Covariance Models for Latent Structure in Longitudinal Data [1]

Marc A. Scott
New York University
and
Mark S. Handcock
University of Washington

Working Paper no. 14
Center for Statistics and the Social Sciences
University of Washington
Box 354322
Seattle, WA 98195-4322, USA

December 2000

[1]Marc A. Scott is Assistant Professor of Education, Department of Education, New York University (Email: ms1098@nyu.edu; Web: www.educ.nyu.edu/mscott), and Mark S. Handcock is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322 (Email: handcock@stat.washington.edu; Web: (www.stat.washington.edu/handcock. This research was supported by the National Science Foundation under grant SES-0088061 and and partially supported by the Russell Sage Foundation and the Rockefeller Foundation. The authors thank Annette D. Bernhardt and Martina Morris whose discussions and insights over the years are deeply embedded in this paper. We also thank James Ramsay for helpful discussions on this topic.

Abstract

We present several approaches to modeling latent structure in longitudinal studies when the covariance itself is the primary focus of the analysis. This is a departure from much of the work on longitudinal data analysis, in which attention is focused solely on the cross-sectional mean and the influence of covariates on the mean. Such analyses are particularly important in policy-related studies, in which the heterogeneity of the population is of interest. We describe several traditional approaches to this modeling and introduce a flexible, parsimonious class of covariance models appropriate to such analyses. This class, while rooted in the tradition of mixed effects and random coefficient models, merges several disparate modeling philosophies into what we view as a hybrid approach to longitudinal data modeling. We discuss the implications of this approach and its alternatives especially on model interpretation. We compare several implementations of this class to more commonly employed mixed effects models to describe the strengths and limitations of each. These alternatives are compared in an application to long-term trends in wage inequality for young workers. The findings provide additional guidance for the model formulation process in both statistical and substantive senses.

Contents

List of Figures

1 Introduction and Motivation

An increasing number of social and behavioral science studies collect information from subjects at several points in time. These longitudinal studies enable researchers to study changes in the phenomena of interest over the life-course of the subjects. At each observation time, at least one response, such as wages earned or the occurrence of a meaningful event, such as graduation from college, is recorded. As with regression, one may collect explanatory covariates in the hope that differences in these inputs will be associated with different levels of response. Each subject is thus associated with their own time series of responses and a corresponding set of potentially time-varying explanatory covariates. Models for longitudinal data attempt to relate those individual time series to an overall group process.

The focus on either individual or group processes plays a key role in how one models longitudinal data. For example, if we are modeling a continuous response, Y, in terms of explanatory covariates, X, then the familiar linear model, for individual i,

$$Y_i = X_i\beta + \epsilon_i \tag{1}$$

could be adopted, but Y_i and ϵ_i would be n_i-vectors, where n_i is the number of observations on individual i. Similarly, X_i would be of dimension $n_i \times p$, where p is the number of explanatory covariates. Note that we are modeling a response *vector*, yet this distinction is not made explicitly with our notation. Alternatively, the model may be written

$$Y_i(t) = X_i(t)\beta + \epsilon_i(t),$$

where the index t identifies a specific element of the response vector Y_i. If we were to proceed with a classical multiple regression, stacking the responses by individual and then by observation within individual, we would obscure an important feature of longitudinal data; namely, we know that some set of observations come from the same individual. And observations within the same individual may be correlated due to unobserved individual characteristics. To see this, let us return to the notation in (1), but now

$$Y_i = X_i\beta + \alpha_i + \epsilon_i^*, \tag{2}$$

where α_i is an unobserved scalar trait for subject i, and the residual variation is mean zero and uncorrelated with the unobserved process, and $E(\epsilon_i^*(t)\epsilon_i^*(t')) = 0$ for $t \neq t'$. Since we do not observe α_i, we tend to use (1) when the underlying process is accurately described by (2), so the residual variation structure ϵ_i is really $\alpha_i + \epsilon_i^*$. The unobserved trait induces a correlation within individual i, since

$$E(\epsilon_i(t)\epsilon_i(t')|\alpha_i) = E((\alpha_i + \epsilon_i^*(t))(\alpha_i + \epsilon_i^*(t'))|\alpha_i) = \alpha_i^2,$$

and $\alpha_i^2 \neq 0$ in general. Note that the unobserved α_i may not correspond to a single measurable characteristic; instead it proxies for all unobserved characteristics.

There are several different ways to think about the correlation structure in longitudinal data. The different perspectives are induced by the nature of the unobserved trait and its relationship to the covariates and residual variation. If substantive interest is on the effects of the covariates on the response averaged over the population then models are usually formulated for the mean response averaged over the unobserved traits. Broadly speaking, the correlation structure is modeled as a nuisance parameter, and regression coefficients represent *population-average* effects (Liang and Zeger 1986, Zeger and Liang 1986, Prentice 1988). Alternatively, individual differences may be of interest, and can be modeled directly as latent variables; for these *individual-specific* models, regression parameters are to be interpreted conditionally on the value of the subject's latent variable. We will discuss these different approaches, certain variations thereof, and their implications in subsequent sections. Along the way, we will introduce a class of models for longitudinal data that merges these two approaches in a new hybrid form, which is conceptually linked to principal components and factor analysis. First, we introduce the substantive problem that motivates this new formulation.

In labor market economics, a rise in cross-sectional measures of wage inequality that began in the 1970s and has persisted into the 1990s is well-documented (Levy and Murnane 1992; Danziger and Gottschalk 1993; McMurrer and Sawhill 1998). This means that there are greater numbers of workers making more and less than ever before. And for many groups of workers, wages have remained stagnant over time. This stagnation is due in part to a disproportionate growth in the lower tail of the wage distribution. Using data from two young adult cohorts in the National Longitudinal Survey (NLS), we find, for example, that 30-35 year old white men have a mean wage of \$17.78 in 1979, while this figure is \$14.27 per hour for a similarly aged group in 1992 (inflation adjusted, 1999 dollars). A measure of inequality is the variance in outcomes; the variance of the logged wages increased 44% over the same period.

This dramatic rise has prompted researchers to look more closely at trends in inequality over the lifecourse of a worker. Cross-sectional data can document a rise in inequality, but since each cross-section is a random sample from the population, one cannot conclude that the same people are making the higher wages in each period. Statements such as "the rich are getting richer while the poor are getting poorer" cannot be definitively made. But longitudinal data can be used to address this type of question. To couch this in labor economic terms, we would like to examine two competing hypotheses that explain the growth in inequality:

I. Wages have become more volatile

II. Wages have become more stratified over time, indicating a reduction in economic mobility

The scenarios are illustrated in Figures 1 - 3 below. Figure 1, represents an economy in which individual "profiles" fan out over time, but not excessively. This is our stylized image of a past economy; in Figures 2 and 3, we changed the covariance structure to reflect at least a doubling of process variance, but we do this in very different ways. In Figure 2, the structured variation has become more stratified, but the residual process is left unchanged. In Figure 3, the structured variation is identical to that used in Figure 1, but the residual variance of the process has been greatly increased. This last figure may seem exaggerated, but the average variation between individuals is actually a bit smaller than in Figure 2.

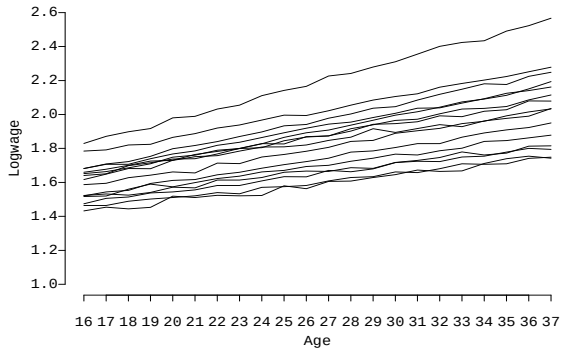

Figure 1: Stylized wage trajectories for a less stratified economy

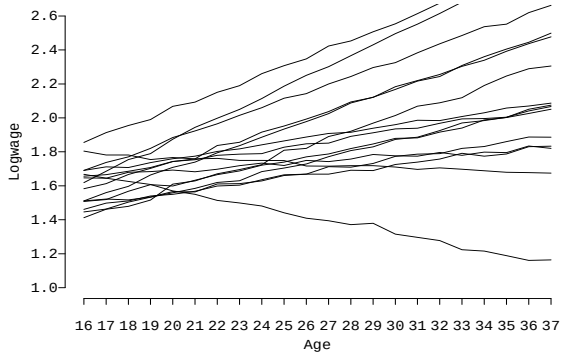

Figure 2: Stylized wage trajectories for a more stratified economy

Based on the figures, the difference between increased stratification (Figure 2) and increased volatility (Figure 3) seems transparent, but in a real application, both hypotheses

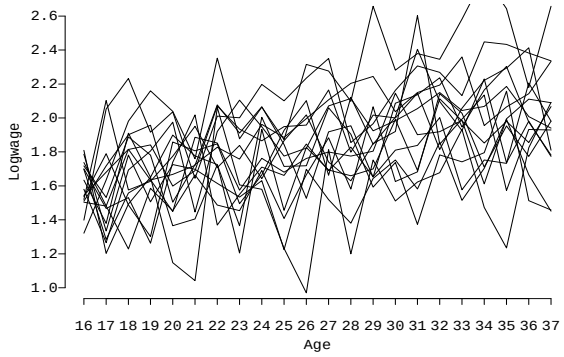

Figure 3: Stylized wage trajectories for a more volatile economy

may be true and the differences may be more subtle. Each possibility has a different substantive interpretation, so sorting out the extent to which each hypothesis describes the changes in wage structure is very important and will have different implications in terms of policy.

