The Annals of Applied Statistics
2018, Vol. 12, No. 1, 222–245
https://doi.org/10.1214/17-AOAS1066

HOW GAUSSIAN MIXTURE MODELS MIGHT MISS DETECTING FACTORS THAT IMPACT GROWTH PATTERNS[1]

BY BRIANNA C. HEGGESETH AND NICHOLAS P. JEWELL

Williams College and University of California, Berkeley

Longitudinal studies play a prominent role in biological, social, and behavioral sciences. Repeated measurements over time facilitate the study of an outcome level, how individuals change over time, and the factors that may impact either or both. A standard approach to modeling childhood growth over time is to use multilevel or mixed effects models to study factors that might play a role in the level and growth over time. However, there has been increased interest in using mixture models, which have inherent grouping structure to more flexibly explain heterogeneity in the longitudinal outcomes, to study growth patterns. While several possible model specifications can be used, these methods generally fail to explicitly group individuals by the shape of their growth pattern separate from level, and thus fail to shed light on the relationships between growth pattern and potential explanatory factors. We illustrate the weaknesses of these methods as they are currently being used. We also propose a pre-processing step that removes the outcome level to focus explicitly on shape, discuss its impact on estimation, and demonstrate its usefulness though a simulation study and with real longitudinal data.

1. Introduction. A key advantage of a longitudinal study is the ability to observe the evolution of an outcome measured repeatedly over time. The path which these measurements take can be viewed as a longitudinal trajectory and has many interesting features that include the overall level and the growth pattern. Most longitudinal methods attempt to model the outcome level across time based on baseline and time-varying variables but fewer explicitly focus on the second of these features, the growth pattern. We attempt to isolate these two features so they can be studied separately, as factors that impact the starting level may differ from those acting on how the outcome grows or changes. For example, birth weight may be determined by a set of factors separate from those that are associated with the subsequent physical growth pattern of a child.

The motivating data example for this article is from the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study. A cohort of pregnant women were enrolled in 1999–2000 and the mother-child pairs have

Received July 2016; revised May 2017.

[1]Supported by the National Science Foundation Vertical Integration of Research and Education Grant DMS-0636667 and National Institute of Environmental Health Sciences Grant R01-ES015493.

Key words and phrases. Finite mixture model, longitudinal data analysis, latent variables, growth curves.

been interviewed about every other year for physical, neurodevelopment, and environmental assessments. One of the study aims is to examine the impact of in-utero exposure to chemicals on childhood physical growth. Most attempts to study this type of relationship have only used physical measurements at a single time point, such as at ages 14 months [Mendez et al. (2011)], 14–22 months [Cupul-Uicab et al. (2010)], 6.5 years [Valvi et al. (2012)], 7 years [Cupul-Uicab et al. (2013), Warner et al. (2013)], and 9 years [Warner et al. (2014)]. Others have collected and analyzed longitudinal data using mixed effects or multilevel linear models but variability in growth patterns cannot be easily explained with a linear model of measured covariates. Therefore, many researchers have turned to Gaussian mixture models to account for heterogeneity in growth patterns in body mass index (BMI) over time [Carter et al. (2012), Garden et al. (2012), Pryor et al. (2011)]. While these methods are more flexible, they may not be able to accurately address the aims of the CHAMACOS study, which are to study whether early life exposure impact growth and development over time. In some instances, the Gaussian mixture model may miss detecting, or may incorrectly estimate, significant relationships with development. We describe our concerns about the standard longitudinal methods and illustrate the weaknesses of the widely used mixture model specifications. To allow for utilization of existing software, we propose a data preprocessing step to improve estimation of relationships with growth patterns using mixture models.

Standard longitudinal analysis approaches, based on generalized linear models and multilevel models [Diggle et al. (2002), Singer and Willett (2003)], can be used to estimate the mean outcome over time while accounting for within-individual correlation. Typically, variation in the growth pattern is explicitly modeled through linear models of slope coefficients which leads to interaction terms in the single composite model [Heo et al. (2003)]. However, the process of modeling nonlinear growth patterns and nonlinear relationships between baseline factors and growth with this type of structure is not trivial and model assumptions are often hard to verify. Additionally, interpretations are not straightforward when growth is not linear.

Due to the limitations of the standard methods, there has been increased interest and usage of a more flexible mixture model approach [Erosheva, Matsueda and Telesca (2014), Nagin and Odgers (2010a, 2010b), Pickles and Croudace (2010)]. This data-driven approach attempts to approximate the distribution of growth patterns with latent discrete variables that separate the population into homogeneous trajectory groups. The group membership probabilities may be modeled based on baseline factors. Unlike most other clustering methods, these models can accommodate unbalanced observation times and missing data, common with longitudinal studies. The most-widely used methods are based on a finite mixture model, some of which date back to Quetelet in 1846 or more famously, the work of Karl Pearson in 1894 [Pearson (1894)]. A more thorough history and introduction can be found in standard textbooks [Everitt and Hand (1981), McLachlan and Basford (1988),

Titterington, Smith and Makov (1985)]. The recent increase in use of these models can be partly attributed to availability of software to estimate two finite mixture model specifications known as growth mixture models (GMM) [Muthén and Shedden (1999), Proust-Lima et al. (2014)] and a special type of GMM referred to as group-based trajectory modeling, or latent class growth analysis (LCGA) [Jones, Nagin and Roeder (2001), Nagin (1999)]. One main difference between these specifications is how dependence between repeated measures is handled. The GMM uses random effects to account for longitudinal dependence while LCGA assumes the measurements are independent conditional on group membership. Both models use a regression structure to model group means across time and use the generalized logit function to allow baseline factors to impact group membership probabilities. These two frameworks have been well studied [Muthén and Asparouhov (2009), Muthén et al. (2002)]. Further details of the model specification are given in Section 2.

While our present focus is the Gaussian mixture model for continuous outcomes, recent work allows nonnormal group densities such as skewed-t distributions [Asparouhov and Muthén (2016), Huang, Chen and Yin (2017), Lu and Huang (2014)]. While these models are more robust to skewed error or random effect distributions, they do not focus directly on the growth pattern over time. As mentioned by Asparouhov and Muthén (2016), the skew-t mixture model can lead to a more parsimonious model with larger groups potentially at the expense of differentiating subtle but interesting differences, thus making it harder to detect possible relationships with baseline factors.

The grouping structure in finite mixture models is defined by the component distributions, whether Gaussian or skewed-t. Therefore, the densities are estimated to explain the most variability in the data, no matter whether the variability is due to between-subject differences in the level or in the growth pattern shape. Thus, we cannot be sure whether resulting groups are based on level, shape, or a combination of both. Related, researchers have noted the importance and difficulty of disentangling these types of effects through statistical modeling. For example, Morin and Marsh (2015) explored ways to separate the level and shape effects (in this case, shape was defined as the pattern of factors, not in terms of growth pattern) using variations of a latent profile analysis model. Closer to our goal, some researchers have proposed dissimilarity measures that remove the level of longitudinal trajectories by defining groups based on the estimated first derivative of the growth trajectory [D'Urso (2000), Möller-Levet et al. (2003)]. While studying "the heterogeneity in the level, shape, and stability of the developmental trajectories" [Morin et al. (2013)] is a commonly stated goal, there has little discussion about explicitly separating these effects in a mixture model for longitudinal data.

In this paper, we introduce the finite Gaussian mixture model as well as the LCGA and GMM specifications in Section 2, and then we illustrate the potential issue of missing the relationship of factors with growth using LCGA or GMM in Section 3. In Section 4, we propose processing continuous outcome data prior to

fitting a mixture model so as to directly group based on shape and discuss our proposal in the context of other data processing and standardizations. Modeling challenges due to the pre-processing step and implementation details are discussed in Sections 5 and 6. A simulation study, demonstrating the merits of the proposed method in comparison to the standard longitudinal clustering methods, LCGA and GMM, is described in Section 7. Then the method is applied to the motivating data example as well as another data set in Section 8.

2. Finite mixture model. In a finite mixture model, the probability density function of a random vector $\boldsymbol{y} \in \boldsymbol{R}^m$ takes the form

$$f(\boldsymbol{y}) = \pi_1 f_1(\boldsymbol{y}) + \cdots + \pi_K f_K(\boldsymbol{y}),$$

where $\pi_k > 0$ for $k = 1, \ldots, K$ and $\sum_k^K \pi_k = 1$. The parameters $\pi_1, \ldots, \pi_K$ are prior group probabilities and the functions $f_1, \ldots, f_K$ are group probability densities. Given our motivating data example, we focus on continuous outcome vectors. In this case, the group densities are assumed multivariate Gaussian. This model can be extended by including a regression structure for the mean and for the prior probability of belonging to a group [Wedel (2002)]. That is, the mixture density for $\boldsymbol{y}$ conditional on a $m \times p$ explanatory matrix, $\boldsymbol{X}$, and a q-dimensional vector of static or baseline variables, $\boldsymbol{w}$, is defined by

$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{w}, \boldsymbol{\gamma}) f_k(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta_k}), \tag{1}$$

where $f_k(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta_k})$ denotes the m-variate Gaussian probability density function with mean $\boldsymbol{X\beta_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$, and the vector $\boldsymbol{\theta_k}$ includes mean and covariance parameters for the kth group with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T, \boldsymbol{\gamma}^T)^T$. With longitudinal data, the matrix $\boldsymbol{X}$ is typically based on a functional basis for time appropriate for the growth pattern (e.g., quadratic, cubic, or spline). With group-specific parameters, each group has their own level, growth pattern, and relationship with other covariates. The prior group probabilities can be parameterized using the generalized logit function $\pi_k(\boldsymbol{w}, \boldsymbol{\gamma}) = \frac{\exp(\boldsymbol{w}^T \boldsymbol{\gamma_k})}{\sum_{j=1}^K \exp(\boldsymbol{w}^T \boldsymbol{\gamma_j})}$ for $k = 1, \ldots, K$ where $\boldsymbol{\gamma_k} \in \boldsymbol{R}^q$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}^T, \ldots, \boldsymbol{\gamma_{K-1}}^T)^T$, and $\boldsymbol{\gamma_K} = \boldsymbol{0}$ for identifiability.

