1991).

If we view the trajectory for class k , denoted $p_{t,k}$, as a continuous function of time $f_k(t)$, then (2) is a discretized version of the square-integrated second derivative $\lambda \int f_k(t)''^2 dt$. This penalty vanishes if $f_k(t)$ is linear, that is, if $f_k(t) = a_k + b_k t$ for class-specific constants a_k and b_k , as long as the time values t occur on a regularly-spaced grid (straightforward generalizations permit the same interpretation if the time values are not gridded). The effect of incorporating the penalty is to draw the p_{tk} estimates toward linear trajectories compared to the maximum likelihood estimate (MLE), which is given when $\lambda = 0$. The limiting trajectories as $\lambda \rightarrow \infty$ are linear functions of t , and the limiting penalized estimates are the MLE for that class of models. Therefore as λ varies from zero to infinity, we can view the penalized estimates as spanning two MLE's: the MLE for the saturated model with T free parameters per class occurs when $\lambda = 0$, and the MLE for the linearly constrained model with 2 free parameters per class occurs when $\lambda = \infty$.

Figure 1 shows the results with real data (further discussed below) of using the penalized estimation procedure for a two-class problem with 13 time points. In each graph, the estimated

probability trajectories $\hat{p}_{tk}$ for the two classes $k = 1, 2$ are plotted as functions of time t . The estimated marginal class probabilities $\hat{Q}_k$ are shown below each graph. The unpenalized fit ($\lambda = 0$) fluctuates dramatically due to the relatively small sample size in any one class at a particular age. As the value of λ increases, the fitted trajectories tend toward linear functions of time (in this case the linear functions have zero slope, but in general the slope is not affected by the penalty). Below we will argue that the upper-right graph, corresponding to a penalty weight of $\lambda = 10$, is the best fit to the data. This estimate clearly demonstrates the value of a penalization approach over a more conventional approach of parameterizing the functions in a convenient way, say by using polynomials. The pattern for the high-risk subgroup, which shows an interesting elevation of symptoms just prior to the onset of adolescence, would not be well approximated by a low order polynomial.

More generally, the use of additive penalties $R(\theta)$ for regularized maximum likelihood estimation of a parameter θ with likelihood L has been useful in many problems (e.g. Fan and Li 2003). Interestingly, it turns out that the maximizers of the additively penalized likelihood function $L(\theta) - \lambda R(\theta)$ are closely related to the somewhat more intuitive formulation of the estimation problem as maximizing $L(\theta)$ subject to a constraint $R(\theta) \leq c$ for a given constant c . In this case the value of c would correspond to the maximum tolerable roughness in the class trajectories as measured by the square integrated second derivative. While the constrained problem is somewhat more interpretable, it poses both computational and inferential challenges, since the solution will usually lie on the boundary of the constrained parameter space. It also turns out that the solutions to the constrained problem are identical to the solutions to the additively penalized problem for some mapping between λ and c .

While we focus on frequentist estimation here, parallels exist to Bayesian estimation. The penalty (2) can be viewed as the log-density of an improper prior for the p_{tk} parameters. Under this interpretation, the maximum likelihood estimate that we use becomes equivalent to the maximum a posteriori (MAP) estimate in the Bayesian setting (also known as the posterior mode). The reason that the penalty (2) corresponds to an improper prior is that it is invariant under addition of a linear trajectory. That is, the penalty evaluated at $f_k(t)$ is the same as the penalty evaluated at $f_k(t) + a + bt$, for constants a and b . To achieve a proper prior for Bayesian analysis, it would be necessary to include penalization terms for the intercept and trend of each trajectory. A Bayesian approach to mixture models for trajectory data was given in Dunson and Herring (2005).

The selection of λ is important. A practical method for selecting λ will be given below. From the standpoint of interpretation, we think it most natural to view the penalization as selecting from a family of estimators that balance bias (resulting from overly smooth trajectory estimates) and variance (resulting from over-parameterized trajectory estimates). Thus we do not penalize in order to impose an *a priori* view that the trajectories are smooth, but rather to adaptively estimate the trajectory patterns on as fine a scale as permitted by the data. If the sample size is large, little smoothing is necessary, local fluctuations in the trajectory estimates are likely to be real, and they are estimable with high precision. On the other hand, a fluctuating trajectory estimate based on a small data set is likely to reflect non-reproducible sampling variation. In this case a smoother estimate should be used to reduce variance, even at the risk of increased bias.

Boundary penalties

Irregular parameter estimates in maximum likelihood or penalized maximum likelihood estimation occur when the parameter estimates lie on the boundary of the parameter space. Latent class models with probability trajectories are prone to two different types of irregular parameter estimates. We will use additional likelihood penalties, beyond the smoothing

penalties discussed above, to ensure that the estimates lie in the interior of the parameter space. These penalties are constructed to be approximately flat away from the boundary. Therefore they should have little effect on the positions of interior modes, but rather serve to direct the classifier toward interior rather than boundary modes of the penalized likelihood surface.

One type of irregular estimate occurs when an estimated p_{tk} probability is equal to zero or one. The likelihood penalty

$$c \sum_{tk} \log(p_{tk}(1 - p_{tk})), \quad (3)$$

where c is a constant, prevents the estimate from taking on this irregular solution. When the solution without (3) converges to a point not on the boundary, (3) has little effect since it is very flat away from the boundary. For example, when the sample size is large and the marginal item response frequencies are not close to zero or one, this term typically has little effect. Moreover, when the smoothing penalty (2) is included, it tends to pull the p_{tk} values away from the boundaries unless a number of consecutive time points have extreme response frequencies. For these reasons, we have found that the value of c does not strongly affect our results. We use the value $c = 0.1$ for all results reported here, although some more data-adaptive ways of selecting c could be proposed, such as tying c to the marginal response frequencies.

For many data sets, the likelihood surface of the finite mixture has modes in which one of the class probabilities vanishes. Since inference for the number of latent classes will be handled by explicit comparison of models having different numbers of classes, as detailed below, it is reasonable to restrict estimation to solutions in which all classes have positive marginal probability. To achieve this, the likelihood penalty

$$c \sum_k \log(Q_k(1 - Q_k)), \quad (4)$$

is used, where Q_k is the marginal probability for class k . As with the penalty (3), this term is approximately flat in the interior of the parameter space, so has little effect unless the solution is converging to a point on or near the boundary. The inclusion of (4) keeps the optimizer from converging to a $k-1$ class solution when a k class solution is sought. Since we always consider a family of fits for between 1 and K classes, nothing is lost by imposing this constraint, except in the unlikely situation that there is no proper k class solution without the penalty (a phenomenon that we have not observed in practice). Again we use $c = 0.1$ for all results reported here, and we find that our results are relatively insensitive to this choice since our simulated and actual data did not exhibit classes with very small marginal probabilities.