To investigate the two hypotheses, Bernhardt, et al. (1997), Gottshalk and Moffitt (1994), Haider (1996) and Baker (1997) decompose the wage into permanent and transient components as follows. Let

$$w(t) = p(t) + u(t), \tag{3}$$

where w is the wage, p is its permanent portion, and u is a residual variation term, capturing short-term, or transient variation. For a specific worker, $p(t)$ can be thought of as their mean wage at time t, with residual variation $u(t)$. Assuming independence of $p(t)$ and $u(t)$, we have that

$$\text{Var}(w(t)) = \text{Var}(p(t)) + \text{Var}(u(t)),$$

and the two hypotheses can be differentiated through this variance decomposition: a rise in wage variance must involve a rise in at least one of the two variance components. Greater stratification implies an increase in the first term, while greater volatility involves the second. If we had a substantial number of observations for each individual, we could estimate $p(t)$ using separate regressions for each. The distribution of these predicted curves would represent permanent variation, while the residuals represent transient variation. The hypotheses of interest describe differences in wage trajectories without any socioeconomic controls, so the only explanatory covariates we include in this analysis are functions of time.

This last point warrants further explanation. Socioeconomic variables such as level of schooling, parent's education, and industry of employment, capture expected returns to individual (supply-side) and employer (demand-side) characteristics. For example, there may be changes in the mean return to obtaining a high school degree which reflects the value of that set of skills in the labor market. Including socioeconomic covariates also

controls for compositional shifts in the labor market. The growth in a specific sector of the economy could induce growing inequality if that sector is typically associated with lower wages. But all of these explanations are necessarily focused on the permanent portion of the wage trajectory, since volatility is associated with residual, rather than mean effects. The first stage in any analysis of wage inequality is the accurate documentation of the growth in inequality, and how it is apportioned with regard to permanent and transient components. Thus, our focus is first on covariance structure, and not on the socioeconomic covariates that might "explain" the structure.

In many longitudinal studies, there are a relatively small number of observations per individual, so estimating separate regressions to assess wage inequality is infeasible. A variance components model (Searle, et al. 1992) using (3) can partition the variance into long-term, permanent variation and short-term, transitory variation. We note that the distinction between long- and short-term trends is fundamentally about economic mobility. The variation between the individuals' permanent components is a measure of mobility relative to one's peers, and a variance components analysis allows one to evaluate this important economic issue.

In matters with such strong policy implications, proper specification of a model that describes the components of variation is crucial. What may be less apparent is the role of the covariance structure in such an analysis. If we generalize the basic model (2) for longitudinal data to allow for more complex individual characteristics, we get the standard mixed effects model (Diggle, Liang and Zeger 1994),

$$Y_i = X_i\beta + Z_i\delta_i + \epsilon_i. \tag{4}$$

We have introduced a random effects component, $Z_i\delta_i$, in which Z_i is an $n_i \times q$ known design matrix, and δ_i is a q-vector of unknown (latent) variables. It is often assumed that δ_i are mean zero multivariate Gaussian. Under this assumption, it is seen that $E(Y_i) = X_i\beta$, while $E(Y_i|\delta_i) = X_i\beta + Z_i\delta_i$. This distinction is important. The latter approach asserts that individuals differ from the population average response in a systematic manner, which is dependent on some latent characteristics. In our application, the growth in wage inequality and evidence for the two competing hypotheses are features of the latent process, $Z_i\delta_i$, not the population-average process $X_i\beta$. If $\delta_i \sim N_q(0, G)$ and $\epsilon_i \sim N_{n_i}(0, R)$, independently, then

$$\mathrm{Cov}(Y_i(t), Y_i(t')) = \{Z_i G Z_i' + R\}_{tt'}.$$

Note that the $X_i\beta$ term is not included. In other words, the covariance structure of the responses, rather than the mean structure, captures the nature of individual differences, and thus inequality, in the labor market. Strictly speaking, the covariance structure captures all

extra-mean variation. Since for the mean we employ only time as an explanatory covariate, all of the contributions to inequality are expressed in the covariance. This establishes a baseline level of inequality that we can later use additional covariates (such as education) to explain.

Models for this covariance structure that can differentiate between hypotheses 1 and 2 will be developed in subsequent sections.

1.1 Data

As alluded to above, we will anchor our presentation by using an example from labor market economics, where proper modeling of covariance structure is of paramount importance. We will be investigating two datasets from the NLS. The first, or original cohort, is a representative sample of young men aged 14-21 first interviewed in 1966 and interviewed annually for the next fifteen years (with the exception of 1972, 1974, 1977 and 1979). The second dataset began with a comparable sample of young men in 1979 who have been interviewed yearly since then for fifteen additional years. For comparability between cohorts, we selected only non-Hispanic whites, with resulting sample sizes of 2,614 and 2,373 respectively for the original and recent cohorts. For a detailed description of these datasets and their comparability, see Bernhardt, et al. (1997). According to Topel and Ward (1992), "the first 10 years of a career will account for 66 percent of lifetime wage growth for male high school graduates and almost exactly the same fraction of lifetime job changes," so it is important to understand trends manifesting themselves in this early period.

In this paper, we present several different ways to model covariance structure with the ultimate goal of addressing questions such as those posed in hypotheses 1 and 2. One of these methods is novel in the literature, so it is be developed in some depth. We begin by discussing several different philosophical perspectives to longitudinal data modeling in Section 2. To address hypotheses like the ones just presented, we argue that a different modeling philosophy is necessary; we develop a hybrid framework with this in mind in Section 3 and illustrate it in Section 4. We apply more traditional models to our labor market data in Section 5 and discuss the strengths and weaknesses of each approach, including the substantive implications of each choice. Section 6 summarizes the discussion and suggests future directions of research.

2 Alternative Modeling Philosophies

The choice of modeling framework should depend on the substantive question of interest. For example, in many medical applications, one may be focussed on how a treatment affects

the population as a whole. However, if there are potentially serious risks involved in treatment, the distribution of outcomes, including information about the extremes, may be of importance. Along with many modeling paradigms comes a modeling philosophy, focussed on the primary goals of the research. We now describe several philosophies in longitudinal data modeling.

2.1 Population-average analysis

Population average models focus on describing the population, rather than individuals within it. Much as in classical regression, the mean response is modeled conditional on the observed covariates. In a linear model, $E(Y|X) = X\beta$, the parameter β describes how changes in the components of X affect the overall population. With longitudinal data, we have seen that the covariance of the responses within an individual influences the response trajectory. For some problems, that covariance structure is effectively a nuisance parameter—it must be included in the model but is of no intrinsic interest in and of itself. Generalized Estimating Equations (GEE) is a methodology that produces consistent estimates of population-average parameters even when the covariance structure is misspecified (Liang and Zeger 1986, Zeger and Liang 1986, Prentice 1988). This technique allows one to pursue the population-average approach to modeling, while accounting for the dependencies due to the longitudinal nature of the data. Since the covariance is viewed as secondary, the method does not yield a variance components analysis, which one might use to address our labor market hypotheses, for example.

2.2 Individual-specific analysis

Individual-specific effect models consist of two key components: the fixed effects, which capture gross differences between individuals based on differences in their explanatory covariates; and the random effects, which reflect the influence of unobserved covariates. These so-called "unobserved" covariates are just a device to capture unexplained but systematic variation in outcomes. Typically there is no single covariate, such as "motivation," that would replace the individual effects in our model, were we able to measure it. Rather, after controlling for what was measured, some systematic differences between individuals are likely to exist for a variety of reasons. Because we are looking at longitudinal data, we can verify that some differences seem to persist throughout an individual's life course, and that these are not simply random disturbances.[1]

[1]Note that this modeling philosophy is agnostic toward the substantive interpretation assigned to the systematic differences.

Under model (4), the fixed effects are captured by the $X_i\beta$ term, while the random effects are modeled via $Z_i\delta_i$. The δ_i vector is indexed with an i to reflect the fact that every individual is expected to have their own value for this "parameter." These models are also referred to as random coefficient models (Longford 1993) because the coefficient on the Z_i terms is allowed to vary. These coefficients introduce extra-mean variation into the response in a systematic manner mediated by the design matrix Z_i.

We interpret these models conditional on the individual specific effects, so we are modeling $E(Y_i|X_i, Z_i, \delta_i)$, rather than $E(Y_i|X_i)$. Using model (4), the interpretation of the fixed effects parameters shifts to the following. Given the individual specific effect δ_i, the expected response for individual i is $X_i\beta + Z_i\delta_i$. We are making statements about individuals, not populations; the regression coefficients reflect this distinction and should be interpreted in this conditional manner. In the standard linear mixed effects model in which all random components are assumed Gaussian, the distinction between population average and individual specific modeling is more philosophical, as the models and their parameter estimates are identical. This is not the case, for a generalized linear mixed model (GLMM, McCulloch 1997, Hu, et al. 1998, Crouchley and Davies 1999).

What may be less immediately apparent about this shift toward an individual-specific perspective is that the parameters that define the distribution of the effects often represent meaningful components of variation. For example, if δ_i is a scalar and Z_i is a column of ones, then the individual differences are being modeled as shifts in the intercept. This implies that the differences between individuals are constant over the lifecourse. The variance of the random effect δ_i is an important model parameter. If it is large, then large differences between individuals exist and persist throughout the lifecourse; if it is small, they do not. The ability to interpret a variance component in terms of a substantive question is a key feature of individual-specific modeling.

Note that not all mixed effects models are oriented toward meaningful variance components analyses. Beyond the fixed effects, the variation is modeled in the random effects and in the residual variation structure. In model (4), $\delta_i \sim N_q(0, G)$ and $\epsilon_i \sim N_{n_i}(0, R)$, and the residual variation structure, R, can be made arbitrarily complex. There is often a tension, in terms of modeling, between these two components. ARMA models (Box and Jenkins 1976) can capture a substantial portion of the within-individual correlation, but they do so via parameters that do not take on individual-specific values. For example, the correlation between observations may be given by ρ, but ρ does not vary between individuals. So we know how variation occurs, but we cannot directly use it to position a curve above or below the mean trajectory.