LCGA and GMM are based both on this general finite mixture model, but they differ in how they accommodate dependence between repeated measures. The LCGA assumes homoscedastic, independent errors conditional on group membership such that $\boldsymbol{\Sigma_k} = \sigma^2 \boldsymbol{I}$, while the Gaussian GMM accounts for dependence between repeated measures through random effects such that $\boldsymbol{y_i} = \boldsymbol{X_i \beta_k} + \boldsymbol{Z_i \alpha_{ik}} + \boldsymbol{\varepsilon_{ik}}$ where $\boldsymbol{\alpha_{ik}} \sim N(0, \boldsymbol{\Psi}_k)$ and $\boldsymbol{\varepsilon_{ik}} \sim N(0, \sigma_k^2 \boldsymbol{I})$, resulting in $\boldsymbol{\Sigma_{ik}} = \boldsymbol{Z_i \Psi_k Z_i^T} + \sigma_k^2 \boldsymbol{I}$. Longitudinal outcome measures have inherent dependence and misspecifying the covariance structure in a Gaussian mixture model can result in incorrect estimation of the number of groups, misclassification of individuals,

and bias in the parameter estimates, depending on group separation and how well the modeling correlation structure approximates the truth [Davies, Glonek and Giles (2015), Diallo, Morin and Lu (2016), Gray (1994), Heggeseth and Jewell (2013)].

Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{im_i})$ denote an outcome vector of m_i repeated observations for the ith individual ($i = 1, \ldots, n$). The vector of corresponding observation times is denoted as $\boldsymbol{t_i} = (t_{i1}, \ldots, t_{im_i})$. We assume that each individual is a member of one group and the probability of being in the kth group depends on baseline factors in $\boldsymbol{w_i} \in \boldsymbol{R}^q$ collected at, or before, time t_{i1}. Then, for an individual in the kth group, their outcome at time t_{ij} is

$$y_{ij} = \lambda_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta_k} + \varepsilon_{ij}, \tag{2}$$

where λ_i is an individual-specific intercept, $\boldsymbol{x}_{ij}^T$ is the jth row in the $\boldsymbol{X_i}$ matrix and $\boldsymbol{\varepsilon}_{ij} \sim N(0, \boldsymbol{\Sigma}_k)$. In this article, we want to consider growth separate from level; therefore, we restrict the matrix $\boldsymbol{X}_i$ to include only time-varying covariates to model growth and let λ_i encompasses the level. While the level is not our focus in this article, one could model the variability in λ_i with a combination of linear models of nontime-varying covariates, random effects, or a grouping structure.

3. Limitations of Gaussian mixture model. The finite Gaussian mixture model and the specifications popularized through the availability of software are not inherently poor models. They are useful in many situations; however, they were not developed to address questions about relationships between growth patterns and time-invariant factors. The Gaussian mixture model implicitly defines dissimilarity between individuals using the Mahalanobis distance which measures the normalized point-wise distance between outcome measurements. In model estimation using maximum likelihood, group densities are determined to minimize dissimilarity within groups. If the level of the outcome measurements dominate the observed variability, then the groups will be determined by level and will not necessarily be homogeneous in terms of growth pattern.

To illustrate how level can drive group estimation, we show a simple example with three different development patterns: increasing linearly, decreasing linearly, and constant over time. We let the starting level be weakly related to the slope of the trajectory; in this case, we deliberately let the intercepts follow a nonnormal distribution. The resulting groups from fitting a three-group Gaussian mixture model with a linear mean function assuming independence (LCGA) or a random slope (GMM) do not correspond with the three underlying development patterns used to generate the data (Figure 1). Rather, the groups are determined only by level and the horizontal group mean trajectories do not accurately reflect the majority of individual growth patterns. The LCGA model is not robust to high variability in the intercepts within groups. A three-group GMM with a random intercept and slope model explains some variation in the intercepts but these mixture models are

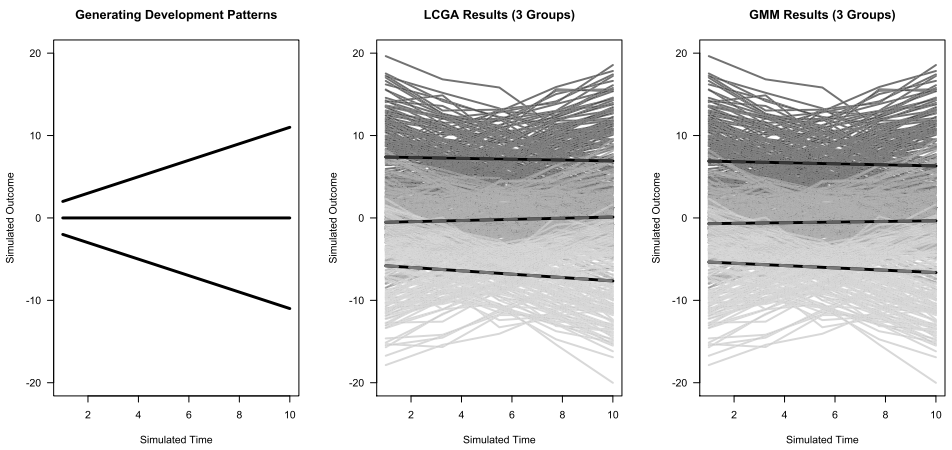

FIG. 1. *Three generating growth patterns* (*left*) *and* 500 *simulated general outcome trajectories shaded according to the estimated classification into three groups with overlaid group mean trajectories estimated with two Gaussian mixture model specifications*: *LCGA* (*center*) *and GMM* (*right*).

not robust to multimodal intercept distributions, which can occur when individuals with drastically different starting levels have the same development pattern.

This issue can also be seen with data that emulates childhood growth patterns. We simulated BMI trajectories based on three growth patterns resembling those observed in our motivating data set (more details in Section 8): a shallow U-shape, a slanted J-shape pattern with a higher rate of change in later childhood, and a linear growth pattern. To mimic real data, we simulated the starting BMI level of each individual dependent on the pre-pregnancy maternal obesity status with a child of an obese mother having a BMI about 3 kg/m^2 higher than a child with a nonobese mother, on average. Then we allowed the growth pattern to be nonlinearly related to a simulated early-life exposure such that the probability of having J-shape growth was the highest amongst those exposed followed by linear and then U-shape. Of those who were not exposed, the probability of having the U-shape was highest followed by linear and then J-shape. Note that maternal obesity only impacts the level and not the shape in this example. Based on 500 simulated growth trajectories generated in the manner explained, we see how the level can drive the group membership, and thus the estimated baseline factor parameters (Figure 2). For the two common specifications, the Gaussian mixture models resulted in estimated groups that were not homogenous in growth pattern. With the LCGA, exposure was not significantly associated with the resulting groups. With the incorporation of random intercepts and slopes into the model, the GMM estimated that both exposure and obesity were significantly associated. We contend that this model does not explicitly separate the effects of level and shape so it is hard to make clear interpretations. These two simple examples illustrate how

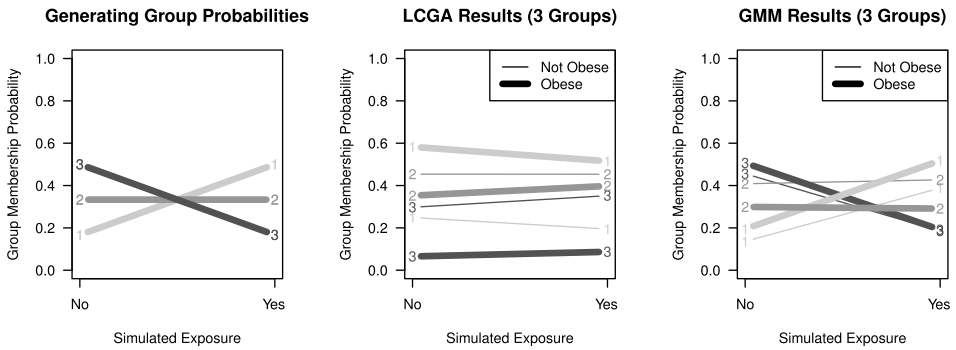

FIG. 2. *Group membership probabilities used to generate three development pattern groups based on the exposure for both normal and obese women* (*left*) *and estimated group membership probabilities based on two baseline factors* (*maternal obesity and exposure*) *from* 500 *simulated BMI developmental trajectories with three growth patterns* (1. *J-Shape*, 2. *Linear*, 3. *U-shape*) *based on three-group Gaussian mixture model specifications*: *LCGA* (*center*) *and GMM* (*right*).

easy it is to incorrectly estimate the relationship between a baseline factor such as early-life exposure and the growth patterns.