Boundary penalties such as are discussed here have previously appeared in the literature. Chen, Chen and Kalbfleisch (2004), and Galindo-Garre and Vermunt (2006), among others, have previously used this approach in finite mixture modeling. A related idea also is used in the Latent GOLD software (Vermunt and Magidson, 2004).

Computation

Penalized maximum likelihood parameter estimates are calculated using a conjugate gradient optimization applied to the penalized likelihood

$$\sum_i \log \left(\sum_k Q_k \prod_t p_{tk}^{Y_{it}} (1 - p_{tk})^{(1-Y_{it})} \right) - \lambda \sum_k \sum_{t \geq 2} (p_{t+2,k} - 2p_{t+1,k} + p_{tk})^2 + c \sum_{tk} \log(p_{tk} (1 - p_{tk})) + c \sum_k \log(Q_k (1 - Q_k)) \quad (5)$$

until the gradient norm reaches a defined point (usually 0.1). As noted above, the likelihood is typically multimodal (although the penalty terms make it more concave than it would otherwise be). Therefore multiple runs could be used to attempt to identify the highest mode. The optimization takes 2-5 minutes to converge for the data sets used in the simulation studies and real data analyses below. Software written in C to estimate the models is available from the first author.

Inference

We are mainly interested in inference about the shape of class mean trajectories, specifically whether the trajectories are constant, linear, or nonlinear. A prerequisite to addressing these issues is inference for the number of latent classes. This issue has been widely discussed in the literature (e.g. McLachlan and Peel 2000, Chen, Chen, and Kalbfleisch 2004). Our inferences about trajectory shapes and the number of latent classes will be based on a generalized AIC criterion. The classical AIC criterion (Akaike 1974) compares models based on unbiased estimates of their Kullback-Liebler (KL) divergence relative to the true population structure. AIC is usually used to compare models fit using maximum likelihood estimation. In that case, the form of the asymptotic bias correction for the KL divergence is particularly simple. Here we will derive the bias correction for parametric models estimated by maximizing a regularized likelihood.

Divergence-based inference procedures such as the AIC have been controversial when applied to inference for the number of classes in a latent class model fit using the MLE (e.g. McLachlan and Peel 2000). Such approaches have been criticized for being inconsistent and for tending to favor overly-complex models, especially when selecting among models with large numbers of parameters relative to the sample size. An alternative perspective (Burnham and Anderson 2002, 2004) is that since any model for a complex system such as human behavior is substantially mis-specified, consistency may not be the most important criterion to consider. Moreover, the situation with penalized estimation may be somewhat more favorable than in other high-dimensional mixture models, since the penalization has the effect of reducing the effective number of parameters for all models. Our simulations, shown below, suggest that in the setting of penalized estimation, the finite sample performance of divergence-based inference is good.

We also use the divergence-based criterion for estimating the structure of the class probability trajectories p_{tk} . Here some parallels can be drawn to model and variable selection in regression and longitudinal modeling. A common approach to modeling curves is to represent them as linear combinations of basis vectors. Selecting which elements of the basis to include in the model is a variable selection problem. Many approaches to this problem focus on calibrating the improved fit of a more complex model relative to the number of additional free parameters. Fan and Li (2003), Shen and Ye (2002) and Pu and Niu (2006) are three recent papers that take this approach. In penalized likelihood, e.g. using (2), one must account for the parameters becoming less “free” as the value of λ is increased. It is relatively straightforward to define an “effective degrees of freedom” accounting for regularization in a regression setting (e.g. Hastie et al. 2001, Bishop 1995), but these approaches are not directly applicable in non-regression settings such as we have here. Therefore we will consider below the issue of defining an

effective degrees of freedom from the perspective of general penalized likelihoods, following the same bias correction strategy underlying the AIC.

We begin by reviewing the AIC for models fit using the MLE. Suppose we are using a parametric family of densities $f(X|\theta)$, where X denotes all observed data and θ denotes the unknown parameter vector. For the K class finite mixture model of interest here, θ will contain the probability trajectory values p_{tk} for each class, followed by the $K - 1$ marginal class probability parameters. Thus θ contains a total of $KT + K - 1$ parameters, where K is the number of classes and T is the number of time points.

Given a parametric estimator $\widehat{\theta} = \widehat{\theta}(X)$ of θ , the density of X is estimated to be $f(\cdot|\widehat{\theta})$. The Kullback-Liebler divergence between the estimated density $f(\cdot|\widehat{\theta})$ and the true density of X is (up to a linear transformation that does not depend on $\widehat{\theta}$)

$$D(\widehat{\theta}(X)) = E_Y \log f(Y|\widehat{\theta}(X)), \quad (6)$$

where E_Y denotes the expected value relative to an independent data set Y with the same distribution as X . Larger values of D indicate that the estimated density $f(\cdot|\widehat{\theta})$ is relatively more similar to the true density of the data.

Since we do not know the true distribution of the data, D cannot be computed in practice. A naive plug-in estimate for D is $\log f(X|\widehat{\theta}(X))$, but this is substantially biased upward. Akaike (1974) noted that when the estimator $\widehat{\theta}$ is the MLE, the bias

$$E_X \log f(X|\widehat{\theta}(X)) - E_X E_Y \log f(Y|\widehat{\theta}(X))$$

is simply the number of free parameters m (i.e. the dimension of θ). This led to the proposal of the Akaike Information Criterion

$$\text{AIC} = \log f(X|\widehat{\theta}(X)) - m$$

as a model selection statistic (often this is scaled by -2 but we will not do that here).

Information based inference for penalized likelihood estimates

Since a penalized MLE does not overfit the observed data to as great a degree as the MLE, the bias of the plug-in estimate of D is less than m in the penalized case, therefore the AIC cannot be directly applied to penalized maximum likelihood estimates. To derive an approximation to the bias in the penalized case, suppose that the penalized MLE is defined as follows

$$\widehat{\theta}_\lambda = \arg\max_{\theta} \log f(X|\theta) - \lambda \theta' R \theta, \quad (7)$$

where $R = FF'$ is a quadratic penalty on the parameter vector. The smoothing penalty (2) is a quadratic penalty, and can be written in the form $\theta' R \theta$ for an appropriate $KT + K - 1 \times KT - K - 1$ matrix F .