Jones (1990) discusses this model formulation issue by comparing a classical random

growth curve model to an AR(1) model for that same structure. He finds that these two approaches typically compete with each other in terms of explaining the variation in the data. We tend to favor models that emphasize structured variation in the $Z_i\delta_i$ term, because these provide direct summaries of differences between individuals.

In sum, mixed effects models may be based on an individual-specific philosophy, but they are not required to do so. Thus, care must be given in the model formulation process as to which philosophical perspective to adopt.

2.3 Latent Curve Models

A related, but philosophically different approach to modeling longitudinal trajectories was developed by Meredith and Tisak (1990).[2] They outline a framework in which each response is a weighted average of a fixed set of curves:

$$Y_i(t) = \sum_k \omega_{ik}\phi_k(t) + \epsilon_i(t), \tag{5}$$

where $Y_i(t)$ represents the response for the i^{th} individual, ω_{ik} is the individual-specific coefficient associated with the k^{th} latent curve $\phi_k(t)$, and $\epsilon_i(t)$ is the residual process. The ϕ_k capture the shape and magnitude of the variation, and the ω_{ik} allow individuals to differ systematically, much in the same way that random coefficients do in mixed effects models. In the above formulation, the mean process will be a specific weighted sum of the latent curves, but it could just as well be parameterized separately, as in the fixed effects portion of a mixed model. For the remainder of this discussion, we will ignore the mean of the process, or assume it is identically zero.

If the latent curves are known, then the formulation is similar to a mixed effects model. If we stack the ϕ_k as columns of a design matrix $Z(t) = \{\phi_1(t), \phi_2(t), \ldots, \phi_K(t)\}$ at T design points $t_1, t_2, \ldots, t_T$, then we can estimate a model:

$$Y_i(t) = Z(t)\delta_i + \epsilon_i(t),$$

again, ignoring the mean process. But when model (5) was originally presented, it was assumed that the latent curves were not known and would be estimated directly from the data. With a few additional assumptions, this would be a factor analysis, which is a particular decomposition of the covariance into structured and residual variation. The former are captured in the factor loadings, while the latter are summarized by the specific variances. The large variability inherent in covariance estimation prompted researchers to impose smoothness constraints on the curves (Rice and Silverman 1991). A basic premise of the new model

[2]We also refer the reader to Raykov (2000), in which latent curve modeling is developed using the Structural Equation Modeling (SEM) approach. SEM emphasizes covariance structure in the model formulation.

that we will propose is that smooth latent curves can go a long way toward describing systematic variation in longitudinal data.

In practice, the latent curve model described above cannot be estimated without further assumptions. If the ϕ_k are known, then this can be estimated as a random growth curve mixed effects model. If they are left completely unspecified, we have a factor analysis formulation. Both of these model-based approaches avoid some technical problems that arise when an estimate of the full unspecified covariance matrix is required.[3] Moving beyond these traditional models will actually open up a whole new way to think about longitudinal data modeling, and we develop this alternative approach at length in Section 3.

2.4 Latent Class Models

So far, we have discussed models for which differences between individuals are expressed as an offset from the mean value in shifts that come from a continuous distribution. For example, the random coefficients in mixed effects models come from a multivariate Gaussian distribution, yielding a wide (actually, infinite) variety of outcomes. If the differences "clump" together in a natural way, then it might make sense to restrict the variation to a finite set of possibilities, in which each represents a clump or cluster of similar outcomes in the population. This is the approach taken by latent class analysis (Clogg 1995).[4] The analyst divides the population into K distinct classes, and typically any variation that exists within a class is of secondary interest.[5] The model can be represented as:

$$Y_i(t) = X_i(t)\beta_k + \epsilon_i(t), \text{ when } C_i = k, \tag{6}$$

where the random variable C_i captures the latent class membership, and β_k represents the regression coefficient for class k.[6] By allowing the regression coefficients to take on several

[3]Missing data at the individual-record level is a key challenge in direct estimation of the covariance matrix but is surmountable. Many algorithms have been developed to estimate pairwise covariances, $\sigma_{tt'} = E((Y_t - \bar{Y}_t)(Y_{t'} - \bar{Y}_{t'}))$, from all available observations. Unfortunately, the resulting full covariance matrix may not be positive semi-definite (Jolliffe 1986). Beale and Little (1975) impose a multivariate normality assumption and estimate the covariance parameters using the method of maximum likelihood. But it is uncertain how robust these methods are to violations of the normality assumption. Other methods have been developed, with similar limitations. For further discussion see Devlin, et al. (1981), Locantore, et al. (1999) and Arminger and Sobel (1990).

[4]Strictly speaking, latent class models capture discrete outcomes, while latent trait models are employed with continuous responses. We view latent class modeling conceptually as a form of mixture modeling (Banfield and Raftery 1993, Muthen and Shedden 1999), and thus make no distinction as to response type for these models.

[5]The model-based clustering work of Banfield and Raftery (1993) is an exception; they try to capture the within-class variation using several different forms for the covariance.

[6]Comparing this to a standard mixed effects model, it is interesting to note that one can view the β_k as random coefficients governed by a multinomial distribution, in which case $E(\beta_k)$ may be different from zero, and $X_i(t)$ assumes the role of both mean and random effects design matrices ($X_i(t)$ and $Z_i(t)$ in (4)). Such

different values, a set of distinct trajectories can be captured, assuming that the data support them. In formulating such models, one typically models membership in one of the K classes as a random process following a multinomial distribution, so individuals are members of exactly one class. Several features of the population are documented in this approach: the shape of the different trajectories, and the probability of membership in each. Extensions of this approach by Muthen and Shedden (1999) and Roeder, et al. (1999) involve estimating a multinomial "choice" model for class membership. For example, we might assume that membership is based on a multinomial logit model in which some subset of the explanatory covariates play a role:

$$\text{logit}\left[P(C_i = k|X_i)\right] = \theta_k + X_i\beta_k, \quad k = 1, \ldots, K,$$

where X_i are explanatory variables influencing membership and θ_k are scalars (for identifiability, we would fix $\theta_1 = 0$ and $\beta_1 = 0$). The full model combines this choice model with a model for the response, conditional on the class membership and the explanatory covariates.

This modeling philosophy focuses on identifying subgroups in the data with similar mean structures. However there are close links to approaches that model the covariance. To see this consider the model as the number of latent classes increases. If there is only one latent class, then this approach is equivalent to regression and is not modeling covariance at all. As the number of classes increases more of the covariation in profiles is attributed to the classification. If there are a large number of classes then the covariation in the profiles is largely explained by the classification and the within-class variation will be reduced. There is a clear tradeoff between mean and covariance modeling as the number of classes increases. However, the present latent class models do not attempt to identify features shared by the entire population (the features are by definition disjoint), and we do not consider them here.

2.5 Discussion

In sum, there are several different ways to think about modeling longitudinal data. One can concentrate on the population average and represent the covariance structure as a foil. Or, one can model individual differences directly by imposing a strict structure on how these differences arise. A relatively general framework is to decompose variation into the sum of curves with different weights. This could be in the form of a factor analysis, but a model-based approach to this is preferred over analyses based on the directly estimated covariance matrix. In some instances, one can separate the variation into similar clusters, with an explicit model for how these are determined by explanatory covariates.

random coefficients follow a discrete, rather than the more classical continuous, distribution. We thank an anonymous reviewer for suggesting this interpretation.

All of these are good ideas, depending on the substantive issues to be addressed. We would use the second and third to explicitly model variation in populations that is quite general in form and consider the fourth when natural clusters are apparent.

3 A hybrid model

The theoretical approaches of Section 2 each have their value and place. In our labor economic example, we wish to extract permanent and transient variance components. We want the permanent component to reflect features of the whole population while still allowing the expression of individual differences. A modeling class that identifies a common, population-level pattern as distinct from short-term effects requires a hybrid modeling philosophy, since both population-average and individual-specific approaches are being employed. When used to address hypotheses 1 and 2, such an approach will provide a highly interpretable and novel variance components decomposition.

3.1 The proto-spline model class

In Scott (1998) and Scott and Handcock (2000), we introduced the hybrid proto-spline class of heterogeneity models. Motivated by a longitudinal study of wage growth, we formulated a class of models that capture long- and short-term features of the covariance structure. The models use a latent curve formulation to identify long-term patterns of variation, and they yield a meaningful variance components decomposition. The proto-spline class is distinguished by the data-adaptive manner in which the curves are estimated. We will now describe this class in detail.

The proto-splines class is derived from the model class (5) of Meredith and Tisak (1990),

$$Y_i(t) = \mu_i(t) + \sum_{k=1}^{K} \omega_{ik}\phi_k(t) + \epsilon_i(t). \tag{7}$$

We have added a general mean process and changed the approach to modeling ω_{ik} as follows. What was formerly an unconstrained individual specific weight, we will now view as a random coefficient from some known distribution. In addition, we will specify a functional form for the $\phi_k(t)$. However, only a functional *form* and not a specific function is necessary for estimation of the proto-spline class.

We restrict the ϕ_k to be orthonormal.[7] This parallels the orthogonality employed in principal components analysis and allows us to interpret each curve's contribution as mathematically distinct from the others. We assume that the random coefficients, ω_{ik} are independent

[7]By orthonormal it is meant that $\int_{-\infty}^{\infty} \phi_j(t)\phi_k(t)dt = \mathcal{I}(j = k)$. For a discussion of orthogonality and spline bases see De Boor (1998).

(for different k), which further uncouples the latent curves. This formulation mirrors many psychological and behavioral models in which a response is the combination of several orthogonal shocks to the system. For the proto-spline class, the way the orthogonality of the ϕ_k is maintained is a departure from techniques used in principal component analysis and in Rice and Silverman (1991), in that no external constraints are placed on the estimation procedure.