Many may argue that while these models are not necessarily detecting development pattern groups, they provide meaningful groups as the level is an important aspect of a longitudinal trajectory. That may be true, especially in the context of physical childhood growth as there are many clinical reasons to distinguish between an underweight and obese child and determine which factors lead to these levels and subsequent health conditions. Many statistical methods are designed to address those questions. We argue that if the effort has been made to collecting data over time, growth pattern is a key characteristic to study on its own. As we have just briefly illustrated, many methods, including the standard Gaussian mixture models, do not explicitly study that change over time. In this article, we propose a pre-processing step that aims to allow the use of widely-used mixture model software while focusing on growth patterns and their relationships with baseline factors.

4. Pre-processing. To explicitly model growth patterns using a Gaussian mixture model, we need to deal with the variability in the level. One could use measured covariates and random effects, but in practice, unexplained variability may be hard to model with small to moderate sample sizes. Therefore, we propose a pre-processing step to treat the level as a nuisance and remove it.

The idea of standardizing data prior to analysis is common in many scientific fields and for many statistical methods. In lab experiments, measurements often are normalized within observational units to compensate for known sources of variability between samples such as with microarray data [Park et al. (2003)]. In multivariate data analysis, variables with varying units and magnitudes often are

standardized when clustering so as to equalize variable contributions [Everitt et al. (2011)]. In classical time series, models are decomposed into deterministic and stochastic components, so removal of the mean trend is almost always advised [Brillinger (1975), Shumway and Stoffer (2010)].

With longitudinal data, the variables are repeated outcome measurements with the same units. Unlike typical multivariate data, standardization of variables is not needed or appropriate. Unlike classical time series, we have many trajectories of repeated measures and we want to explicitly model the trend pattern over time. But normalizing measurements within observational units, like microarray data, treats the magnitude of the level as a nuisance and highlights change over time.

One possible adjustment involves subtracting the first observed outcome from each subsequent observation to model the change over time akin to estimating the longitudinal rather than cross-sectional effect [Diggle et al. (2002)]. This process directly removes the intercept, but it is sensitive to measurement error in the first outcome and works only when the first observation time is consistent across individuals. To remedy these issues, our proposed pre-processing step removes the level by subtracting the within-individual mean outcome from each measurement. This adjustment accommodates different baseline observation times and is less sensitive to individual measurement errors. Subtracting a mean is not novel, but we believe that its untapped application within the context of mixture models provides an appreciable improvement in estimating relationships with growth patterns.

In notation, let $\bar{y}_i = m_i^{-1} \sum_{j=1}^{m_i} y_{ij}$ be the mean of the vector of outcome measurements for individual i. This measure of the overall outcome level of the individual can be removed by applying the centering matrix, $\boldsymbol{A_i} = \boldsymbol{I_{m_i}} - m_i^{-1} \mathbf{1_{m_i}} \mathbf{1_{m_i}}^T$, to the vector of observations. Based on the model in equation (2), the centered vector for individual i is denoted as $\boldsymbol{y_i^*} = \boldsymbol{A_i y_i} = \boldsymbol{\mu_i} - \bar{\mu}_i + \boldsymbol{\varepsilon_i} - \bar{\varepsilon}_i$ where $\boldsymbol{\mu_i} = \boldsymbol{X_i \beta_k}$, $\bar{\mu}_i$, and $\bar{\varepsilon}_i$ are the mean of their respective vectors. Standardizing the data in this manner directly removes the individual intercept, λ_i, and thus there is no need to model the between-individual variability in the level.

5. Challenges. Finite mixture models involve some well-known challenges such as identifiability which requires minor constraints for estimation [McLachlan and Peel (2000)]. Here, we discuss important modeling and estimation challenges that result from the proposed pre-processing. To illustrate these issues, let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ be a random vector observed at times $\boldsymbol{t} = (t_1, \ldots, t_m)$ such that $\boldsymbol{Y} = \lambda \mathbf{1_m} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ where for simplicity, $\lambda \sim G$, $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_m))$, $\mu(t)$ is a smooth deterministic function of observation time, and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma} = \boldsymbol{V^{1/2} R}(\rho, \boldsymbol{t}) \boldsymbol{V^{1/2}}$ where $\boldsymbol{R}(\rho, \boldsymbol{t})$ is an $m \times m$ correlation matrix based on a parameter ρ and observation times $\boldsymbol{t}$, and $\boldsymbol{V}$ is a $m \times m$ matrix with variances along the diagonal. Then $\boldsymbol{Y}^* = \boldsymbol{AY}$ is the centered random vector where $\boldsymbol{A} = \boldsymbol{I_m} - m^{-1} \mathbf{1_m} \mathbf{1_m^T}$.

One important characteristic of the proposed pre-processing step is that it is noninvertible, the effects of the centering matrix cannot be reversed by multiplying

another matrix. While the centering does not impact modeling the trend over time, it has important consequences on the dependence structure within the data. The covariance of the errors after centering equals

$$\operatorname{Cov}(\boldsymbol{Y}^* - \boldsymbol{\mu}) = \operatorname{Cov}\big((\boldsymbol{A} - \boldsymbol{I}_m)\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\varepsilon}\big).$$

If the observation times are fixed, then the mean vector $\boldsymbol{\mu}$ is constant. If the observation times are random, then $\boldsymbol{\mu}$ is a random vector that contributes to the variability in the errors. From this point on, $\boldsymbol{I_m}$ will be written as $\boldsymbol{I}$ and $\mathbf{1_m}$ as $\mathbf{1}$ for ease of use.

If the observation times $\boldsymbol{t}$ are fixed, then the covariance matrix is singular with the form $\operatorname{Cov}(\boldsymbol{Y^*} - \boldsymbol{\mu}) = \boldsymbol{A\Sigma A^T}$ where $\boldsymbol{\Sigma} = \operatorname{Cov}(\boldsymbol{\varepsilon})$. If the original data have equal variance across time, $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, then the resulting covariance after centering is a linear combination of the original correlation, column and row means, and the overall mean correlation,

$$\begin{aligned}\boldsymbol{A\Sigma A^T} = \sigma^2\big(&\boldsymbol{R}(\rho, \boldsymbol{t}) - m^{-1}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{R}(\rho, \boldsymbol{t}) - m^{-1}\boldsymbol{R}(\rho, \boldsymbol{t})\mathbf{1}\mathbf{1}^{\boldsymbol{T}} \\ &+ m^{-2}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{R}(\rho, \boldsymbol{t})\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\big).\end{aligned}$$

When the errors are independent, $\boldsymbol{R}(\rho, \boldsymbol{t}) = \boldsymbol{I}$, the resulting covariance matrix is $\operatorname{Cov}(\boldsymbol{Y^*} - \boldsymbol{\mu}) = \sigma^2(\frac{m-1}{m})(a\mathbf{1}\mathbf{1}^{\boldsymbol{T}} + (1-a)\boldsymbol{I})$ where $a = \frac{-1}{m-1}$. If the correlation structure of $\boldsymbol{\varepsilon}$ is exchangeable with correlation ρ, then $\operatorname{Cov}(\boldsymbol{Y^*} - \boldsymbol{\mu}) = (1-\rho)\sigma^2(\frac{m-1}{m})(a\mathbf{1}\mathbf{1}^{\boldsymbol{T}} + (1-a)\boldsymbol{I})$. So in either circumstance, subtracting the mean induces a negative exchangeable correlation of $\frac{-1}{m-1}$ between vector elements. On the other hand, if the correlation of the errors is exponential, then the covariance matrix has negative values when the mean correlation within columns and rows is positive and large.

The true matrix structures, while they can be written down in closed-form, cannot be used for modeling the covariance as the singularity will disrupt estimation. Therefore, an approximation for the covariance matrix is needed. If there is short range dependence in the errors, independence or a decaying autocorrelation structure may be adequate.

In practice, individuals in a longitudinal study are not typically observed at fixed times, but rather there is variation in the observation times. This irregular spacing adds variability to the centered random vector. If $\operatorname{Cov}(\boldsymbol{\varepsilon})$ does not depend on observation times, then $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ are independent and the covariance is approximately

$$\operatorname{Cov}(\boldsymbol{Y^*} - \boldsymbol{\mu}) \approx m^{-2}\left(\sum_{j=1}^{m} \operatorname{Var}(t_j)\big[\mu'\big(E(t_j)\big)\big]^2\right)\mathbf{1}\mathbf{1}^{\boldsymbol{T}} + \boldsymbol{A\Sigma A^T}$$

by the delta method assuming the times $t_1, \ldots, t_m$ are pair-wise independent. In this case, the covariance depends on the variability of observation times as well as the derivative of the mean function. If the data are sampled with random variation, there is limited evidence to suggest that common correlation structures may

provide an adequate approximation. However, more research is needed to develop structures to better approximate the resulting correlations.