Let θ_0 be the parameter such that $f(\cdot|\theta_0)$ is closest in the sense of divergence to the true density of X . The mean squared deviation of $\widehat{\theta}_\lambda$ around θ_0 , denoted Σ_λ below, can be partitioned into variance ($\Sigma_{\lambda,1}$) and squared bias ($\Sigma_{\lambda,2}$) terms, as follows

$$\begin{aligned}\Sigma_\lambda &\equiv E(\widehat{\theta}_\lambda - \theta_0)(\widehat{\theta}_\lambda - \theta_0)', \\ &= E(\widehat{\theta}_\lambda - E\widehat{\theta}_\lambda)(\widehat{\theta}_\lambda - E\widehat{\theta}_\lambda)' + (E\widehat{\theta}_\lambda - \theta_0)(E\widehat{\theta}_\lambda - \theta_0)', \\ &= \Sigma_{\lambda,1} + \Sigma_{\lambda,2}.\end{aligned}$$

If the smoothing parameter grows proportionally with the sample size n , so that $\lambda/n \rightarrow d$, the variance term scaled by n has a limit as the sample size grows, as does the squared bias term, as follows

$$n\Sigma_{\lambda,1} \rightarrow d^{-2} \left(\bar{I}(\theta_0)/d + 2R \right)^{-1} \bar{J}(\theta_0) \left(\bar{I}(\theta_0)/d + 2R \right)^{-1} \equiv \Sigma_{d,1} \quad (8)$$

$$\Sigma_{\lambda,2} \rightarrow 4 \left(\bar{I}(\theta_0)/d + 2R \right)^{-1} R \theta_0 \theta_0' R \left(\bar{I}(\theta_0)/d + 2R \right)^{-1} \equiv \Sigma_{d,2}, \quad (9)$$

where $\bar{I}(\theta)$ is the second derivative matrix of the unpenalized log-likelihood function, and $\bar{J}(\theta)$ is the expected outer product of the gradient of the unpenalized log-likelihood function with itself. Note that since the data are iid (unconditionally on the class membership), both $\bar{I}(\theta)$ and $\bar{J}(\theta)$ are computed for a single observation, and the limits in (8) and (9) are numerical, not probabilistic.

A calculation using Taylor's expansions leads to the following value for the asymptotic bias for the penalized MLE:

$$\begin{aligned}E_y E_x \log g(y|\widehat{\theta}_d(x)) &\approx \\ &E_x \log g(x|\widehat{\theta}_d(x)) - 2d \text{tr} R \Sigma_{d,1} - \text{tr} \bar{I}(\theta_0) \Sigma_{d,1}/2 - \text{tr} \bar{I}(\theta_d^*) \Sigma_{d,1}/2 + \\ &n \text{tr} \left(\left(\bar{I}(\theta_0) - \bar{I}(\theta_d^*) \right) \Sigma_{d,2} \right) / 2.\end{aligned} \quad (10)$$

The error in equation (10) has order $O(n)$, and also makes the assumption that the penalty term is quadratic in θ , as in (7). The logarithmic boundary penalties (3) and (4) can be handled in this framework using quadratic approximations. In cases where the penalized MLE does not approach the boundary, the boundary penalties can be safely ignored.

In practice, a simpler bias correction can be obtained if the following two assumptions are made: (i) $\bar{I}(\theta_0) = \bar{I}(\theta^*)$, and (ii) $f=g(\cdot|\theta_0)$. Letting $T_1, \dots, T_q$ denote the eigenvalues of $F' \bar{I}(\theta_0)^{-1} F$, the limiting bias simplifies to

$$m - \sum_{i=1}^q \frac{2dT_i}{1+2dT_i}, \quad (11)$$

where m is the total number of parameters, and q is the rank of R ($m = KT + K - 1$ and $q = KT - 2K$ in the finite mixture model for trajectories).

It is easy to see that when $d = 0$ (no smoothing), the classical AIC bias correction for the MLE of the saturated model is recovered (i.e. the bias correction is the number of free parameters, m). At the opposite extreme, when $d = \infty$ the trajectory estimates are linear, and $\hat{\theta}_\lambda$ becomes the MLE for a model with two parameters per class. Since the rank of R is $K(T - 2)$, again the AIC bias correction is recovered: $KT + K - 1 - K(T - 2) = 2K + K - 1$, where the $2K$ summand accounts for the intercept and slope of the linear probability trajectory within each class, and the $K - 1$ summand accounts for the marginal class probabilities. When $d < 0 < \infty$, the bias correction (11) lies between $2K + K - 1$ and m .

Since (11) provides a bias correction for $D(\hat{\theta}_\lambda)$ that is applicable to the penalized maximum likelihood estimates (7), inference can be approached by comparing the bias corrected divergence estimates for the models of interest. We will continue to refer to these bias corrected divergence estimates as “AIC” values, although they are not calculated using the standard AIC formula. Models with differing numbers of classes, with some or all class mean trajectories constrained to be constant, and/or with different λ values can be fit and their AIC values compared. The number of classes and the number of constant trajectories are inferred from the structure of the selected model.

The interpretation of the inferred λ value differs from the interpretation of quantities relating to the model structure itself (e.g. the number of latent classes or the number of time-invariant trajectories). The value of λ does not affect the parametric family of models being fit to the data, but rather indexes a family of estimators. The value of λ selected by the AIC is interpreted as defining an estimator that provides the most accurate estimation of the likelihood in the sense of KL divergence. Below we will provide some empirical evidence that this also provides the most precise trajectory estimates in the sense of mean squared error. It is important to note that the true population trajectories may be less smooth than those estimated under the selected λ value. In such a case, given the noise level and sample size of the available data, a smoothed estimate that forgoes small-scale features gives the most precise estimates. If the sample size is large, a small λ value will be selected (unless the population trajectories are linear), allowing estimation of fine-scale features.

Simulation studies

We used four simulation models to study the sampling properties of the proposed estimation and inference procedures. Figure 2 shows the four probability trajectory pairs on which these models are based. They include a model in which both trajectories are linear (model 1), and three models in which one or both trajectories are nonlinear. For each trajectory pair, five-hundred data sets were simulated from the one-class model in which all subjects belong to the second trajectory class. Then five-hundred additional data sets were simulated from the two-class model in which each individual was assigned with probability 1/2 to either trajectory class one or two.