Consider first the case where the stochastic variation can be described by one curve, ϕ_1 (this is a single latent curve model). We must specify the functional form of ϕ_1, and this is done by choosing an appropriate functional space.[8] For example, we can assume that ϕ_1 is a cubic spline with knots at four equispaced time points. Cubic splines are smooth functions that have a tremendous degree of flexibility in terms of the possible set of shapes that they describe (see Green and Silverman 1994 for details). In our theoretical development (Scott and Handcock 2000), we employed the cubic spline function space because of its smoothness features and flexibility, and this is where the "spline" portion of the proto-spline class is derived. We are not restricted to the class of cubic splines; for example, we can specify that ϕ_1 has the form of a jump process, well-described by wavelets (Ogden 1996). To keep the discussion on familiar ground, ϕ_1 will come from the function space of all quadratic curves for most of the remainder of this paper.

We denote the chosen function space by $\mathcal{H}$, and proceed with the specification of the proto-spline model class. Let $\psi_1(t), \ldots, \psi_T(t)$ be an orthogonal basis for $\mathcal{H}$. Then $\phi_1 \in \mathcal{H}$ is a specific linear combination of those bases, just as is an element of a vector space:

$$\phi_1(t) = \sum_{j=1}^{T} \eta_j \psi_j(t), \tag{8}$$

where the η_j are T *non-random* parameters that define the curve. If $\mathcal{H}$ is smooth, then so is $\phi_1(t)$. Extending this to the response variable, for the full model, we have

$$\begin{aligned} Y_i(t) &= \mu_i(t) + \omega_{i1}\phi_1(t) + \epsilon_i(t) & (9) \\ &= \mu_i(t) + \omega_{i1} \sum_{j=1}^{T} \eta_j \psi_j(t) + \epsilon_i(t), & (10) \end{aligned}$$

where ω_{i1} is a mean zero random coefficient with variance one and $\epsilon_i(t)$ are i.i.d. Gaussian random variables with variance σ_ϵ^2.[9]

[8] Function spaces have properties similar to vector spaces and are collections of curves that share a set of well defined properties. We use them to limit the set of possible shapes that define the structured variation in a process (see Jain et al. 1995 for details of function space theory).

[9] Note that the magnitude of the variation associated with ϕ_1 is contained in its norm, and the residual variation structure $\epsilon_i(t)$ can be made more complex.

Our model has two variance components, the variance of ω_{i1} and the residual variance, σ_ϵ^2, and it is a *parametric* covariance model that defers specification of the curve $\phi_1(t)$ to the estimation phase. The uncertainty in the form of $\phi_1(t)$ (until estimation) is a distinguishing feature of proto-spline models. Note that the parameters η_j define the shape of ϕ_1, which is fixed. These parameters do not represent the variance of a random coefficient, but taken as a whole, they determine the magnitude of the random curve ϕ_1 and thus the variance in the process.

The uncertainty allowed in the proto-spline class deserves further attention. In a standard mixed effects model, the random effects design matrix is fixed. Systematic variation takes a known form mediated by that design. The proto-spline class is a departure from this paradigm because it allows the shape of the design, given by ϕ_1 in our example, to be determined from all of the information in the data. In this sense, the curve ϕ_1 is a population-average value—the whole population influences its shape, and it can be considered a population "feature." Individual-specific differences are directly modeled using the random coefficient ω_{i1}. So this model is a hybrid between population-average and individual-specific philosophies and it belongs to the latent curve class of models.

In fact, the only philosophy not employed here is that of latent class modeling. As previously discussed, there is always a tension between modeling the covariance and modeling the mean, and since our emphasis is on covariance modeling, we do not utilize the latent class modeling philosophy, in which mean processes dominate.

3.2 Extensions

In this section we extend the development of the single proto-spline model in (7)-(10) to the general multiple proto-spline model. We note that the choice of our function space implies that ϕ_1 has a nonparametric interpretation, since it is a curve lying in a potentially smooth, continuous function space. Note that the model is not restricted to the space of any particular functions. Any finite-dimensional function space (or vector space) can be employed. An advantage to the proto-spline class of models is that the bases may be chosen to reflect the form expected in the substantive process without knowing which specific version of that form is present. If we choose models that result in latent curve estimates with a functional interpretation, features such as the derivative become available.

We define the full proto-spline class by extending the single curve example to more than one curve without introducing additional parameters. The main idea is to use only a subset of the T bases ψ_j to construct each curve ϕ_k. For this development, we let $\mathcal{H}$ be a basis for cubic splines with an appropriate set of knots. Let $\mathcal{I}_k$ be an indexing function defined on the integers $1, \ldots, T$, which selects the basis functions used to construct the k^{th} curve. In

our simple example, ϕ_1 uses all T basis functions, so $\mathcal{I}_1 = \{1, \ldots, T\}$. We construct latent curves as a deterministically weighted sum of the basis functions specified by the indexing function $\mathcal{I}$ so

$$\phi_k(t) = \sum_{j \in \mathcal{I}_k} \eta_j \psi_j(t). \tag{11}$$

In order to insure orthogonality of the ϕ_k, the index sets given by $\mathcal{I}_k$ must be disjoint. This restriction implies that once we decide to estimate more than one latent curve, the curves are highly constrained elements from the class $\mathcal{H}$, using only a subset of the bases for each. Since in the theoretical development $\mathcal{H}$ was chosen to be the natural cubic splines, we named the resulting ϕ_k *proto-splines*, because they are *partial* versions of a full spline fit. This method requires T parameters to build all K curves; if we do not normalize the curves, then for identifiability the random coefficients ω_{ik} are all presumed to have variance one.

To place this model in the context of those previously developed, we examine it for two extreme cases. First, if $K = T$, then each proto-spline is just a rescaled version of the basis function. This is essentially the model proposed in Brumback (1996) and Brumback and Rice (1998), although the form of their model was chosen to produce cubic spline *predictions* for individual curves. If $K = 1$, then we are estimating a smooth principal component in the presence of noise, and it is constrained to be a natural cubic spline.[10]

A more useful approach is to choose K to be small in relation to T, so that for equal-sized index sets, T/K bases are available for each latent curve. Equations (7) and (11) still apply, but the "proto-spline" nature of the curve estimates becomes more apparent. This intermediate case is similar to a principal functions analysis (Ramsay and Silverman 1997), in which we expect that most of the variation in the process is captured in a few of the largest principal functions. We are enforcing a small number of these by our choice of K, and we maintain the orthogonality requirement by the way the model is constructed. Note that this model differs from a principal functions analysis in that we can choose our function spaces with substantive features in mind, rather than simple smoothness constraints. We then build our model directly around these structures.

3.3 Link to Mixed Effects models

The standard mixed effects model can be expressed in the following form:

$$Y_i(t) = X_i(t)\beta + Z_i(t)\delta_i + \epsilon_i(t). \tag{12}$$

A key feature of this model is that $X_i(t)$ and $Z_i(t)$ are *prespecified* designs. The single latent curve proto-spline model is precisely the above model, with $Z_i(t) = \phi_1(t)$ and $\delta_i = \omega_{i1}$, only

[10] Another name for smooth principal components, principal functions, comes from the functional data analyses arena and is discussed at length in Ramsay and Silverman (1997).

$Z_i(t)$ is *specified from the data*. This illustrates a conceptual distinction between proto-spline models and other mixed effects models.

To explore the conceptual difference, we will consider three random quadratic models. For Model I, we assume that we know the exact quadratic curve that describes the structured covariation about the mean. Let $Z_i(t) = t + \frac{1}{2}t^2$, so $Z_i(t)$ is a scalar-valued function describing a *particular* growth structure. Further, let the random effect, δ_i, be a Gaussian random variable with unknown variance (the variance is one of the model's variance components). For a specific individual,

$$Y_i(t) = X_i(t)\beta + (t + \frac{1}{2}t^2)\delta_i + \epsilon_i(t). \tag{13}$$

Every subject gets some random multiple of the fixed curve $t + \frac{1}{2}t^2$.

For Model II we consider a mixed effects model in which each individual has their own quadratic perturbation as follows. Let the elements forming the three columns of $Z_i(t)$ be given by the vector $(1, t, t^2)$, and let $\delta_i = (\delta_{1i}, \delta_{2i}, \delta_{3i})$ be a vector of random coefficients, with a multivariate Gaussian distribution. Then for an individual specific curve,

$$Y_i(t) = X_i(t)\beta + \delta_{1i} + \delta_{2i}t + \delta_{3i}t^2 + \epsilon_i(t). \tag{14}$$

While this is quite flexible, the variance components analysis requires a full description of the estimated covariance structure of the random effects, which is contained in a 3×3 matrix that includes important covariance as well as variance components. We must use all of this information when describing any variance partitioning.

Model I is highly inflexible in that we must impose an exact form for growth beyond the mean. However, the variance component for δ_i is highly interpretable—it is the variance of the coefficient of precisely determined shocks to the system, so a larger variance means there is greater dispersion in individual growth, and that all structured growth follows the same form. It would be difficult to make a similar statement about Model II.

For Model III we consider a single latent curve proto-spline model, which offers the interpretability of the simpler model (I), and the flexibility of the more complex model (II). Let $\phi_1(t) = \eta_1\psi_1(t) + \eta_2\psi_2(t) + \eta_3\psi_3(t)$, where $\psi_1(t) = 1$, $\psi_2(t) = t$ and $\psi_3(t) = t^2$.[11] While this might resemble model II, the vector (η_1, η_2, η_3) is common to each individual and does not represent individual-specific random effects. Every individual curve has the following form:

$$Y_i(t) = X_i(t)\beta + \delta_i(\eta_1 + \eta_2 t + \eta_3 t^2) + \epsilon_i(t), \tag{15}$$

with the parameters (η_1, η_2, η_3) fixed and identical across individuals; this is a reparameterization of (10) that keeps the notation consistent. Each of these models is different, and

[11] We are ignoring the orthogonality requirement in this example for illustrative purposes.

we claim that the proto-splines offer an effective compromise between the rigidity and flexibility of Models I and II, respectively, while remaining highly interpretable from a variance components perspective.