So far, we have assumed all subjects have the same number of observations. However, the number of observations per subject plays a role in the covariance of the centered vector. If subjects have unequal numbers of observations, the assumption of a common covariance structure within groups does not hold. Limited simulations with unequal observations due to missing data suggest little impact on the resulting groups if the number of observations per subject is moderately large and adequately cover the observation time period.

6. Implementation.

6.1. *Mean and covariance*. Due to the possible irregularity of observation times, we impose structure on the mean and covariance within the groups, with or without pre-processing the data. As noted in Section 5, pre-processing by removing the individual level does not change how we model the mean growth pattern, but it does change the dependence so we must be mindful when choosing a correlation structure.

A polynomial basis provides structure but may be too restricting for complex growth patterns. Alternatively, the growth pattern can be modeled by a more flexible functional basis such as a B-spline basis [Curry and Schoenberg (1966), De Boor (1976, 1978)]. This basis accommodates more complex shapes by dividing the codomain with internal knots and fitting piecewise polynomials. Knots can be placed at local extrema and inflection points of the overall trends [Eubank (1999)] or at sample quantiles based on all observation times [Ruppert (2002)]. Care must be taken when selecting the number of knots and polynomial order so as to moderate the number of parameters. Basis function values at observation times for individual i are used in the matrix, $\boldsymbol{X_i}$, to model the growth patterns.

The irregular nature of observation times in longitudinal studies often places modeling limitations on the covariance matrix [Jennrich and Schluchter (1986)] but adequate approximations are needed to avoid bias [Heggeseth and Jewell (2013)]. The LCGA specification assumes conditional independence and equal variance ($\boldsymbol{\Sigma_k} = \sigma^2 \boldsymbol{I}_{m_i}$), and GMM utilizes random effects to restrict the covariance structures. One common random effect model is a simple random intercept model which is equivalent to a model with an exchangeable covariance structure where all repeated measures are equally correlated with correlation ρ_k [$\boldsymbol{\Sigma_k} = \sigma_k^2(\rho_k \mathbf{1}_{\boldsymbol{m_i}} \mathbf{1}_{\boldsymbol{m_i}}^{\boldsymbol{T}} + (1-\rho_k)\boldsymbol{I}_{m_i})$ where $-\frac{1}{m_i} < \rho_k \leq 1$]. The covariance structure for a random slope model depends on covariate values as well as distributional assumptions about the random effects. Beyond LCGA and GMM, other specifications allow the covariance to decay as the time between observations increases. With a stationary exponential or autoregressive model, the covariance between the jth and lth observation equals $\sigma_k^2 \exp(-|t_{ij} - t_{il}|/r_k)$ where $r_k > 0$ captures the range of the dependence. If the range r_k is small, the correlation decays quickly,

but large r_k will induce long-range dependence among measurements within an individual. The covariance structure can be selected from a set of options through a model selection procedure. If any choices result in nonconvergence on preprocessed data, it may be due to the singularity of the covariance matrix and we recommend trying another structure.

6.2. *Estimation and inference.* Under the assumption that $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*$ are independent realizations from a mixture distribution, $f(\boldsymbol{y}^*|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\theta})$, defined in (1), the log likelihood function for the parameter vector, $\boldsymbol{\theta}$, is given by $\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i^*|\boldsymbol{X}_i, \boldsymbol{w}_i, \boldsymbol{\theta})$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained by finding an appropriate root of the score equation, $\partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. Solutions of this equation corresponding to local maxima can be found iteratively through the expectation-maximization (EM) algorithm [Dempster, Laird and Rubin (1977)]. The EM algorithm guarantees convergence to a local maximum; global convergence may be attained by running the algorithm multiple times with initial random group assignments or parameter estimates and using estimates associated with the highest log likelihood. The algorithm also returns posterior probability estimates of group membership, written as $\alpha_{ik} = \pi_k(\boldsymbol{w}_i, \boldsymbol{\gamma}) f_k(\boldsymbol{y}_i^*|\boldsymbol{X}_i, \boldsymbol{\theta}_k) / \sum_{j=1}^{K} \pi_j(\boldsymbol{w}_i, \boldsymbol{\gamma}) f_j(\boldsymbol{y}_i^*|\boldsymbol{X}_i, \boldsymbol{\theta}_j)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, which are used to partition individuals into groups by selecting the group with the highest posterior probability for each individual.

Estimation requires the number of groups with unique growth patterns, K, to be known. In practice, K is chosen by setting a maximum value such that $K_{\max} < n$, fitting the model under all values of $K = 2, \ldots, K_{\max}$, and choosing the value that optimizes a chosen criteria [Celeux and Soromenho (1996), Fraley and Raftery (1998)]. In this article, we select K using the Bayesian Information Criterion (BIC) [Schwarz (1978)] defined as $\text{BIC} = -2 \log L(\hat{\boldsymbol{\theta}}) - d \log(n)$ where d is the dimension of $\boldsymbol{\theta}$. Additionally, the BIC can be used to choose a covariance structure during the model selection process.

In the context of mixture models, statistical inference often comes in the form of hypothesis tests to choose K and confidence intervals for parameter estimates. Standard asymptotic likelihood theory fails to produce an adequate null distribution for the likelihood ratio statistic as the test sits on the boundary of the parameter space [Aitkin, Anderson and Hinde (1981)]. Bootstrapping approaches [Efron (1979, 1982)], both parametric [Feng and McCulloch (1996), McLachlan and Basford (1988)] and nonparametric [Schlattmann and Böhning (1997)], have been offered as alternatives to provide an approximate null distribution. For confidence intervals, most software provide standard errors from likelihood theory which can be used to calculate Wald-type confidence intervals for parameters estimates of the growth patterns. Alternatively, bootstrapping may be used to calculate standard errors or bootstrap confidence intervals.

Currently, many software packages can estimate a Gaussian mixture model with regression mean structure and a choice of covariance structures. `Proc Traj` in

SAS and Stata fits LGCA models that assume independence and equal variance [Nagin (1999), Jones, Nagin and Roeder (2001)] and Mplus fits GMM that use random effects to provide covariance structure [Muthén and Muthén (1998–2010), Muthén and Shedden (1999)]. The `hlme` function in the `lcmm` R package [Proust-Lima et al. (2014)] allows random effects as well as an an autoregressive process for residual autocorrelation but restricts residual variance to be equal across groups and the `flexmix` R package [Grün and Leisch (2008), Leisch (2004)] assumes independence but allows variances to differ across groups. The authors used the `lcmm` R package for most examples in this paper.

7. Simulation study. To study the performance of the standard mixture methods, with and without proposed pre-processing, in detecting groups with homogeneous growth patterns, we completed a simulation with three trajectory shapes and three outcome levels in the population similar to the first simple example of Figure 1 in this paper. Data were generated with three distinct mean growth patterns, $\mu_1(t) = -1 - t$, $\mu_2(t) = 0$, $\mu_3(t) = -11 + t$. We note that if the model is unable to distinguish among these three very distinct but simple growth patterns, then likely it will fail with more subtle differences observed in practice. We let $c(i) \in \{1, 2, 3\}$ indicate the growth pattern for individual i. The probability of following a growth pattern at a particular starting level depended on two binary factors. The first factor, $w_{1i} \in \{0, 1\}$, determined the growth pattern whereas the second, $w_{2i} \in \{0, 1\}$, impacted the starting level. Individual values for these factors were randomly assigned independently by simulated tosses of a fair coin.

We let the random intercept equal $\lambda_i = \lambda_{1i} + \lambda_{2i}$ where $\lambda_{1i} \in \{0, 12\}$ and $\lambda_{2i} \sim N(0, \sigma_\lambda^2)$. The probability of $\lambda_{1i} = 0$ was about 0.05 when $w_{2i} = 1$ and 0.95 when $w_{2i} = 0$ if $c(i) = 1$ or 3 and 1 if $c(i) = 2$. This effectively resulted in a bimodal Gaussian mixture distribution for λ_i within mean pattern group 1 and 3 and a Gaussian distribution in group 2.

The probability of individual i following the kth pattern was $P(c(i) = k|w_{1i}) = \frac{\exp(\gamma_{0k} + \gamma_{1k} w_{1i})}{\sum_{l=1}^{3} \exp(\gamma_{0l} + \gamma_{1l} w_{1i})}$ for $k = 1, 2, 3$ where $\gamma_{01} = 2, \gamma_{11} = -4, \gamma_{02} = 1.5, \gamma_{12} = -2, \gamma_{03} = \gamma_{13} = 0$ and $w_{1i} \in \{0, 1\}$. It can be shown that each growth pattern had about an equal chance, marginally.

For individual i, the outcome at the jth observation time was a realization of $y_{ij} = \lambda_{1i} + \lambda_{2i} + \mu_{c(i)}(t_j) + \varepsilon_{ij}$ at times $t = 1, 3.25, 5.5, 7.75, 10$ where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $\lambda_{2i} \sim N(0, \sigma_\lambda^2)$, σ_ε was the standard deviation of the measurement error, and σ_λ was the standard deviation of the level perturbation.