Each of the four models was considered for a sample size of 600 individuals over 13 observed time points under two sampling schemes. In the “dense sampling scheme” each individual is assessed at all 13 time points. In the “sparse sampling scheme”, which follows the design of the MLS analyzed below, individuals were randomly assigned to be initially assessed at time point 1, 2, or 3 (with equal probabilities), and every third time point following the starting point was observed (thus for 2/3 of the sample, 5 of the 13 possible measurement points are observed, and for the remainder of the sample 4 measurement points are observed). The dense sampling scheme is more typical of developmental studies, since observing only four points over the 13 year time span makes resolving changes very challenging. Note in particular that models 3 and 4 are very challenging under the sparse sampling scheme. For model 4, at most one dichotomous measure per subject will be substantially informative about class membership.

For each simulated data set, penalized maximum likelihood estimates were calculated for $\lambda = 0, 10, 100$, and 1000 , for both the one class and two class model. Thus a total of eight models were compared for each simulated data set based on the AIC. Using $\lambda = 0$ and $\lambda = 1000$ gives MLE's for, respectively, an unpenalized model with 13 free parameters per class, and (approximately) a model with linear class mean trajectories having two free parameters per class. Using $\lambda = 10$ and $\lambda = 100$ gives intermediate levels of regularization.

For illustration, Figure 3 shows some estimation results from the simulation study. The top row shows the mean trajectory estimates for both classes over all 1000 simulation replications. The bottom row shows the trajectory estimates based on five randomly selected data sets from the simulation study. The left column shows results for model 3 and the right column shows results for model 4. All results shown are based on densely sampled data using $\lambda = 10$.

Performance for estimation of the number of latent classes

Table 1 shows the results of the simulation study in terms of the estimated number of latent classes. Powers for inferring the correct number of latent classes for each fixed value of λ are given in columns 6 and 10 for the dense and sparse schemes, respectively. The overall power (not conditioned on the selected value of λ) would be obtained by averaging the values in columns 6 and 10 using the values in columns 3 and 7 (which show the frequency that each λ value was selected), as weights. For the linear model 1, power exceeding 0.9 for identifying the correct number of latent classes occurs for both the dense and sparse schemes. For the nonlinear models 2-4, the dense scheme always has power substantially greater than 0.8. Models 2-4 in the sparse scheme are very challenging, and show only moderate power.

Columns 4 and 5 contain, respectively, the proportion of true one-class models that are inferred to be one class, and the proportion of true two-class models that are inferred to be two-class. Columns 8 and 9 contain the corresponding values for the dense sampling scheme. The values in columns 4 and 8 are of special interest since they are characterize a “type I” error in which a more complex model is incorrectly favored over a more parsimonious one. For the sparse sampling scheme, type I error rates are well under 0.05, giving “hypothesis test-like” inference. For the dense sampling scheme, in which the sample size is effectively three times larger, type I error rates are around 25% for models 2 and 4, 5% for model 1, and 15% for model 3 (all for the most frequently-selected λ value). As a practical heuristic, by moving to the next higher λ value over that favored by the AIC, type I error rates are always below 0.05, with little or no decrease in power. A more principled way to achieve this would be to adapt the method proposed by Shen and Ye (2002).

Interestingly, in some cases, such as model 2 in the dense sampling scheme, the power actually increases when λ is increased beyond its estimated value. Evidently, for purposes of estimating the number of latent classes, it is sometimes beneficial to use an over-simplified and misspecified model. The only exception to this heuristic is model 3, in which the classes can only be distinguished through their highly nonlinear features, and hence large values of λ hide the key distinguishing information. The structure of model 3 is interesting to consider, but it is not very plausible for most growth curve mixture modeling applications.

Performance for trajectory estimation

Our major interest is inference about the time-varying pattern of class-specific symptom-endorsement probabilities p_{tk} . Columns 3 and 7 of Table 1 show the frequencies with which each possible value of λ is selected by the AIC. Large values of λ are correctly favored in model 1 where the population trajectories are linear ($\lambda = 1000$ is selected 82% and 89% of the time for the dense and sparse schemes respectively). For the other three models, where the population structure is not linear, smaller λ values tend to be selected. The plurality of fits for

models 2-4 favor $\lambda = 10$. The fact that the AIC procedure tends to select the same penalties for models 2-4 may suggest that these three models are similar in estimation difficulty. Interestingly, the MLE ($\lambda = 0$) is selected less than 11% of the time, except for the highly nonlinear model 3. This suggests that with a sample size of 600 subjects observed over 13 regularly-spaced points, fully non-parametric estimation of trajectory patterns rarely yields the highest possible estimation precision.

Next we considered the bias and variance of parameter estimates $\hat{p}_{tk}$ calculated using each of the four penalized estimators (for $\lambda = 0, 10, 100$, or 1000). The key principle of penalized maximum likelihood estimation is the bias/variance tradeoff — more accurate trajectory estimates (in the sense of mean squared error) result from using a biased estimation procedure when the variance is reduced substantially. We therefore calculated integrated mean squared errors (IMSE's) $\sum_{tk} (\hat{p}_{tk} - p_{tk})^2$, where t runs over the 13 time points and k indexes the two classes.

The IMSE values are shown in Table 2. The values corresponding to model structure estimates with greater than 15% frequency are underlined. For the fits using our proposed method, there is substantial agreement between the λ values selected by AIC and the λ values giving the best IMSE. For the dense sampling scheme the agreement is perfect: the optimal IMSE occurs for $\lambda = 1000, 10, 10$, and 10 for models 1, 2, 3, and 4 respectively, and these are also the λ values most frequently selected by the AIC, as shown in table 1. Under the sparse sampling scheme, for models 1 and 4 there is a unique best λ value in terms of IMSE, which coincides with the most frequently-selected λ value ($\lambda = 1000$ for model 1 and $\lambda = 10$ for model 4). For models 2 and 3, there are multiple λ values that give approximately optimal performance, and the AIC strongly tends to select one of these values. Specifically, for model 2 the IMSE values for $\lambda = 10, 100$, and 1000 are quite similar, and together these are selected 98% of the time using the AIC. For model 3, the AIC prefers $\lambda = 10$ and $\lambda = 1000$ but not $\lambda = 100$, while the IMSE values for $\lambda = 10, 100$, and 1000 are roughly similar.