To understand the link between proto-splines and other mixed effects models it is important to understand their technical distinctions. Scott and Handcock (1999) discuss estimation for the proto-spline model and show that there is a likelihood-equivalent but *non-standard* mixed effects model corresponding to the proto-spline class. In this section we describe the ways in which the proto-spline model is non-standard.

For notational convenience we suppress reference to time in the functions, representing $\psi_j(t)$ and $Z_i(t)$ as ψ_j and Z_i, respectively. Let $Z_i = \{\psi_1, \psi_2 \ldots, \psi_T\}$ be a design matrix constructed using the basis functions for the space $\mathcal{H}$. The coefficients η_j are assumed to be ordered so that if there are K different groups used in the model, with the k^{th} group given as $\gamma_k = (\eta_{k1}, \ldots, \eta_{kn_k})^T$, then the coefficients can be stacked into a $T \times K$ matrix $\Gamma = \bigoplus_{k=1}^{K} \gamma_k$. The difference between these two models can be understood by examining their representations. Our proto-spline model class is:

$$Y_i = X_i\beta + Z_i\Gamma\delta_i + \epsilon_i, \tag{16}$$

where $\delta_i \sim N_K(0, I_K)$. The likelihood-equivalent mixed effects model is

$$Y_i = X_i\beta + Z_i\delta_i^* + \epsilon_i, \tag{17}$$

where $\delta_i^* \sim N_T(0, \Gamma\Gamma')$.

The proto-spline formulation (16) has K random effects, while (17) has T. In (16), the random effect distribution is completely known ($N(0,1)$), while in (17) the parameters governing the effects (the γ_k's), must be estimated for us to know the structure of the random effects. In (16), the design $Z_i\Gamma$ represents the latent curve ϕ_1 and is estimated, while in (17), the design Z_i is prespecified. These distinctions are convenient ways to interpret the components of the models; they have the same likelihood and set of unknown parameters. In principal, the likelihood equivalence means that any software that can estimate a mixed effects model can be used to estimate the parameters of the proto-spline class. However, the covariance structure associated with model (17) would not be implemented in standard statistical software for mixed effects models, such as `SAS PROC MIXED` or `lme` in `Splus`.

While we have some evidence that these are different models, is this really the case? Restricting our attention to the *single* proto-spline model, formulation (16) contains a scalar random effect, δ_i, while the vector δ_i^* defined in (17) contains T effects. It is interesting to note that the covariance structure $\Gamma\Gamma'$ governing δ_i^* is degenerate, since $\Gamma\Gamma'$ is not positive definite. This does not introduce problems with estimation, however, because the degeneracy

is removed in the full likelihood, once residual variation is included. If one examines the structure $\Gamma\Gamma'$ more closely, it is apparent that each element of the vector δ_i^* is linearly dependent on each of the others, so in essence only *one* random effect is generated by this covariance structure.[12] So the likelihood-equivalent model (17) is a non-standard mixed effects model, which is equivalent to our proto-spline model, even in terms of the observations that would be generated from it, if we consider the limit of its degenerate covariance matrix.

What this means is that our proto-spline formulation effectively "corrects" the degeneracy in (17) by modeling the random effects in a simpler manner, without direct reference to the relationships indicated by the latter model's $\Gamma\Gamma'$ covariance structure. The γ_k parameters contained in the Γ matrix are essential to each formulation of the model, but should not be confused with what is actually random in the process. By viewing the design as estimated, rather than prespecified, our formulation (16) correctly separates the model into a portion driven by *population* features contained in the γ_k and individual features represented by δ_i.

In sum, the proto-spline class provides an interpretation for an interesting class of non-standard mixed effect models.[13] This interpretation is a philosophically distinct, hybrid modeling approach, and thus not only generates new knowledge with its use, but also establishes a new way to "allocate" the information provided in longitudinal data. The parameters contained in γ_k partition the variance as follows: they set the overall level of variation, since this is given by the sum of the components' squared values; and they describe the correlation structure because they define a shape which relates observations at different points in time. This formulation thus captures two things simultaneously in a full modeling class—orienting a modeling class to have these philosophical properties is to our knowledge novel in the literature.[14] By developing this class using likelihood-based procedures, a complete set of inferential tools is at the analyst's disposal. Scott and Handcock (2000) establish the asymptotic properties of this class and discuss inferential techniques. Being able to differentiate between population and individual effects is crucial to the formation of comparative statements in the policy domain.

[12]Due to the degeneracy, this analysis involves taking the limit as the covariance structure approaches $\Gamma\Gamma'$ from a "nearby" positive definite structure. The dependent relationships correspond to the relative magnitudes of the components η_{kj} of γ_k.

[13]Given their degeneracies, it is unclear if they can truly be labeled as mixed effects models; they are certainly non-standard.

[14]The proto-spline class is closely linked to a factor analysis model, which could be interpreted similarly, but we have not seen a formulation that explicitly attempts to guide the shape of the factors using something as flexible as a function space. The work of Rice and Silverman (1991) was pioneering by constraining principal components to be smooth, but was not presented as a modeling class, and inferential properties were not fully developed.

4 Illustration

For ease of exposition, we illustrate our model class by fitting a single latent quadratic curve proto-spline model to longitudinal wage data from the NLS. For the fixed effects, $X_i(t)$, we use a simple quadratic in age; this yields Model III of Section 3.3. In Figure 4, we display the cross-sectional mean of the process.[15] It provides the center from which the curves deviate. In Figure 5, we superimpose 4 simulated realizations from the proto-spline model fit, with the residual process $\epsilon_i(t)$ suppressed. The fitted curve $\hat{\phi}_1$ used in that simulation is presented in Figure 6. ¿From this figure, one can see that the growth of wages near the college years of 18 to 22 sets the extent of growth for the later years as well. The shape of the single latent curve describes the long-term trend in variation—strong growth in the 20s, followed by steady but diminished growth in the 30s. Each realization is simply the mean curve plus some random multiple (positive or negative) of the latent curve $\hat{\phi}_1$.

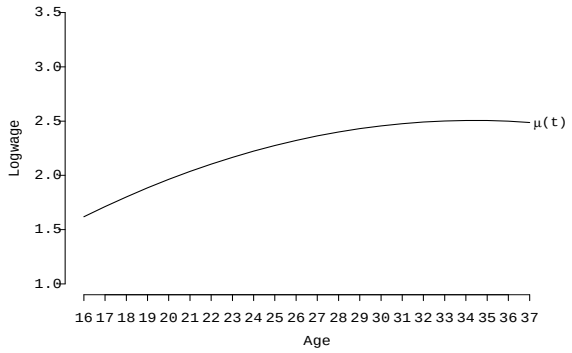

Figure 4: Mean curve for single proto-spline model

One might be concerned that imposing a quadratic latent curve is overly restrictive and essentially forces the decomposition into the shape indicated above. However, within the class of quadratic curves there are pure linear and constant curves, so if there were no change in the *growth rate* at early and later ages, then we would expect a different fitted latent curve. By forming a single latent curve spanning all ages, we are specifying that we want this curve to represent long-term structure, within the quadratic class. This is in part how we can model the covariance structure for the entire age span even though only segments of the full trajectory are observed for a specific individual.[16] More complex spaces

[15] As discussed, there is no choice for the mean that does not depend on the choice of covariance structure. We found the mean to be fairly robust to alternative formulations, and chose a quadratic mean for ease of exposition.

[16] Longitudinal data are often unbalanced or incomplete, sometimes by design. For example, the NLS cohort aged 14-21 during the initial survey year is tracked for 16 years, so complete individual records would

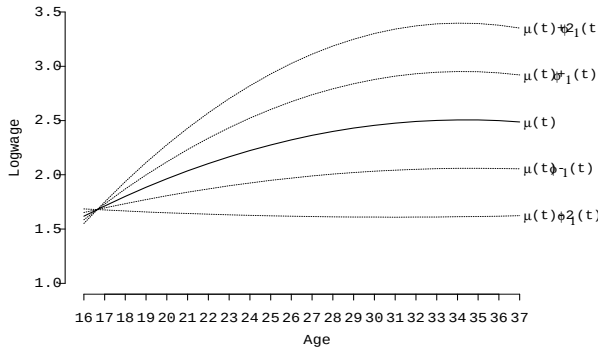

Figure 5: Curves for random coefficients one and two standard deviation from the mean for single proto-spline model

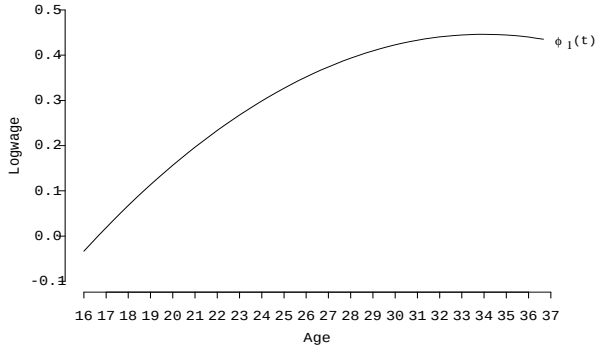

Figure 6: Fitted latent curve for single proto-spline model

may reveal more complicated dependencies, should they exist, and they should be considered in the model formulation process.