To test the impact of overlap between groups, we adjusted the variability of the level and let $\sigma_\lambda = 2$ or 3. For the signal to noise ratio, we let $\sigma_\varepsilon = 0.5$ or 2 to adjust the magnitude of measurement error. The four possible combinations of these parameter values represents the conditions of the data-generating process in this simple simulation study.

For each condition, we generated a data set of $n = 500$ individuals using the process described above and fit five different Gaussian mixture models: independent mixture (similar to LCGA), random intercept GMM, random intercept and slope GMM, independent mixture on pre-processed data, and mixture with exponential correlation on pre-processed data. The variance parameters were allowed to differ between groups in all but the random intercept and slope model due to model complexity. The two correlation structures for the processed mixture were chosen based on an empirical variogram [Diggle et al. (2002)] from one simulated data set. Note that a random intercept model cannot be estimated for this generated pre-processed data due to the singularity of the estimated covariance matrix.

Even though we generated the data with linear growth patterns, we used a quadratic polynomial basis to create the matrix $\boldsymbol{X}_i$ so as to avoid imposing a priori knowledge while limiting the number of parameters. For each model, we estimated the model parameters for $K = 2, 3, 4$, and 5, chose the optimal K using the BIC. We calculated the adjusted Rand Index [Hubert and Arabie (1985), Rand (1971)] to measure the agreement among estimated groups based on optimal K and the three generating growth pattern groups and the misclassification rate when $K = 3$ as the percentage of individuals not correctly classified using the true generating pattern groups as a reference. Since a mixture is identifiable up to a permutation of group labels, we mapped the classifications from the model to the true labels such that the misclassification rate was minimized. This process was repeated 500 times under each condition.

7.1. *Results*. Table 1 summarizes the simulation results in terms of frequency of number of groups chosen as optimal, the mean BIC and the mean adjusted Rand Index when the optimal number of groups is chosen, and the mean misclassification rate when $K = 3$ over the 500 replications. With the raw data, the average BIC suggests that a GMM is preferable over the independent Gaussian (LCGA). However, neither standard model specifications consistently detected the three development patterns in the simulated data sets. The independent mixture frequently selected five groups while the number of groups under the random effect mixtures (GMM) varied among three, four, and five groups. One might argue that discovering five groups in this simulation is ideal because we have five combinations of patterns and levels due to the multimodal intercept distribution, but for all conditions, the mean adjusted Rand Index for the independent Gaussian was less than 0.5, whereas a value of 1 indicates agreement among groupings. Our goal is to have a method that explicitly detects the growth pattern similarity in groups to improve the estimation of the growth pattern and the relationships with baseline factors.

When constrained to form three groups, the standard model specifications correctly classified about 50–85% of the data on average into the true shape groups. Amongst these three models, the random intercept mixture is the closest to the true data generating distribution, and thus performs the best in terms of misclassification when measurement error is small.

TABLE 1

Frequency table of the number of groups chosen, mean BIC, mean adjusted Rand Index (ARI), and mean misclassification rate (MR) ($K = 3$) for 500 replications from a variety of finite mixture model specifications applied to data generated under different values for σ_ε and σ_λ

σ_ε	σ_λ	$K = 2$	$K = 3$	$K = 4$	$K = 5$	BIC	ARI	MR
			Independent Gaussian Mixture					
0.50	2.00	0	0	10	490	12,008	0.46	0.41
2.00	2.00	0	0	8	492	12,766	0.48	0.39
0.50	3.00	0	0	4	496	13,186	0.31	0.45
2.00	3.00	0	0	4	496	13,611	0.32	0.45
			Random Intercept Gaussian Mixture					
0.50	2.00	0	1	178	321	8642	0.79	0.16
2.00	2.00	1	2	131	366	11,694	0.68	0.38
0.50	3.00	0	3	222	275	9007	0.81	0.15
2.00	3.00	0	0	104	396	11,798	0.74	0.39
			Random Slope Gaussian Mixture					
0.50	2.00	0	4	116	380	8778	0.64	0.44
2.00	2.00	0	1	116	383	11,383	0.63	0.47
0.50	3.00	0	1	11	488	8821	0.67	0.49
2.00	3.00	0	1	38	461	11,590	0.63	0.49
			Pre-Processed Independent Mixture					
0.50	2.00	0	499	0	1	5777	0.99	0
2.00	2.00	0	499	1	0	9222	0.98	0.01
0.50	3.00	0	499	0	1	5776	0.99	0
2.00	3.00	0	499	1	0	9221	0.98	0.01
			Pre-Processed Exponential Mixture					
0.50	2.00	0	500	0	0	5795	1	0
2.00	2.00	0	500	0	0	9240	0.98	0.01
0.50	3.00	0	500	0	0	5795	1	0
2.00	3.00	0	500	0	0	9240	0.98	0.01

If the data are processed prior to fitting the mixture model with either correlation structure, three groups are detected as the optimal number of groups about 99% of the time. Additionally, after centering the data, a mixture model discovered the common shape patterns with little to no misclassification when fixing $K = 3$. When comparing the two correlation assumptions, the independence structure is preferred over the exponential structure because the additional correlation parameters did not drastically improve the fit as measured by the BIC.

It is important to note how group membership impacts the estimated relationship with baseline factors. Figure 3 shows the group probabilities based on mean parameter estimates of $\boldsymbol{\gamma}$ for one condition ($\sigma_\varepsilon = 2$, $\sigma_\lambda = 3$, and $K = 3$). We note that both the random intercept GMM and the pre-processed mixture correctly indicates that the groups are related to w_1 and not the second baseline factor w_2,

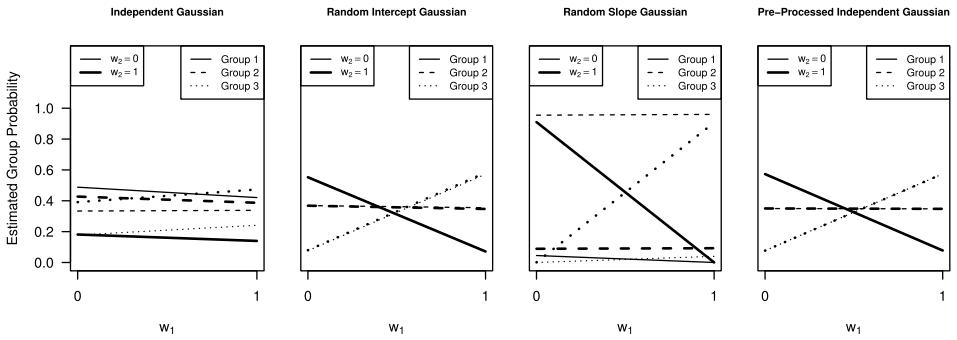

FIG. 3. *Estimated group probabilities for baseline factor values of* w_1 *and* w_2 *based on mean parameter estimates of* $\boldsymbol{\gamma}$ *for* 4 *out of the* 5 *models from the simulation study.*

which was used to determine starting levels. On the other hand, the independent Gaussian (LCGA) indicates no relationship between group probabilities and the first baseline factor w_1, which was used to determine growth patterns. The random slope mixture is a compromise and groups are determined by both baseline factors. These results are not surprising based on the mean misclassification rates and the illustration in Figure 1.

This simulation was designed to emulate a situation in which the level and shape were weakly related such that the starting value is not necessarily predictive of the subsequent trend. In these situations, centering the data prior to modeling provides the most benefit for detecting growth patterns and the factors that impact them. We used linear patterns in this simulation for the sake of simplicity, but in general, for any mixture model, including all of those discussed in this paper, it will be more difficult to discriminate between nonlinear growth patterns that are similar especially when the signal to noise ratio is small. For additional comparison, we ran a simulation study with growth patterns similar to the childhood growth data example, described in the next section, and found that the misclassification rates increase slightly, but the pre-processed model still outperformed mixture models without the pre-processing when attempting to detect a known number of distinct nonlinear development patterns. See the supplemental article [Heggeseth (2018a)] for more details.

8. Data examples.

8.1. *CHAMACOS.* The Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) Study is a longitudinal birth cohort study designed to assess the health effects of pesticides and other environmental exposures on the growth and development of low-income children living in the agricultural Salinas Valley, CA [Eskenazi et al. (2004, 2005)]. Of 601 pregnant women enrolled in the study in 1999–2000, a total of 527 mostly Latino mother-child singleton pairs were followed through a live-birth delivery and 327 pairs continued to

be followed through the 9-year interview. Baseline maternal characteristics were measured at the start of the study and maternal urine and blood samples were taken twice during pregnancy and then again shortly after delivery to measure levels of pesticide and chemical exposure. Child height and weight were measured by trained staff at interviews that occurred at birth and when the child was approximately 1, 2, $3\frac{1}{2}$, 5, 7, and 9 years of age. The BMI was calculated as weight (kg) divided by height squared (m^2) for interviews starting at age 2. The exact age of the child also was recorded. For this article, we limit our analysis to 247 children who have four or five recorded BMI values during the observational time period as well as complete maternal exposure data. Details of the study are published elsewhere [Eskenazi et al. (2003)].