We also considered IMSE values for each class separately. In general, the MSE for estimating p_{tk} will be affected by three factors: (i) the value of p_{tk} (with values closer to $1/2$ having greater MSE), (ii) the curvature of p_{tk} near time t (with greater curvature leading to greater bias induced by smoothing, and hence greater MSE), and (iii) the ability to correctly assign individuals to latent classes. For models 1 and 3, the IMSE's for the two classes are nearly identical (the difference between the two class-specific IMSE's is less than 5% of their mean value). For model 1, since both population trajectories have no curvature (they are linear) and have similar marginal frequencies averaged over time, they are approximately equally easy to estimate. In the case of model 3, the equivalence of the class-specific IMSE's is due to the symmetry. For models 2 and 3, the IMSE's turn out to be substantially class-specific, with greater IMSE values in the non-flat class. For model 2, the IMSE ranges from being about 10% greater in the non-flat class ($\lambda = 0$) to being about twice as great in the non-flat class ($\lambda = 1000$). For model 4, the IMSE ranges from being about 50% greater in the non-flat class ($\lambda = 0$) to being almost completely concentrated in the non-flat class ($\lambda = 1000$). These are mostly driven by factor (ii) above, where bias is introduced by forcing the fitted trajectories to be approximately linear when the truth is not linear. It is important to note that AIC strongly tends to select λ values for which the IMSE is approximately equally distributed between the two classes.

Basis methods are widely used in longitudinal data analysis. Specifically, polynomial methods are widely used in longitudinal modeling, and spline methods have been previously used in latent class time series analysis (Winsberg and De Soete, 1999) and in mixed effects longitudinal analysis (Jacqmin-Gadda, 2002). For regression models it has been shown that penalized least squares estimation with (2) is equivalent to fitting the trajectories with a basis of cubic splines (Wahba 1990). However the spline coefficients are not the standard least squares

estimates, but are shrunken toward zero depending on the value of λ . For more complex models such as finite mixtures, there is no simple theoretical correspondence between penalized likelihood estimates and cubic spline fits.

For comparing our method to basis methods, we computed IMSE values for latent class models using two different sets of basis vectors for modeling p_{tk} . These models are fit using the MLE instead of the penalized MLE, but somewhat the same effect as penalization is attained by using bases of smooth functions. In practice, basis methods can be applied by either pre-specifying the dimension, as is often done when quadratic or cubic polynomial bases are used, or by searching over a family of bases $B_1, B_2, \dots$ of increasingly higher dimension to optimize a criterion like the AIC. The latter approach is more common with spline bases. Note that using the T -dimensional basis B_T for trajectories observed at T points is equivalent to applying our approach when $\lambda = 0$ (both are parameterizations of the saturated model). Typically much lower dimensional bases than B_T are used unless a very large data set is available.

A basis method achieves a good MSE when the true trajectories are approximately in the span of B_k for a low value of k . Comparing basis methods to our approach is therefore not straightforward — the result of a comparison depends strongly on the choice of population structure, since the basis approach will have lower MSE if the true values of p_{tk} are specified to lie in a B_k for a small value of k , and our approach will have lower MSE if this is not the case. However empirical evidence, described below, suggests that in the former case the advantage of the basis approach over our method is slight, whereas in the latter case the advantage of our method can be very substantial. To illustrate this, we fit models using a cubic polynomial basis (a common choice in practice) and a five dimensional B-spline basis. The IMSE values for these two fits are given in Table 2, with the polynomial IMSE's in the column denoted P_3 and the spline IMSE's in the column denoted S_5 .

In the challenging “sparse” sampling scheme, with at most four observations per subject over the 13 time points, all methods perform similarly for models 1, 2, and 3. For model 4, a strong parametric assumption that happens to be approximately correct yields lower MSE's for the basis methods compared to our method, which aims to adapt to a wide variety of trajectory shapes. In model 4, the polynomial and spline methods are about 10% and 30% more efficient than our approach, respectively. The reason for this is that the population trajectory (Figure 2) happens to be reasonably well approximated by an element of S_5 .

For the dense sampling scheme, bias and variance contribute more equally to trajectory estimation. In models 3 and 4, where the bias is substantial, we see that the low dimensional polynomial and spline models perform much worse than our method. All methods perform comparably in models 1 and 2, where the bias is minimal. Of course a higher dimensional spline basis would have less bias. As the dimension of the spline basis is increased, the MSE would converge to the result of our method using $\lambda = 0$. Overall, the simulation analysis suggests when the population structure favors basis methods, our method also performs well, while when the population structure is not compatible with the chosen basis, our method gives substantially better estimation.

Application to the Michigan Longitudinal Study

The Michigan Longitudinal Study (MLS, Zucker et al. 1996, 2004) is a prospective long-term study of initially intact families at high risk of substance abuse. Behavioral assessments of children began at ages 3-5, and the study is ongoing today with the core group now at ages 18-20. The data used here are binary parent ratings on 34 behavioral items (coded “never true” or “sometime/always true”) from the aggression and anxiety/depression subscales of the Child Behavior Checklist (CBCL) (Achenbach & Edelbrock 1983), the most widely used child behavioral assessment instrument in the United States. Data were obtained for 598 children in

349 families observed over some part of the age range 3 to 15. Interviews and questionnaire assessments with the parents of each child were carried out at 3-year intervals, with up to four assessments per child obtained. The time variable t for an assessment was taken to be the child's age, rounded to the nearest year. For most children, behavioral assessments from both the mother and father are available. For simplicity, we only use the mothers' assessments here. The 598 children in our sample include some siblings. Dependencies resulting from family structure are not considered here.

Our goal was to use the MLS data to illustrate our proposed regularized finite mixture estimation procedure over a variety of different measured trajectories. Thus, while a comprehensive analysis of the MLS data would require multi-item analyses, here we fit separate models to each of the 34 items. Since each subject is assessed at most four times during the 13 year span of the study, this data set is very challenging as a basis for uncovering longitudinal patterns on a time scale of one year. Thus we expect the penalization method to select a relatively small "effective degrees of freedom," and uncertainty levels for the fits will be relatively high. Nevertheless we demonstrate below that a number of interesting findings can be rigorously substantiated.

We fit a total of nine models per item. Specifically, we considered models with one, two, or three latent classes, and for which any subset of the latent classes have constant trajectories over time (thus, we fit two one-class models, three two-class models, and four three-class models). The reason for allowing constant classes is that the presence or absence of significant temporal variation in a class is one of the most important questions from the perspective of developmental psychology. For model selection, we calculated the AIC values adjusted for penalized fits as discussed above, and used them to rank the nine models for each item. Figure 4 shows plots of the fitted trajectories for the most favored model for each item.

For inference, we considered the range of AIC values produced for the nine models. When working with AIC values, it is common to consider two models to fit equivalently well if their AIC values differ by less than two units (e.g. Burnham & Anderson 2002). Since we do not include the factor of two in our calculation (see comment above), this becomes one unit for us. We also used bootstrapping, as discussed below.