In Figure 5, we see that the effect of the single latent curve crosses the mean at age 17. The zero value for $\hat{\phi}_1$ near this age corresponds to a negligible amount of permanent variance, but even this is not predetermined by our choice of basis. If below-mean wages at those ages were to lead to much larger gains later on, the "crossover" would be at some later age. Random quadratics do limit us to a single change in the direction of growth (positive or negative), while higher order polynomials would not. Finally, note an important difference between this model and Model II. In Model II, each individual has a uniquely

span ages 14-30 and 21-37 in the extremes. No information on the correlation between responses at age 14 and 37 is available at the individual level. Estimating the covariance between ages 14 and 37, without making further assumptions about the covariance, is not possible. This concern and others noted by Jolliffe (1986) and Beale and Little (1975) are sufficient to warrant a model-based approach; the gains in precision associated with the imposition of a model (Altham 1984) further justify this. For a more sophisticated treatment of missing data issues in covariance structure estimation, the reader is also referred to Arminger and Sobel (1990).

shaped quadratic curve, so it may rise quickly and not level off, or it may level off quickly. In our model, which is basically Model III, every individual's variation beyond the mean has the same shape, given by $\hat{\phi}_1$—only the magnitude of that variation is allowed to vary.

5 Application and comparison of models

To illustrate how the models differ in practice, we apply several different covariance models to labor market data from the NLS. After a preliminary analysis, we found that the mean structure in this data resembles a quadratic curve, so we set the columns of the fixed effects design matrix to correspond to constant, linear and quadratic growth over time.[17] More complex mean structures can describe the influence of additional covariates on aggregate wage growth; our goal in this study is to understand the degree of long-term wage stratification, so the overall divergence of these curves over time in comparable samples yields important substantive information. Next, we must select a form for the structured portion of the variation. For wage data, the structured portion consists of long-term, or permanent, differences between wage trajectories.

We will compare three models. The first is a random quadratic mixed effects model similar to Model II and to that used by Bernhardt, et al. (1997) in their analyses. The second is a single latent curve proto-spline model, similar to Model III, and the third is an extension of proto-spline models that includes a second non-orthogonal latent curve. Beyond the structural variation just described, the residual variation is modeled simply as independent with constant variance.

5.1 Random Quadratics

The strength of a random quadratic model, such as that given by (14) is the flexibility provided by the three random coefficients, δ_{1i}, δ_{2i} and δ_{3i}. Note that the quadratic basis we use is an orthogonalized and normalized version of $(1, t, t^2)$, which is also the fixed effects basis. The random coefficients are globally constrained to come from a multivariate Gaussian density. This choice yields a broad range of curves of various shapes and intensities. This distributional form does, however, require that there are no clusters of curves, or other multimodalities.

We assume that the δ_i are distributed as $N(0, G)$, where G is a completely unspecified 3×3 covariance matrix defined by six distinct parameters. We fit the model on data from

[17] Note that we orthogonalize and normalize these basis vectors, so instead of a vector of ones, the first column is a vector consisting entirely of $1/\sqrt{22}$, based on the 22 ages from 16 to 37.

the recent NLS cohort[18] using maximum likelihood estimation, and find that

$$\hat{G} = \begin{pmatrix} +2.019 & +0.899 & -0.265 \\ +0.899 & +1.312 & +0.096 \\ -0.265 & +0.096 & +0.529 \end{pmatrix}$$

and $\hat{\sigma}^2_\epsilon = 0.0719$.[19] Unfortunately, these results are somewhat hard to interpret. The structured portion of the covariance is given by $Z_i \hat{G} Z_i^T$, where Z_i is the random effects design matrix. Since the rows of Z_i correspond to the subject's age, this matrix product describes individual wage differences at each age and how they relate to each other. For example, the diagonal of $Z_i \hat{G} Z_i^T$ represents the structured, or permanent, wage variance at each age. These values are plotted against age in Figure 7 below. The initially larger variance

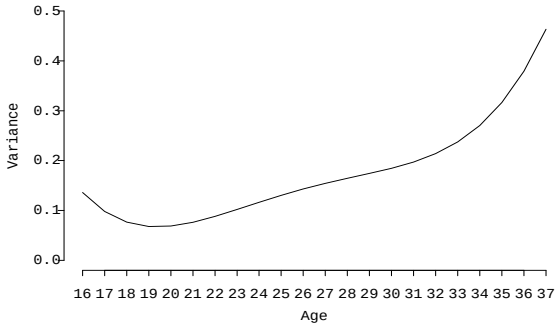

Figure 7: Permanent wage variance for random quadratic model

at the earliest ages indicates some initial stratification between individual trajectories that seems to diminish by age 20, only to increase substantially from that point forward, with a dramatic rise after age 32. Had permanent differences in trajectories been limited to an intercept shift, this graph would have consisted of a horizontal line some distance above the axis. The result above indicates that wages fan out quite dramatically as individuals age, and gives some indication of how the process accelerates. We can infer that the trajectory fans out as a whole because the partition is based on a model employing a continuous curve for the permanent portion of the trajectory.

[18] All of the following results are based on recent cohort data; we make cross-cohort comparisons in Section 5.4.

[19] Differences in these findings and those presented by Bernhardt, et al. (1997) are due to a different choice for the quadratic basis, but are otherwise comparable.

5.2 Single latent curve proto-spline

In this model, we assume that most of the structured variation takes a specific form, but we let the exact shape be determined by the data. The explicit model is

$$Y_i(t) = X_i(t)\beta + \delta_i\phi_1(t) + \epsilon_i(t), \tag{18}$$

where ϕ_1 is the single latent curve, assumed to lie in the space of quadratic curves, and δ_i is the random coefficient for the i^{th} individual. This is the same model as the one used for our illustrative example in Section 4. A look at Figure 5 (prior page) reveals the strength of this model. A wide range of outcomes are easily represented by the mean plus a random multiple of the single latent curve, $\hat{\phi}_1$. Figure 6 (prior page) displays this curve for the recent cohort.

In this figure, the interpretability of this class of longitudinal data models becomes apparent. The single latent curve reveals most of what we need to know about structured variation. Contrast this to the covariance matrix $\hat{G}$, which along with design matrix Z_i provides the equivalent information in a less accessible form. The random coefficient on our proto-spline model is standard Gaussian, so we have an immediate sense of the range of impact of the single latent curve.

The restriction to a single latent curve does limit our ability to model more complex structured variation. In Figure 8 below, we see that the permanent variation, the squared version of $\hat{\phi}_1$, describes a very simple growth structure.[20] Two features stand out in comparison to random quadratic models: the permanent variation starts out lower at the youngest ages and it does not grow as dramatically as individuals age. We believe that the initial variation is less important from a likelihood perspective, so it is effectively being ignored in the estimation process. Had we used a higher order polynomial, we might have discovered persistent initial wage differences. If this were the case, we would expect $\hat{\phi}_1$ to begin higher, possibly decrease somewhat and then increase again, in a shape similar to the permanent variance graph from the random quadratic model.

5.3 Double latent curve model

The limitations of a single curve model prompts us to explore a model with two latent curves. Fitting such a model under the pure proto-spline formulation would require the fitted curves to be orthogonal, and this restricts the function spaces in which each may lie. We propose a new model that effectively "reuses" the basis for each latent curve. The model, abstractly, is given by

$$Y_i(t) = X_i(t)\beta + \delta_{1i}\phi_1(t) + \delta_{2i}\phi_2(t) + \epsilon_i(t), \tag{19}$$

[20] The permanent variance calculation is straightforward in this case because the random coefficient δ_i is standard Gaussian.

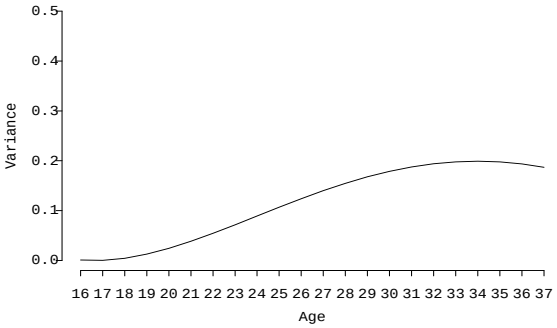

Figure 8: Permanent wage variance for single latent curve model

where ϕ_1 and ϕ_2 are latent curves and δ_{1i} and δ_{2i} are i.i.d. standard Gaussian random coefficients. We construct each curve from the same basis.

$$\phi_k(t) = \sum_{j=1}^{T} \eta_{kj}\psi_j(t), \tag{20}$$

with $k = 1$ or 2, effectively doubling the number of parameters used by the single latent curve model. After adding some identifiability constraints to our estimation procedure, we were able to fit this more complex model.

The double latent curve model can best be understood as the combination of a common mean process and two independent "shocks" taking some functional form. We choose to continue to employ the space of quadratic polynomials for ease of exposition. Looking at Figure 9, we find that the fitted curves are quite different from each other. These are the forms for the two shocks, $\hat{\phi}_1$ and $\hat{\phi}_2$. We see that $\hat{\phi}_1$ is quite similar to its counterpart in the single latent curve model, although it starts out further below the origin. The latter feature will induce greater permanent variation at the youngest ages, and then this will subside, as the curve crosses the origin between ages 18 and 19 (contrast this to the crossing at age 17 in the single curve model). The second curve introduces a whole new feature to the covariation. It appears that individuals who start out earning more are penalized as they age. This is indicated by $\hat{\phi}_2$'s initially positive level of about 0.2 at age 16, which sinks to -0.3 by age 37. Of course, negative random coefficients are just as likely as positive ones, so this curve could also represent later growth for young workers who initially accept lower wages. There is mild evidence that this is capturing an "education effect," in which individuals who defer fully entering the labor market (and possibly pursue education or training) benefit with larger wage growth in the long run.[21]

[21]This effect was based on an analysis of the final level of education attained by each individual and the predicted value of the coefficient $\hat{\delta}_{2i}$ of $\hat{\phi}_2$.