In addition to detecting common growth patterns for boys and girls in this cohort, we hope to estimate any relationships between early-life environmental factors and the growth patterns. To illustrate the differing results when using a mixture model with and without pre-processing, we fit a variety of Gaussian mixture models to the data. We used a Gaussian mixture with independent correlation (Model 1), a random intercept model (Model 2), and a random intercept and slope model (Model 3) on the BMI data and then a Gaussian mixture model after applying our proposed pre-processing to the data using independent and exponential correlation structures (Model 4 and 5) to model the BMI development patterns over time between 2 and 9 years of age, separately for boys and girls. To allow for nonlinear growth patterns, the mean structure was based on a B-spline basis of degree 2 with a knot at the median age. The number of groups for each mixture model without any baseline covariates was chosen amongst $K = 2, \ldots, 7$ to minimize the BIC.

After determining the number of groups to explain the variability in the BMI trajectories, we refit the model allowing baseline factors to impact group membership probabilities though a generalized logit function. To illustrate the behavior of the methodology, we focused on two factors, maternal pre-pregnancy BMI and maternal o,p'-DDT (dichlorodiphenyltrichloroethane) exposure during pregnancy, and estimated the model with them separately. Maternal pre-pregnancy BMI has been well studied as a factor that is strongly associated with child weight at delivery and at later ages. On the other hand, in-utero exposure to a chemical such as DDT has been hypothesized as being associated with the metabolic system that controls weight and BMI, and thus potentially related to the growth pattern over time.

Figure 4 shows the estimated mean growth patterns and classified individual trajectories based on mixture models fit to BMI data of boys in the CHAMACOS study. The chosen number of groups without considering baseline factors using BIC is $K = 4$ for Model 1 (BIC = 2152.33), $K = 4$ for Model 2 (BIC = 1994.86), $K = 3$ for Model 3 (BIC = 2304.69), $K = 4$ for Model 4 (BIC = 1555.69), and $K = 4$ for Model 5 (BIC = 1572.53). The model that utilizes random effects for

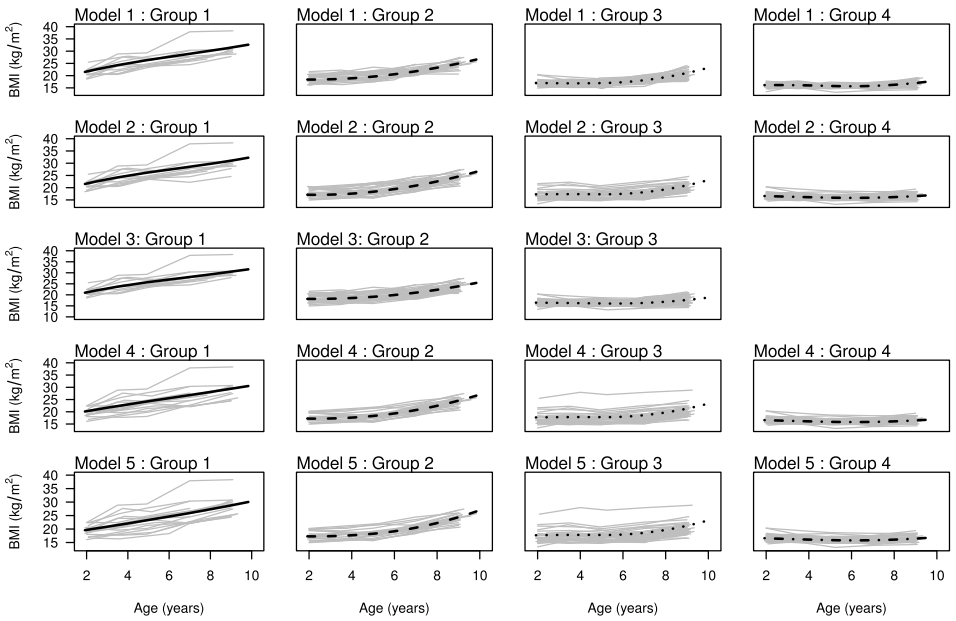

FIG. 4. *Clustered CHAMACOS BMI trajectories for boys separated according to the group assignment made by maximizing the posterior probability and group mean functions for five mixture models* (*Model* 1: *independence*, *Model* 2: *random intercept*, *Model* 3: *random slope and intercept*, *Model* 4: *pre-processed with independence*, *Model* 5: *pre-processed with exponential*) *without any baseline factors*.

both the slope and intercept (Model 3) results in the smallest number of groups because it seeks to model variability in growth pattern within clusters with the random effects. Based on BIC, Model 2 is the best model for the raw data and Model 4 has the lowest BIC for pre-processed data.

At first glance, it may seem as though all five models provide essentially the same results, but a a closer look reveals differences in group assignments for individuals. In this data example, the group means are fairly robust to the model specification but differences in group membership noticeably impact the estimated relationship between the growth pattern and baseline factors in terms of point estimates and inference.

Figure 5 presents the estimated group probabilities for maternal pre-pregnancy BMI and $\log_2$ maternal o,p'-DDT exposure for Models 1, 2, and 4. The group probability estimates for pre-pregnancy BMI are similar across the three models and indicate that higher maternal pre-pregnancy BMI is associated with the group that has the highest rate of growth and the highest average starting BMI at age two (Group 1). Thus both level and growth are associated with pre-pregnancy BMI. The relationship with level is confirmed by regressing the removed mean level on pre-pregnancy BMI and the relationship with growth is estimated based on the data with the level removed.

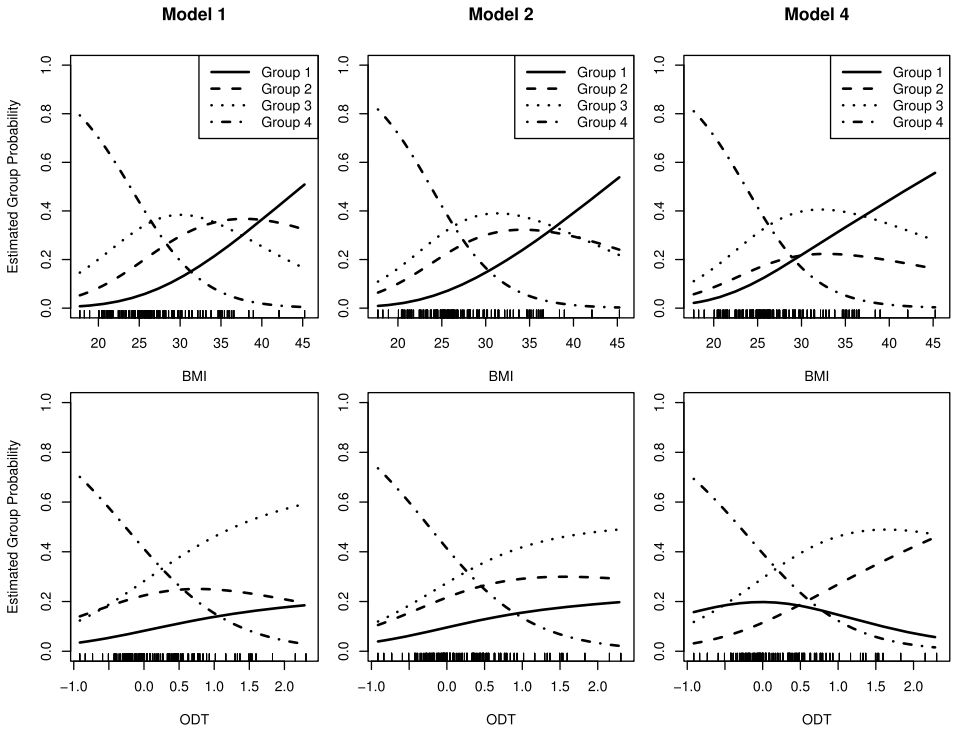

FIG. 5. *Estimated group probabilities by pre-pregnancy maternal BMI and log*-10 *o,p'-DDT* (*ODT*) *maternal exposure for three mixture models* (*Model* 1: *independent mixture*, *Model* 2: *random intercept mixture*, *Model* 4: *pre-processed independent mixture*) *fit to BMI trajectory data of CHAMACOS boys*.

However, we notice more discord between the resulting models for maternal o,p'-DDT exposure. Models 1 and 2 suggest that the probability of having the highest rate of growth (Group 1) increases with increased exposure while Model 4 suggests children with higher DDT exposure in-utero have a smaller chance of having the highest rate linear growth pattern. The probability of Group 2 also differs amongst the three models. These differences also can be highlighted through estimated relative risk estimates with corresponding confidence intervals, provided in the supplemental article [Heggeseth (2018b)]. The magnitude of the differences is small, but, in the field of environmental exposure and obesity where signal to noise ratios often are weak, it indicates a need for further research into DDT exposure and growth.

8.2. *CD*4 *counts*. Another data example that illustrates the limitations of the standard mixture models comes from an AIDS study at University of California San Francisco [Deeks et al. (1999)]. CD4 cells are a type of white blood cell that play an important role in our immune system and are the target for the HIV virus.