Of the 34 behavioral items, five (all in the aggression domain) were deemed to have three latent classes, and 29 were deemed to have two latent classes. different λ values were selected by the AIC for different items: $\lambda = 1000$ was selected for 21 items, $\lambda = 100$ was selected for 9 items, and $\lambda = 10$ was selected for 4 items. The MLE for the saturated model ($\lambda = 0$) was never selected. Fifteen of the items had one latent class deemed to have a constant trajectory. These were generally classes with a very low symptom level (e.g. "threatens") or a very high symptom level (e.g. "showing off"). More than a single time-invariant trajectory was never selected for one item.

As seen in Figure 4, the estimated trajectory patterns show a range of patterns. Importantly, several of the fitted class means would not be well-represented by the typical quadratic growth curve parameterization. Two clear examples are "fears might do something bad" and "others out to get her/him," which show relatively narrow symptom level peaks at age 9 and age 10, respectively. In the case of "fears might do something," rather than falling off the peak according to a global quadratic pattern, the levels stabilize at around 0.4 for ages outside of the peak. In the case of "others out to get her," a stabilization also occurs, but at a substantially higher level following the peak age compared to preceding it. Several other items have a concave class which is markedly asymmetric, and hence would be incorrectly represented under a quadratic parameterization. These items include "worrying," "unhappy," "complains of loneliness," and "has to be perfect."

We evaluated the uncertainty in our model selection using AIC values. Figure 5 shows the results of all AIC calculations for each of the 34 CBCL items. Each of the 34 columns in the figure corresponds to one of the CBCL items. The 9 rows indicate the rankings of the models for a given item (models with any marginal class probability less than 0.05 are not shown). Each of the 9×34 cells in the figure contains two sets of boxes. The boxes on the left indicate the total number of classes, and the boxes on the right indicate the number of latent classes with constant trajectories. Filled boxes correspond to models that are equivalent to the best model for a given item, in the sense that the AIC gap is less than two, as discussed above. Unfilled boxes indicate inferior models in this sense. Based on Figure 5, we see that for 23 items, one of the models was clearly superior to the others, while for the other 11 items, either two, or in one case (“Argues”) three models fit the data to a similar degree. We also see strong evidence for time-varying trajectories, in that the models with no time-varying trajectories are strongly concentrated toward the lower rows of the chart.

While Figure 5 shows an inferential analysis for model selection, Figure 6 shows an inferential analysis for parameter estimation. Here we used the bootstrap, holding the model structure (i.e. the total number of classes and the number of classes with constant trajectories) fixed at the configuration favored by the AIC. We used the bootstrap to construct 95% pointwise confidence intervals for each trajectory. For illustration purposes, results are shown for a single item, “bragging,” which had a three-class solution. The confidence intervals are fairly wide, due to the sample size and sparse assessment schedule. Nevertheless the intervals are sufficiently narrow to justify interpretations of the three classes as comprising approximately constant low and high classes, along with a much more time-varying class in which peak symptoms occur for seven to nine year olds. We also used the bootstrap to construct 95% confidence intervals for the marginal class probabilities Q_k . The intervals, in the same order as the trajectories appear in Figure 6, are: 0.40(0.01,0.46), 0.34(0.27,0.46), and 0.27(0.23,0.60).

In terms of implications for research on child development, several observations of potential importance can be made. First, there is a great deal of heterogeneity in the fits for different behavioral items, both in terms of the apportionment of subjects to the lower versus higher symptom classes, and in the trends and nonlinearities in the fitted trajectories. Although some of this may be due to sampling variation, and hence not reproducible, it is striking how items thought to belong to a common higher level behavioral domain often show substantially different trajectory patterns. One such example is “disobedient at home” compared to “disobedient at school.” Both items are related to aggression, and in particular to a facet of aggression related more to rule breaking than to physical violence. Given its public quality, disobedience at school is a more severe variant of disobedience at home, which is consistent with the pattern of class assignments: 243 children were rated as disobedient at home but not at school, while only 21 were rated as disobedient at school but not at home (class assignments were made according to the individual posterior probabilities). This helps to explain the lower marginal probability of the high-symptom class for disobedience at school (0.16 versus 0.41). In addition, the high marginal probability of the high “disobedient at home” symptom trajectory indicates its more normative quality, and is also consistent with the well-known decline in undercontrolled behavior over the course of childhood (Loeber & Hay, 1997). However it is more difficult to explain why the high symptom trajectory of “disobedient at school” is dramatically increasing. This example illustrates how item-level temporal patterning can be obscured in higher-level scales, which may in turn obscure lower frequency developmental processes.

Another notable observation is that relatively few class mean trajectories were inferred to be constant. In results not shown here, we found that if the saturated MLE ($\lambda = 0$) were used for every item, then nearly every item would include one, and often several constant classes. This illustrates an important property of penalized model estimation: by increasing the precision of

trajectory estimates via regularization, more subtle deviations from time-invariance are able to be identified with rigorous statistical support. An example of this would be the high symptom class of “threatens,” which declines gradually by a total of 0.2 on the probability scale over the 13 year age-span studied. Under a saturated MLE, the AIC would have set this trajectory to be constant over time.

As noted above, the analysis presented here considers each item separately. While there are certainly dependencies among the items, our goal was to evaluate our procedure for fitting trajectories of a single measure, so we do not consider the dependencies in depth here. The results shown in Figure 4 suggest some commonalities but also reveal many differences in how the various items operate. As expected, many items have a class with low symptom levels comprising a large fraction of the population. A check of the underlying individual assignment probabilities reveals, as expected, a substantial but not complete overlap of the individuals exhibiting these trajectories.

Discussion

Estimation of growth mixture models has generally proceeded using maximum likelihood (ML) estimation, although difficulties with identifying stable and interpretable solutions have frequently been reported. Here we illustrate how a penalized likelihood approach, much like that used in Functional Data Analysis (e.g. Ramsay & Silverman 2005) can be used to reduce the imprecision of trajectory estimates in growth mixture models, thereby making them more interpretable, and allowing more subtle inferences to be made with statistical rigor.

A key advantage of the proposed approach is that it is able to adapt to different sample sizes, temporal sampling schemes, and patterns of linearity or nonlinearity in the population structure. If the sample size permits precise estimation of fine-scale nonlinear patterns in the class mean trajectories, this method has good power for detecting them. On the other hand, if the sample size is small so that only the overall trend can be precisely estimated, then the method will adapt itself by imposing greater regularization constraints on the estimates.