Note also the similarities and differences of our fitted model to a principal components analysis (PCA). The proto-spline restriction to a smooth function space means that short-term variability is definitely removed, and each curve represents a permanent component of variation. With a model-based approach, we can precisely describe how the latent curves are added to the response process. This is less immediate with the components in a PCA, because the PC scores have no predetermined distributional form. Further, the proto-spline process is well-defined under the entire age range of interest without either the use of an ad hoc procedure or requiring a balanced design.

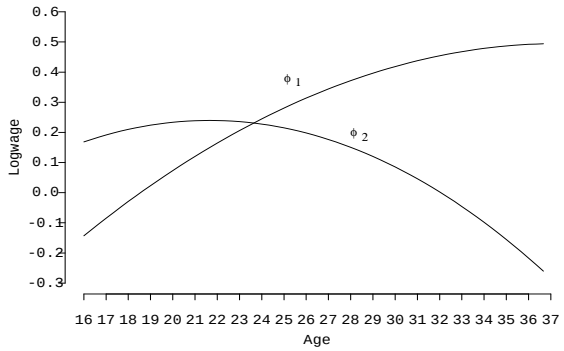

Figure 9: Fitted proto-splines for double latent curve model

The permanent variance partitioning for this model is given in Figure 10, below. By

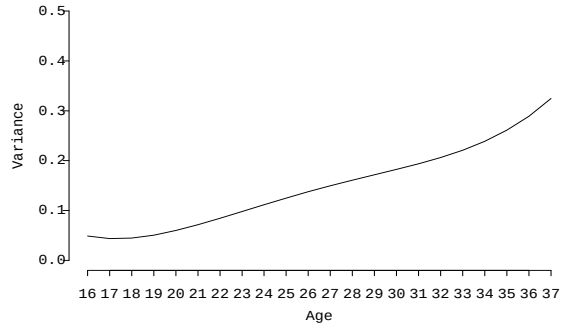

Figure 10: Permanent wage variance for double latent curve model

including two curves additively and independently, this model allows for larger early and later year variation. The effects are permanent, in that they persist over the lifetime of a worker, but their independence points to a subtlety of these variance decompositions. Two curves, along with their coefficients describe the systematic portion of a trajectory, but the independence of the coefficients severs any link between the two. In terms of generating

mechanisms, this only makes sense if two different features of the wage growth process are being captured, such as an overall growth (often attributed to returns to job tenure and experience) and an education effect.

The above comment also points to a limitation of the random quadratic model. Namely, it is hard to describe an underlying process (often thought of as a latent characteristic) that is driving the three coefficients forming the curves. It is hard to imagine that a social or economic generating mechanism involves intercept, slope and acceleration components. The latent curve models provide simpler explanations, which is an advantage in this case.

5.4 Comparing variance partitions

While related, these three models provide different variance decompositions. We display the permanent variation plots in Figure 11, below, and include 95% confidence intervals at each age. We discuss the construction of those intervals in Section A.1 of the appendix. Notable differences exist for the youngest and oldest ages, with strong agreement in the middle range. The single latent curve model does not pick up much structured wage variation at the youngest ages. If initial differences in wages persist during the youngest ages, but then diminish, then this model will have to choose between the initial and later year effects, and since the latter are larger, they tend to dominate. The double proto-spline model picks up this extra variation in $\hat{\phi}_2$, and this is reflected in larger permanent variance for the younger ages. The random quadratic model picks up more variation in both younger and older ages and labels it permanent. We contend that the additional flexibility of the random quadratic model allows it to follow the raw data more closely, capturing less rigid forms of variation. This is indirectly confirmed by examining the residual variation, which is 0.072 for random quadratics, and 0.078 and 0.098 for double and single latent curve models, respectively.

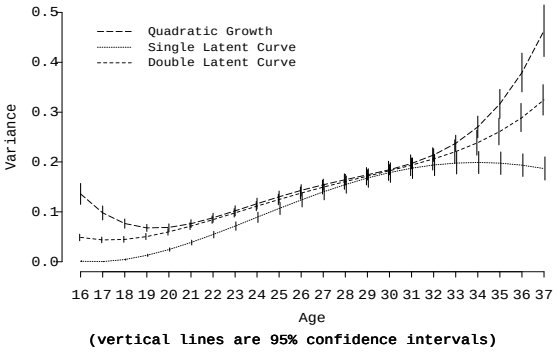

Figure 11: Permanent wage variance for all three models

Below we present the variance decomposition for each cohort to address an important

question. While each model partitions the variance differently, do these differences have substantive impact? That is, how sensitive are the answers to the substantive questions to the choice of model? In our application, the question of interest is whether or not the permanent wage variance between the cohorts differs, and if so, by how much. Any model we use will only be an approximation, but if the answer to our question is consistent across models, we can have more confidence in any conclusions we draw.

In Figures 12 through 14 below, we make a cross-cohort comparison and display the model-based permanent variance for each model along with 95% confidence intervals at each age. All of the models indicate a significantly larger permanent variance in the recent cohort, starting sometime in the mid-twenties. The difference is most dramatic in the random quadratic model and least so in the single latent curve model. There is some between-model discrepancy in what portion of the variance is permanent at the youngest ages, and in how the cohorts differ. Both latent curve models contain a crossover, in which the original cohort starts out more stratified until the early twenties, at which point the opposite it true. In contrast, the random quadratic model posits that both cohorts are more permanently stratified initially and to a comparable extent. If we are interested in the absolute magnitude of permanent wage stratification, we must look more closely at all of these models and determine which is more justified on substantive grounds. If we were concerned about wage stratification at the younger ages, a deeper understanding of each model's characteristics is warranted, since these models tell three different stories.

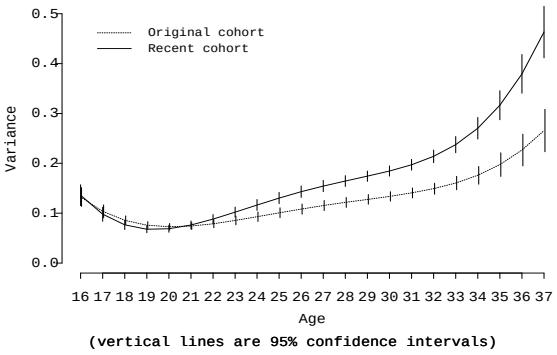

Figure 12: Permanent wage variance for random quadratic model

5.5 Discussion of findings

All three of these models indicate a significant increase in permanent wage variation in the recent cohort for the older ages. But the magnitude of these differences varies greatly between models, and strong differences in the partitions exist at the younger ages.

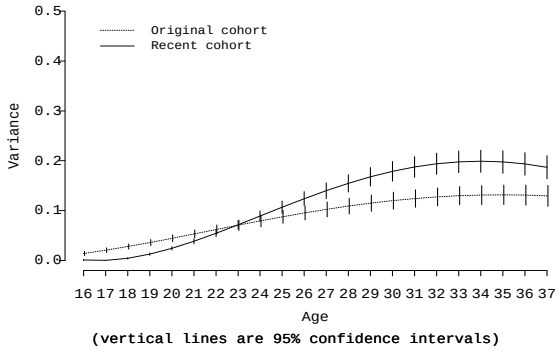

Figure 13: Permanent wage variance for single latent curve model

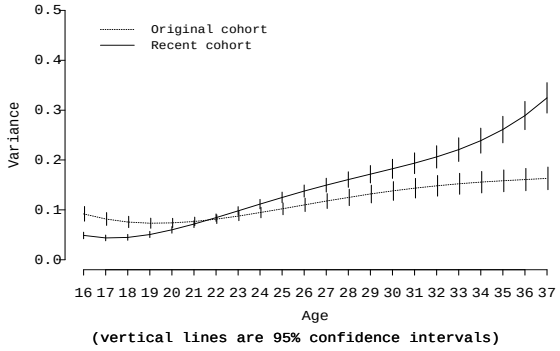

Figure 14: Permanent wage variance for double latent curve model

Since the random quadratic model labeled more variation as permanent, it may be overfitting that feature, in some sense. The flexibility of random quadratics admits even a U-shaped curve, but is it desirable to use such a shape to describe *permanent* wage gains? U-shaped curves, in which initial and final wages are nearly identical, involve a shift in the direction of wage change, from loss to gain, so in what sense is this indicative of a permanent, or lasting, trend? We must understand how the choice of model is reflected in the variance partition if we intend to make an informed assessment of social phenomena.

Latent curve models stand out as a philosophically mixed approach to creating variance partitions. They are highly interpretable, with independent components acting as shocks to the process. The shocks may be readily interpretable in the context of the generating mechanisms for the social processes under study. They offer a handy form of rigidity compared to random quadratic models, yet are inherently adaptive to overall patterns of structured variation. The hybrid nature of this model class provides a new type of analysis to which results from other classes can be compared. Thus, these models can be viewed as excellent foils to the classical random quadratic model.

All three of the models describe the structured portion of variance in such a way that "permanent" is a reasonable label to apply. That is, the model describes smooth versions of the curves in space that are reasonable attempts to separate the analyst-defined signal from noise; and the signal is non-stochastic, conditional on the parameters that describe it. The differences in these definitions allow different aspects of variation to be identified.

6 Conclusion

We embarked on this analysis to determine how different models for covariance affect variance component partitioning. Along the way, we introduced a new, hybrid class of latent curve models, proto-splines, that offer an interpretable paradigm for describing covariation, which is well-suited to formulating substantive questions directly. These models locate population-level persistent covariance structure and reflect it in the shape and size of latent curves. We view proto-splines as covariance function smoothers; they are non-parametric in the sense that the estimated curve lies in a function space, yet the model formulation provides a straightforward interpretation of the curves that is often missing in other non-parametric techniques. In the model formulation, the researcher imposes a class of functions to capture substantively meaningful structure. The restriction to a particular class of functions forces proto-spline models to be conservative in the way they fit the data—they are less susceptible to outliers, which in other models may influence both prediction and fit. This makes them invaluable in comparisons with more traditional models; the ways in which they differ point out characteristics of each, with the clearly defined behavior of our models acting as a foil for the others.