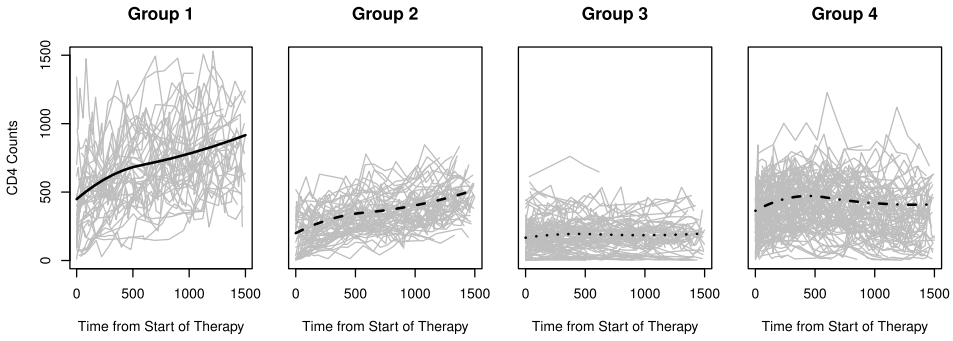

FIG. 6. *Clustered CD4 trajectories separated according to the group assignment made by maximizing the posterior probability and group mean functions for pre-processed independent mixture model* (*Model* 4) *fit to the AIDS data.*

HIV viral load is another important clinical measure as it is used to diagnose AIDS, and it is generally thought that viral load directly impacts CD4 cell count, irrespective of the count level. We can test this theory by applying the same methodological approach to CD4 counts of HIV positive patients starting at the initiation of antiretroviral therapy to investigate the relationship between average baseline viral load in the first sixty days and CD4 development over time.

We fit the same five models as we did with the CHAMACOS data with a quadratic B-spline basis with a knot at the median time and used the BIC to select the number of groups. The random intercept model (Model 2) with $K = 4$ is preferred for raw data and the independent mixture (Model 4) with $K = 4$ is the best for pre-processed data based on BIC. Figure 6 shows the individual CD4 trajectories clustered by group with overlaid mean trajectories based on a independent Gaussian mixture model fit to the pre-processed data (Model 4).

The four group mean growth patterns can be described as a steep increase over time, a gradual increase over time, flat, and a temporary increase which then stabilizes. In this context, the ideal group would be the one with a steep increase in CD4 counts over time as that would indicate improvement in health. Like the CHAMACOS data example, the group means are generally more robust to model specification as compared to the estimated group probabilities and the relationships with baseline factors. The estimated group probabilities for baseline viral loads differ substantially across these models (Figure 7). They generally agree that a steep increase of CD4 cells over time is less likely than other development patterns. However, the mixture based on pre-processed data (Model 4) is the only model that suggests a significant, complex relationship between groups and baseline viral load. A more moderate baseline viral load at the beginning of therapy increases your likelihood of the treatment not being as effective in terms of long-term growth of CD4 (flat or temporary increase only) with reference to low baseline viral load.

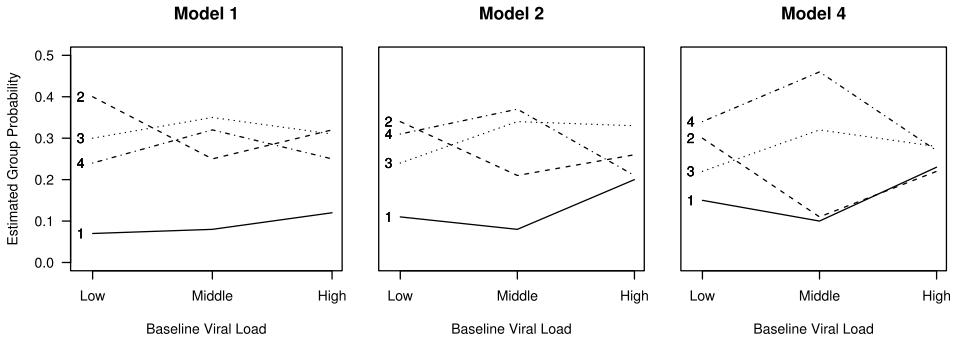

FIG. 7. *Estimated group membership probabilities by baseline viral load for three mixture models* (*Model* 1: *independence*, *Model* 2: *random intercept*, *Model* 4: *pre-processed independent*) *fit to the AIDS data. Group labels and line type correspond to the CD*4 *group means in Figure* 6.

9. Discussion. By using simulated data sets as well as real longitudinal data, we have illustrated some of the limitations of popular Gaussian mixture model specifications to study growth patterns. Longitudinal studies are expensive and time-consuming; hence, the set of available statistical methods for longitudinal data should include some approaches that are able to focus on the main feature of the trajectory, the growth over time. Standard multilevel linear models, which can model variability in the growth pattern with a hierarchy of linear models, pose challenges for nonlinear relationships. In contrast, Gaussian mixture models, which are used to flexibly model extra variability in the outcome, have the potential for allowing nonlinear relationships between baseline factors and the outcomes by introducing a group structure. However, if the raw outcome measurements are used as the response variable in a finite mixture model, the estimation procedure will not necessarily lead to groups defined by the growth pattern. The standard mixture specifications such as LCGA and GMM applied to the raw data do not directly group trajectories based on shape or change over time but rather on the feature that explains the most variability, typically the level. Unfortunately, many researchers who use this model blindly believe that the resulting groups are homogeneous in terms of growth patterns and describe groups according to the mean pattern and discuss relationships with baseline factors in terms of those patterns. The lack of knowledge about the behavior of these models may continue to result in not detecting or incorrectly estimating the relationship between a baseline factor and growth patterns.

To remedy the situation while utilizing existing software, we propose a pre-processing step to focus on the growth pattern. The proposed processing only shifts the data by removing the level and does not impact the relative magnitude within each trajectory. Since it treats the level as a nuisance, the proposed method does not require accurately modeling the intercept distribution, which can be hard in practice. One limitation with the proposed pre-processing is the consequence on

the covariance structure. Future research is required to investigate the extent of the improvement in performance with more accurate modeling of the underlying dependence.

We have shown that a pre-processing step that shifts the data prior to fitting a mixture model allows researchers to use available methods and technology to better focus on growth pattern of the longitudinal trajectory. The comparison of models with and without the processing highlights the fact that level and growth pattern may have different forces acting upon them. In many applications, both growth and level variability are important aspects to study and model in tandem. There are a myriad of approaches to explore level differences in longitudinal data, and more work is needed to increase the possible methodological approaches focusing on growth to complement the study of level and investigate the intertwined relationships between them.

Acknowledgments. We thank Kim Harley, Brenda Eskenazi, and the entire CHAMACO group for access to the childhood growth data, and we thank the anonymous reviewers and Editors for their thoughts and comments that greatly improved the manuscript. See Supplement A [Heggeseth (2018a)] for the supplementary simulations and see Supplement B [Heggeseth (2018b)] for additional CHAMACOS results.

SUPPLEMENTARY MATERIAL

Supplement A: Growth simulation (DOI: 10.1214/17-AOAS1066SUPPA; .pdf). The supplement includes a description and the results an additional simulation study that mimics real childhood growth data.

Supplement B: Additional CHAMACOS results (DOI: 10.1214/17-AOAS1066SUPPB; .pdf). The supplement includes the relative risk ratio estimates from the CHAMACOS data example.

REFERENCES

AITKIN, M., ANDERSON, D. and HINDE, J. (1981). Statistical modelling of data on teaching styles. *J. Roy. Statist. Soc. Ser. A* **144** 419–461.

ASPAROUHOV, T. and MUTHÉN, B. (2016). Structural equation models and mixture models with continuous nonnormal skewed distributions. *Struct. Equ. Model.* **23** 1–19.

BRILLINGER, D. R. (1975). *Time Series*: *Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

CARTER, M. A., DUBOIS, L., TREMBLAY, M. S., TALJAARD, M. and JONES, B. L. (2012). Trajectories of childhood weight gain: The relative importance of local environment versus individual social and early life factors. *PLoS ONE* **7** e47065.

CELEUX, G. and SOROMENHO, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification* **13** 195–212.

CUPUL-UICAB, L. A., HERNÁNDEZ-AVILA, M., TERRAZAS-MEDINA, E. A., PENNELL, M. L. and LONGNECKER, M. P. (2010). Prenatal exposure to the major DDT metabolite 1,1-dichloro-2,2-bis(p-chlorophenyl)ethylene (DDE) and growth in boys from Mexico. *Environ. Res.* **110** 595–603.

CUPUL-UICAB, L. A., KLEBANOFF, M. A., BROCK, J. W. and LONGNECKER, M. P. (2013). Prenatal exposure to persistent organochlorines and childhood obesity in the US collaborative perinatal project. *Environmental Health Perspectives* **121** 1103–1109.

CURRY, H. B. and SCHOENBERG, I. J. (1966). On Pólya frequency functions IV: The fundamental spline functions and their limits. *Journal d'Analyse Mathématique* **17** 71–107.

D'URSO, P. (2000). Dissimilarity measures for time trajectories. *Stat. Methods Appl.* **9** 53–83.

DAVIES, C. E., GLONEK, G. F. V. and GILES, L. C. (2015). The impact of covariance misspecification in group-based trajectory models for longitudinal data with non-stationary covariance structure. *Stat. Methods Med. Res.* Preprint. Available online doi:10.1177/0962280215598806.