Using simulation studies, we have shown that a generalized version of the Akaike Information Criterion (AIC) that accounts for the penalization in the likelihood estimation can be used to make inferences about the number of latent classes, and about the time-varying patterns in the class mean trajectories. These simulation studies demonstrate that the AIC criterion works well. While it is somewhat permissive in estimating the number of classes for large data sets, consistent with previous reports in the literature, this can be addressed by inferring the number of classes based on an over-penalized likelihood. Based on integrated mean square error (IMSE), we found that the AIC consistently selects the optimal value of λ for efficient estimation of the class-mean probability trajectories, under a variety of model structures and sampling schemes. In spite of the good performance of the AIC, criteria for regularized estimation based on the Bayesian Information Criterion (BIC, Schwarz 1978) would also be useful, since such criteria have the desirable property of being consistent when the model is exactly-specified.

We applied the proposed method to 34 separate dichotomous longitudinal measures of behavior in a long term family study data set. The resulting class trajectory estimates are interpretable, and many would not be adequately represented using a traditional quadratic parameterization. Our analysis demonstrates that a variety of non-constant and nonlinear class mean trajectories are likely to be present in the MLS population, and we found that many of these nonlinear patterns would not have been reliably found using maximum likelihood estimation. This result has important implications for understanding behavioral development in this high-risk population of children.

The proposed estimation procedure is somewhat complex, but the model itself is relatively straightforward to understand, and ordinary users are shielded from calculational issues by well-written software. Among many possible issues for future work, we will close by briefly discussing three here. One issue is whether alternatives to the smoothing penalty (2) would outperform our proposed approach. Two possible alternatives would be to penalize the log odds $\log p_{tk}/(1 - p_{tk})$ rather than the actual probabilities, or to replace the square in (2) with an absolute value. Either of these can be justified in principle, although the extent of any practical benefit would need to be investigated. The AIC analysis would also become more complex.

A second possible issue for future work is whether substantial gains could be made by using class-specific penalty coefficients λ_k in place of the universal coefficient λ (e.g. in 2). One aspect of this issue is that we are not interpreting λ as a meaningful parameter of the population, but rather as a means of selecting an estimator in a data-adaptive way that gives a good balance between bias and variance in parameter estimation. Thus the use of a constant λ should not be taken to indicate that an inference is being made about the structure of the population. Nevertheless it remains possible that more precise estimates could be obtained if class-specific λ values were used. While a systematic analysis of this question is beyond the scope of this work, there are reasons to believe that not much is lost by using a common λ value when the number of time points is in the range considered here. For example, suppose the truth is that a nearly constant class trajectory and a fairly nonlinear class trajectory exist (this is typical in the MLS results). In the AIC, the cost of a few extra parameters for the flat class will be more than offset by the reduction in bias for the nonlinear class. However if the time grid were very fine, the balance may shift in favor of class-specific λ values, albeit at a heavy computational cost.

Another important issue for future work is whether the increased complexity of modeling serial correlations between an individual's measurements given class membership is warranted. For continuous response data, finite mixture models with Gaussian random effects (Verbeke & Lesaffre 1996) have been used to account for such residual correlations. In the dichotomous case, serial correlations could be handled by allowing a Markov transition structure within each class. Nagin et al. (2001) argued in a similar setting that the considerable additional complexity of modeling residual correlations may not justify the possible benefits. A separate more conceptual issue is that in behavioral studies it is generally unrealistic to aim for the true data generating model — a finite mixture model is always a working model allowing a potentially meaningful group structure to be identified. If class-specific serial correlations are modeled, they may partly or completely define a class. It will be important to consider whether it is desirable that a grouping should be influenced by serial correlation properties rather than mean properties, which have traditionally been the main determinant of inferred latent class structures.

Acknowledgments

This work was supported in part by NIAAA grants R37 AA07065 and R01 AA12217 to RAZ.

References

- Achenbach, T.M.; Edelbrock, C. Manual for the Child Behavior Checklist and Revised Child Behavior Profile. University of Vermont, Department of Psychiatry; Burlington. VT: 1983.
- Akaike H. A new look at the statistical model identification. *IEEE Transactions* 1974;19:716–723.
- Bauer DJ, Curran PJ. The Integration of Continuous and Discrete Latent Variable Models : Potential Problems and Promising Opportunities. *Psychological Methods* 2004;9(1):3–29. [PubMed: 15053717]
- Bishop, C. *Neural Networks for Pattern Recognition*. Clarendon Press; Oxford: 1995.

- Dunson DB, Herring AH. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* 2005;6(1):11–25. [PubMed: 15618524]
- Chen H, Chen J, JD, Kalbfleisch. Testing for a Finite Mixture Model with Two Components. *Journal Royal Statistical Society B* 2004;66:96–116.
- Burnham, KP.; Anderson, DR. Vol. 2nd. Springer-Verlag; New York, NY: 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach*.
- Burnham KB, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 2004;33:261–305.
- Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of American Statistical Association* 2004;99:710–723.
- Fergusson D, Horwood LJJ, Nagin DS. Offending Trajectories in a New Zealand Cohort. *Criminology* 2000;38:525–552.
- Galindo-Garre F, Vermunt JK. Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika* 2006;33:43–59.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. Springer-Verlag; 2001.
- Jacqmin-Gadda H, Joly P, Commenges D, Binquet C, Chene G. Penalized likelihood approach to estimate a smooth mean curve on longitudinal data. *Statistics in Medicine* 2002;21(16):2391–402. [PubMed: 12210623]
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. John Wiley; New York: 1987.
- Loeber R, Hay D. Key issues in the development of aggression and violence from childhood to early adulthood. *Annual Review of Psychology* 1997;48:371–410.
- McLachlan, GJ.; Peel, D. *Finite Mixture Models*. John Wiley & Sons, Inc.; New York: 2000.
- Moos, RH.; Moos, BS. *Family Environment Scale*. Vol. 3rd. Consulting Psychologists Press; Palo Alto, CA: 1994.
- Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 1999;55(2):463–9. [PubMed: 11318201]
- Nagin DS. Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach. *Psychological Methods* 1999;4(2):139–157.
- Nagin DS, Tremblay RE. Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-Based Method. *Psychological Methods* 2001;6(1)
- Nagin, DS. *Group-based modeling of development*. Harvard University Press; Cambridge MA: 2005.
- WJ, Pu; Niu, XF. Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis* 2006;97:733–758.
- Ramsay, J.; Silverman, BW. *Functional Data Analysis*. Vol. 2nd. Springer Series in Statistics; 2005.
- Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B* 1991;53:233243.
- Raudenbush, S. Toward a coherent framework for comparing trajectories of individual change. In: Sayers, A.; Collins, L., editors. *New methods for the analysis of change*. American Psychological Association; Washington DC: 2001.
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978;6:461–464.
- Shen XT, Ye JM. Adaptive model selection. *Journal of the American Statistical Association* 2002;97:210–221.
- Titterton, DP.; Smith, AFM.; Makov, UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley and Sons; 1985.
- Verbeke G, Lesaffre E. A linear mixed effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association* 1996;91:217–221.
- Vermunt, JK.; Magidson, J. *Technical guide for Latent GOLD 4.0*. Statistical Innovations; 2004.
- Wahba, G. *CBMS-NSF Regional Conference Series*. SIAM; Philadelphia: 1990. *Spline Models for Observational Data*.
- Winsberg S, De Soete G. Latent class models for time series analysis. *Applied Stochastic Models in Business and Industry* 1999;15(3):183–194.
- Yung YF. Finite mixtures in confirmatory factor-analysis models. *Psychometrika* 1997;62:297–330.