In future work, we will consider relaxing the independence assumption for the proto-spline model class. For example, our double latent curve model could include a term for the correlation between curves. This extension would open up the possibility of very different latent curves, since the independence constraint ultimately lowers the likelihood of certain shapes for the fitted curves. Including more complex residual structures, such as age-specific variances, as a check on the homogeneous variance assumption could prove useful.

In many socio-economic processes, there are jump points that are not smooth, but have important substantive meaning. Adapting the proto-spline class to allows for uncertainty in the timing of the change-point could prove useful. Raftery (1994) explores this issue; integrating his approaches with ours is a research direction of interest.

Relaxing the Gaussianality assumption is worth investigating, but we would limit this to forms that remain interpretable, such as parametric forms. One approach that has been suggested by several researchers is a latent class, or mixture formulation (see Clogg 1995,

Banfield and Raftery 1993, Muthén and Shedden 1999, Roeder 1999, Verbeke and Lesaffre 1997, Xu, et al. 1996). Under this paradigm described earlier, individuals belong to *one* latent class, and then conditional on class membership they follow a certain structure. An important point is that the remaining structure could be flexibly captured in the proto-spline models just introduced; most models currently in use do not offer such directly interpretable covariance formulations.

In work in progress, we are examining diagnostics for these models in greater detail. Model selection criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978) can be applied here. These are discussed in Vonesh and Chinchilli (1998) and Pinheiro et al. (1994). Recent extensions to the AIC discussed in Simonoff and Tsai (1999) appear to be especially promising in in the context of these variance component models. An alternative to model selection is the use of Bayesian model averaging (Hoeting, et al. 1998). A developed set of diagnostic techniques will add to our understanding of how each model captures and partitions variation.

A Appendix

A.1 Construction of confidence intervals

Confidence intervals for the model-based variances, such as the permanent variation, are constructed from the asymptotic covariance matrix of the model parameters. For proto-spline models, explicit forms for these covariance matrices are given in Scott and Handcock (2000). The construction begins by finding the variance associated with each point on the latent curve. Each curve is a linear combination of a set of basis functions, with the coefficients specified by the model parameters. If these coefficients are given by column vector $\ell_k = (\eta_{k1}, \ldots, \eta_{kT})^T$, the basis functions by matrix $Z = [\psi_1, \ldots, \psi_T]$ and the asymptotic covariance matrix of ℓ_k by H_k, then the resulting latent curve is given by $\phi_k = Z\ell_k$ and the covariance of $Z\ell_k$ is ZH_kZ^T. So the variance of the estimate of the curve $\hat{\phi}_k$ at each time point (and their covariances) are contained in the diagonals (and off-diagonal elements) of ZH_kZ^T. Confidence intervals for the permanent variance (the squared value of the curve estimate) at a particular time point can be constructed via a delta-method approximation involving that same asymptotic covariance matrix.[22]

[22]Confidence intervals for extra-mean variation for random quadratic models involve a slightly different calculation and will not be discussed in this paper.

References

[1] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control,* **19**, 716-723.

[2] Altham, P. (1984) Improving the Precision of Estimation By Fitting a Model. *Journal of the Royal Statistical Society, Series B, Methodological*, **46**, 118-119.

[3] Arminger, G. and Sobel, M. E. (1990). Pseudo-Maximum Likelihood Estimation of Mean and Covariance Structures With Missing Data. *Journal of American Statistical Association* **85**, 195-203.

[4] Baker, M. (1997). Growth-Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings. *Journal of Labor Economics,* **15**, 338-375.

[5] Banfield, J. D. and Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics,* **49**, 803-821.

[6] Bernhardt, A., Morris, M., Handcock, M. and Scott, M. (1997). Work and Opportunity in the Post-Industrial Labor Market. Final report to the Russell Sage and Rockefeller Foundations. Institute on Education and the Economy, Columbia University New York.

[7] Beale, E. and Little, R. (1975). Missing Values in Multivariate Analysis. *Journal of the Royal Statistical Society, Series B, Methodological,* **37**, 129-145.

[8] Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

[9] Brumback, B. A., and Rice, J. A. (1998). Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves. *Journal of American Statistical Association* **93**, 961-976.

[10] Brumback, B. A. (1996). Statistical Models for Hormone Data. Ph.D Thesis, University of California, Berkeley.

[11] Clogg, C. C. (1995). Latent Class Models. In G. Arminger, C. C. Clogg and M. E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, 311-359. Plenum Press.

[12] Crouchley, R. and Davies, R. B. (1999). A Comparison of Population-Averaged and Random-effect models for the analysis of longitudinal count data with base-line information *Journal of the Royal Statistical Society, Series A*, **162**, 331-347.

[13] Danziger, S. and Gottschalk, P., eds. *Uneven Tides: Rising Inequality in America.* Russell Sage Foundation.

[14] De Boor, C. (1998). *A Practical Guide to Splines*, Vol. 27. New York: Springer-Verlag.

[15] Devlin, S. J. and Gnanadesikan, R. and Kettenring, J. R. (1981). Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association*, **76,** 354-362.

[16] Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

[17] Gottschalk, Peter and Moffitt, Robert (1994). The Growth of Earnings Instability in the US Labor Market. *Brookings Papers on Economic Activity*, **2**, 217-272.

[18] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

[19] Haider, Steven (1996). Earnings Instability and Earnings Inequality in the United States: 1967-1991. Manuscript, University of Michigan.

[20] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian Model Averaging. *Statistical Science*, forthcoming.

[21] Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R. and Pentz, M. A. (1998). Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes *American Journal of Epidemiology*, **147**, 7, 694-703.

[22] Jain, P. K., Ahuja, O. P. and Ahmad, K. (1995). *Functional Analysis* New York: Wiley.

[23] Jolliffe, I. T. (1986). *Principal Component Analysis*, New York: Springer-Verlag.

[24] Jones, M. C. (1990). Serial Correlation or Random Subject Effects? *Communications in Statistics Series B*, **19**(3), 1105-1123.

[25] Levy, Frank and Murnane, Robert (1992). U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations *Journal of Economic Literature,* **30**, 1333-1381.

[26] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22.

[27] Locantore, N. , Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999). Robust Principal Component Analysis for Functional Data. *Test*, **8,** 1-73.

[28] Longford, N. T. (1993). *Random Coefficient Models* New York: Oxford.

[29] McCulloch, Charles E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models *Journal of the American Statistical Association*, **92**, 162-170.

[30] McMurrer, D. and Sawhill, I. (1998). *Getting Ahead: Economic and Social Mobility in America* Washington, DC: The Urban Institute Press.

[31] Meredith, W. and Tisak, J. (1990). Latent Curve Analysis. *Psychometrika,* **55**, 105-122.

[32] Muthén, B. and Shedden, K. (1999). Finite Mixture Modeling with Mixture Outcomes using the EM Algorithm. *Biometrics*, **55**, 463-469.

[33] Ogden, R. T. (1996). *Essential Wavelets for Statistical Applications and Data Analysis*, Boston: Birkhäuser.

[34] Pinheiro, J. C., Bates, D. M., and Lindstrom, M. J. (1994). Model Building for Nonlinear Mixed Effects Models Technical Report 931, University of Wisconsin, Madison, Dept. of Statistics.

[35] Prentice, R. L. (1988). Correlated Binary Regression With Covariates Specific to Each Binary Observation. *Biometrics*, **44**, 1033-1048.

[36] Raftery, A. E., (1994). Change Point and Change Curve Modeling in Stochastic Processes and Spatial Statistics. *Journal of Applied Statistical Science*, **1**, 403-424.

[37] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer.

[38] Raykov, T. (2000). Modeling Simultaneously Individual and Group Patterns of Ability Growth or Decline. In Little, T. D., Schnabel, K. U. and Baumert, J. (eds.), *Modeling Longitudinal and Multilevel Data*, 127-146. Mahwah, NJ: Lawrence Erlbaum.

[39] Rice, J. A. and Silverman, B. W. (1991). Estimating the Mean and covariance Structure Nonparametrically when the Data are Curves. *Journal of the Royal Statistical Society Series B,* **53**, 233-243.

[40] Roeder, K., Lynch, K. G. and Nagin, D. S. (1999). Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology *Journal of the American Statistical Association*, **94**, 766-776.

[41] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics.* **6**, 461-464.

[42] Scott, M. A. and Handcock, M. S. (2000). Latent Curve Covariance Models for Longitudinal Data. Working Paper, Center for Statistics and the Social Sciences, University of Washington, Seattle.

[43] Scott, M. A. (1998). Statistical Models for Heterogeneity in the Labor Market. Ph.D Thesis, New York University.

[44] Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* Wiley.

[45] Simonoff, J. S. and Tsai, C. (1999). Semiparametric and Additive Model Selection Using an Improved AIC Criterion. *Journal of Computational and Graphical Statistics* **8**(1), 22-40.

[46] Topel, R. H. and Ward, M. P. (1992). Job Mobility and the Careers of Young Men. *Quarterly Journal of Economics,* **107**, 439-479.

[47] Verbeke, G. and Lesaffre, E. (1996). A Linear Mixed-Effects Models with Heterogeneity in the Random-Effects Population. *Journal of American Statistical Association* **91**, 217-221.

[48] Vonesh, E. F. and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker.

[49] Xu, W., Hedeker, D., and Ramakrishnan, V. (1996). Mixtures in Random-effects Regression Models. Manuscript, School of Public Health, University of Illinois at Chicago.

[50] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, **42**, 121-130.