DEEKS, S. G., HECHT, F. M., SWANSON, M., ELBEIK, T., LOFTUS, R., COHEN, P. T. and GRANT, R. M. (1999). HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: Response to both initial and salvage therapy. *AIDS* **13** 35–43.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38.

DE BOOR, C. (1976). Splines as linear combinations of B-splines. A survey. In *Approximation Theory II* (G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds.) 1–47. Academic Press, New York.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DIALLO, T. M. O., MORIN, A. J. S. and LU, H. (2016). Impact of misspecifications of the latent variance? Covariance and residual matrices on the class enumeration accuracy of growth mixture models. *Struct. Equ. Model.* **23** 507–531.

DIGGLE, P., HEAGERTY, P., LIANG, K. Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford Univ. Press, New York.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA.

EROSHEVA, E. A., MATSUEDA, R. L. and TELESCA, D. (2014). Breaking bad: Two decades of life-course data analysis in criminology, developmental psychology, and beyond. *Annual Review of Statistics and Its Application* **1** 301–332.

ESKENAZI, B., BRADMAN, A., GLADSTONE, E. A., JARAMILLO, S., BIRCH, K. and HOLLAND, N. (2003). CHAMACOS, a longitudinal birth cohort study: Lessons from the fields. *Journal of Children's Health* **1** 3–27.

ESKENAZI, B., HARLEY, K., BRADMAN, A., WELTZIEN, E., JEWELL, N. P., BARR, D. B., FURLONG, C. E. and HOLLAND, N. T. (2004). Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population. *Environmental Health Perspectives* **112** 1116–1124.

ESKENAZI, B., GLADSTONE, E. A., BERKOWITZ, G. S., DREW, C. H., FAUSTMAN, E. M., HOLLAND, N. T., LANPHEAR, B., MEISEL, S. J., PERERA, F. P., RAUH, V. A., SWEENEY, A., WHYATT, R. M. and YOLTON, K. (2005). Methodologic and logistic issues in conducting longitudinal birth cohort studies: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environmental Health Perspectives* **113** 1419–1429.

EUBANK, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Dekker, New York, NY.

EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions*. Chapman & Hall, London.

EVERITT, B. S., LANDAU, S., LEESE, M. and STAHL, D. (2011). *Cluster Analysis*, 5th ed. Wiley, London.

FENG, Z. D. and MCCULLOCH, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 609–617.

FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

GARDEN, F. L., MARKS, G. B., SIMPSON, J. M. and WEBB, K. L. (2012). Body mass index (BMI) trajectories from birth to 11.5 years: Relation to early life food intake. *Nutrients* **4** 1382–1398.

GRAY, G. (1994). Bias in misspecified mixtures. *Biometrics* **50** 457–470.

GRÜN, B. and LEISCH, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* **28** 1–35.

HEGGESETH, B. (2018a). Supplement to "How Gaussian mixture models might miss detecting factors that impact growth patterns." DOI:10.1214/17-AOAS1066SUPPA.

HEGGESETH, B. (2018b). Supplement to "How Gaussian mixture models might miss detecting factors that impact growth patterns." DOI:10.1214/17-AOAS1066SUPPB.

HEGGESETH, B. C. and JEWELL, N. P. (2013). The impact of covariance misspecification in multivariate Gaussian mixtures on estimation and inference: An application to longitudinal modeling. *Stat. Med.* **32** 2790–2803.

HEO, M., FAITH, M. S., MOTT, J. W., GORMAN, B. S., REDDEN, D. T. and ALLISON, D. B. (2003). Hierarchical linear models for the development of growth curves: An example with body mass index in overweight/obese adults. *Stat. Med.* **22** 1911–1942.

HUANG, Y., CHEN, J. and YIN, P. (2017). Hierarchical mixture models for longitudinal immunologic data with heterogeneity, non-normality, and missingness. *Stat. Methods Med. Res.* **26** 223–247.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JENNRICH, R. I. and SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805–820.

JONES, B. L., NAGIN, D. S. and ROEDER, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol. Methods Res.* **29** 374–393.

LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* **11** 1–18.

LU, X. and HUANG, Y. (2014). Bayesian analysis of nonlinear mixed-effects mixture models for longitudinal data with heterogeneity and skewness. *Stat. Med.* **33** 2830–2849.

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models*: *Inference and Applications to Clustering*. Dekker, New York.

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

MENDEZ, M. A., GARCIA-ESTEBAN, R., GUXENS, M., VRIJHEID, M., KOGEVINAS, M., GOÑI, F., FOCHS, S. and SUNYER, J. (2011). Prenatal organochlorine compound exposure, rapid weight gain, and overweight in infancy. *Environmental Health Perspectives* **119** 272–278.

MÖLLER-LEVET, C., KLAWONN, F., CHO, K. H. and WOLKENHAUER, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *Proceedings of the Fifth International Conference on Intelligent Data Analysis* (M. R. Berthold, H.-J. Lenz, E. Bradley and C. Borgelt, eds.) 330–340.

MORIN, A. J. S. and MARSH, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers? Multidimensional profiles of effectiveness. *Struct. Equ. Model.* **22** 39–59.

MORIN, A. J. S., MAÏANO, C., MARSH, H. W., NAGENGAST, B. and JANOSZ, M. (2013). School life and adolescents' self-esteem trajectories. *Child Dev.* **84** 1967–1988.

MUTHÉN, B. and ASPAROUHOV, T. (2009). Multilevel regression mixture analysis. *J. Roy. Statist. Soc. Ser. A* **172** 639–657.

MUTHÉN, L. K. and MUTHÉN, B. O. (1998–2010). Mplus User's Guide, 6th ed., Los Angeles.

MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.

MUTHÉN, B., BROWN, C. H., MASYN, K., JO, B., KHOO, S. T., YANG, C. C., WANG, C. P., KELLAM, S. G., CARLIN, J. B. and LIAO, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3** 459–475.

NAGIN, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods* **4** 139–157.

NAGIN, D. S. and ODGERS, C. L. (2010a). Group-based trajectory modeling (nearly) two decades later. *J. Quant. Criminol.* **26** 445–453.

NAGIN, D. S. and ODGERS, C. L. (2010b). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology* **6** 109–138.

PARK, T., YI, S.-G., KANG, S.-H., LEE, S., LEE, Y.-S. and SIMON, R. (2003). Evaluation of normalization methods for microarray data. *BMC Bioinform.* **4** 1–13.

PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **185** 71–110.

PICKLES, A. and CROUDACE, T. (2010). Latent mixture models for multivariate and longitudinal outcomes. *Stat. Methods Med. Res.* **19** 271–289.

PROUST-LIMA, C., PHILIPPS, V., DIAKITE, A. and LIQUET, B. (2014). lcmm: Estimation of extended mixed models using latent classes and latent processes. R package version 1.6.4.

PRYOR, L. E., TREMBLAY, R. E., BOIVIN, M., TOUCHETTE, E., DUBOIS, L., GENOLINI, L., XUECHENG, C., FALISSARD, B. and CÔTÉ, S. M. (2011). Developmental trajectories of body mass index in early childhood and their risk factors: An 8-year longitudinal study. *Archives of Pediatrics & Adolescent Medicine* **165** 906–912.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757.

SCHLATTMANN, P. and BÖHNING, D. (1997). On Bayesian analysis of mixtures with an unknown number of components. Contribution to a paper by S. Richardson and PJ Green. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **59** 782–783.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHUMWAY, R. H. and STOFFER, D. S. (2010). *Time Series Analysis and Its Applications*: *With R Examples*. Springer Science & Business Media, New York.

SINGER, J. D. and WILLETT, J. B. (2003). *Applied Longitudinal Data Analysis*: *Modeling Change and Event Occurrence*. Oxford Univ. Press, New York, NY.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

VALVI, D., MENDEZ, M. A., MARTINEZ, D., GRIMALT, J. O., TORRENT, M., SUNYER, J. and VRIJHEID, M. (2012). Prenatal concentrations of polychlorinated biphenyls, DDE, and DDT and overweight in children: A prospective birth cohort study. *Environmental Health Perspectives* **120** 451–457.

WARNER, M., AGUILAR SCHALL, R., HARLEY, K. G., BRADMAN, A., BARR, D. and ESKENAZI, B. (2013). In utero DDT and DDE exposure and obesity status of 7-year-old Mexican-American children in the CHAMACOS cohort. *Environmental Health Perspectives* **121** 631–636.

WARNER, M., WESSELINK, A., HARLEY, K. G., BRADMAN, A., KOGUT, K. and ESKENAZI, B. (2014). Prenatal exposure to dichlorodiphenyltrichloroethane and obesity at 9 years of age in the CHAMACOS study cohort. *Am. J. Epidemiol.* **179** 1312–1322.

WEDEL, M. (2002). Concomitant variables in finite mixture models. *Stat. Neerl.* **56** 362–375.

DEPARTMENT OF MATHEMATICS AND STATISTICS
WILLIAMS COLLEGE
WILLIAMSTOWN, MASSACHUSETTS 01267
USA
E-MAIL: bch2@williams.edu

DIVISION OF BIOSTATISTICS
AND DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: jewell@berkeley.edu