- Zucker, RA.; Fitzgerald, HE.; Refior, SK.; Puttler, LI.; Pallas, DM.; Ellis, DA.; D.A.. The clinical and social ecology of childhood for children of alcoholics: Description of a study and implications for a differentiated social policy. In: Fitzgerald, HE.; Lester, BM.; Zuckerman, BS., editors. *Children of Addiction: Research, Health and Policy Issues*. Routledge Falmer; New York: 2000. p. 109-141.
- Zucker RA, Ellis DA, Bingham CR, Fitzgerald HE. The development of alcoholic subtypes: Risk variation among alcoholic families during early childhood years. *Alcohol Health & Research World* 1996;20:46–54.

Figure 1.

Fitted class probability trajectories for one item of the Michigan Longitudinal study, using four different values of the penalization parameter λ .

Figure 2.
True probability trajectories for the simulation studies for correctly-specified models.

Figure 3.

Top row: mean trajectory estimates for model 3 (left) and model 4 (right). Bottom row: five randomly selected trajectory estimates for one of the classes of model 3 (left) and model 4 (right). In all cases estimates from “densely sampled” data sets were used (see text for details).

Figure 4.

Fitted trajectories for 34 CBCL behaviors assessed over time in the MLS. For each behavior, the trajectory showing the age-specific probability of a "sometimes/always true" assessment are shown for each class, with the probability given on the vertical axis, age on the horizontal axis, and marginal class probabilities given below each graph. Items from "Argues" to "Unusually loud" belong to the aggression subscale while items from "Complaints of loneliness" to "Worrying" belong to the anxiety/depression subscale.

Figure 5.

AIC ranking of nine models for each of the 34 CBCL items in the MLS. The figure shows a 9×34 array of cells, where each column shows the ranking of models from best fitting (top row) to worst fitting (bottom row) for a given CBCL item. The boxes in the left part of a cell indicate the total number of latent classes, and the boxes in the right part of a cell indicate the number of constant latent classes. Unfilled boxes represent models that are inferior to the best ranked ranked model for a given item, based on the AIC gap. Fits in which any class' marginal probability was less than 0.05 are not shown. The item numbers correspond to the same order as the plots shown in Figure 4.

Figure 6.
Bootstrap 95% confidence intervals for the class probability trajectories of the “bragging” CBCL item in the MLS sample.

Simulation study results. For each of four model structures, 500 replicates with one class ($K = 1$) and 500 replicates with two classes ($K = 2$) were generated. The values in columns 4-6 and 8-10 denote the proportion of simulated data sets for which the true number of classes was correctly identified based on the given λ value in column 2. Columns 3 and 10 denote the proportion of the 1000 simulated data sets in which the given value was selected based on AIC. Each simulation was carried out for dense sampling (13 time points), and for sparse sampling in which every third of the 13 time points is observed (see text for details).

Table 1

Model structure	λ	Sparse sampling				Dense sampling			
		Prop. λ selected	$K = 1$ called correctly	$K = 2$ called correctly	K called correctly	Prop. λ selected	$K = 1$ called correctly	$K = 2$ called correctly	K called correctly
1	00.002		0.960	0.752	0.856	0.015	0.698	1.000	0.849
	100.043		0.880	0.952	0.916	0.018	0.770	1.000	0.885
	1000.135		0.944	0.964	0.954	0.076	0.890	1.000	0.945
	10000.820		0.976	0.968	0.972	0.891	0.954	1.000	0.977
2	00.021		0.972	0.156	0.564	0.109	0.708	0.998	0.853
	100.577		0.904	0.360	0.632	0.675	0.772	0.998	0.885
	1000.295		0.978	0.332	0.655	0.211	0.970	0.998	0.984
	10000.107		0.998	0.240	0.619	0.005	0.988	1.000	0.994
3	00.093		0.964	0.220	0.592	0.340	0.710	1.000	0.855
	100.474		0.976	0.402	0.664	0.660	0.838	1.000	0.919
	1000.045		0.986	0.008	0.497	0.000	0.982	0.984	0.983
	10000.388		0.998	0.000	0.499	0.000	1.000	0.016	0.508
4	00.007		0.968	0.046	0.507	0.093	0.758	0.984	0.871
	100.731		0.908	0.134	0.521	0.846	0.768	0.994	0.881
	1000.262		0.994	0.034	0.514	0.061	0.984	0.978	0.981
	10000.000		1.000	0.008	0.504	0.000	0.996	0.642	0.819
Col. num.	1	2	3	4	5	6	7	8	

Table 2

Trajectory IMSE's. For each true two-class model given by the four model structures, and for each of the four penalized estimators determined by the value of λ , the estimated and population trajectory patterns were compared based on the IMSE. Settings of λ selected more than 15% of the time by AIC are underlined. IMSE values are also shown for a cubic polynomial fit (P_3) and a ve-dimensional B-spline t (S_5).

λ	Sparse sampling					Dense sampling							
	0	10	100	1000		P_3	S_5	0	10	100	1000	P_3	S_5
1	0.65	0.12	0.09	0.07		0.10	0.10	0.03	0.02	0.02	0.01	0.01	0.01
2	0.66	0.22	0.22	0.24		0.18	0.24	0.06	0.04	0.04	0.07	0.06	0.06
3	0.81	0.39	0.35	0.33		0.38	0.37	0.07	0.05	0.10	0.30	0.28	0.15
4	0.63	0.46	0.46	0.90		0.41	0.32	0.10	0.05	0.10	0.34	0.23	0.09