A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects

by

Andrew Karl

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved February 2012 by the
Graduate Supervisory Committee:

Sharon Lohr, Co-Chair
Yan Yang, Co-Chair
Ming-Hung Kao
Douglas Montgomery
Jeffrey Wilson

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

Value-added models (VAMs) are used by many states to assess contributions of individual teachers and schools to students' academic growth. The generalized persistence VAM, one of the most flexible in the literature, estimates the "value added" by individual teachers to their students' current and future test scores by employing a mixed model with a longitudinal database of test scores. There is concern, however, that missing values that are common in the longitudinal student scores can bias value-added assessments, especially when the models serve as a basis for personnel decisions – such as promoting or dismissing teachers – as they are being used in some states. Certain types of missing data require that the VAM be modeled jointly with the missingness process in order to obtain unbiased parameter estimates.

This dissertation studies two problems. First, the flexibility and multimembership random effects structure of the generalized persistence model lead to computational challenges that have limited the model's availability. To this point, no methods have been developed for scalable maximum likelihood estimation of the model. An EM algorithm to compute maximum likelihood estimates efficiently is developed, making use of the sparse structure of the random effects and error covariance matrices. The algorithm is implemented in the package GPvam in R statistical software. Illustrations of the gains in computational efficiency achieved by the estimation procedure are given.

Furthermore, to address the presence of potentially nonignorable missing data, a flexible correlated random effects model is developed that extends the generalized persistence model to jointly model the test scores and the missingness process, allowing the process to depend on both students and teachers. The joint model gives the ability to test the sensitivity of the VAM to the presence of nonignorable missing data. Estimation of the model is challenging due to the non-hierarchical dependence structure and the resulting intractable high-dimensional integrals. Maximum likelihood estimation of the model is performed using an EM algorithm with fully exponential Laplace approximations for the E step. The methods are illustrated with data from university calculus classes and with data from standardized test scores from an urban school district.

To Laura, Oliver, and the rest of my family.

ACKNOWLEDGEMENTS

I owe a heartfelt thanks to my advisors, Dr. Lohr and Dr. Yang, who have dedicated so much of their time and effort to helping me develop this dissertation. They were selfless in their willingness to respond quickly to all of my emails and to meet with me whenever I needed their assistance. I am indebted to them for significant portions of my graduate education that they went out of their way to provide me with, including encouragement and funding to travel to JSM in Vancouver and Miami Beach, help with job searching, guiding me through the publication and peer-review process, and in general teaching me how to work in an academic environment. Sharon and Yan made a great team as co-advisors, and it was clear that they wanted to help me accomplish this goal.

I am also very grateful to my other committee members – Dr. Kao, Dr. Montgomery, and Dr. Wilson – for their valuable time and feedback, as well as their mentorship with coursework and professional development. Likewise, I would like to recognize Dr. Young and Dr. Eubank, with whom I had a thoroughly enjoyable time researching parallel random number generation. All of these faculty members from Arizona State have been extremely friendly and willing to share their knowledge. Finally, I need to thank the graduate program coordinator Debbie Olson for her help in navigating many administrative hurdles.

Many others have helped me along the path to a Ph.D. I would like to thank my family – Tom, Paula, and Betsy – for their tremendous support, and my wife, Laura, and son, Oliver, for their patience and encouragement. Dr. Dennis Snow (University of Notre Dame) sparked my interest in statistics as an undergraduate and showed me how valuable it is to have a good advisor. Although I decided during his REU that I was better suited to be a statistician than a geometer, Dr. Justin Corvino (Lafayette College) introduced me to mathematical research: I am grateful for the patience he had with me. Finally, although a decade has passed, I would be remiss in failing to acknowledge Andrea Hodges and Dr. Brian Reeves for inspiring me while I was in high school.

This research was partially supported by the National Science Foundation under grant DRL-0909630. In addition, my fourth year of studies was funded by an ASU Dissertation Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the National Science Foundation or Arizona State University.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Value-added models (VAMs) are increasingly recommended as a method of estimating the effectiveness of schools and teachers. It is widely recognized that comparing teachers and schools based solely on their students' test scores unfairly penalizes those that teach students from disadvantaged backgrounds. VAMs attempt to partially account for student background by examining student growth on standardized tests and estimating teacher or school effectiveness using the growth of the students who have been instructed by that teacher or school. The VAM scores for a teacher from the models are intended to estimate the "value added" by that teacher to the student's growth—how much more (or less) students' scores changed under that teacher than they would be expected to change under an "average" teacher.

The federal Race to the Top program requires that participating states use longitudinal student achievement data as a component of teacher and principal evaluation. Value-added models are increasingly being used to estimate the portions of individual students' growth that can be attributed to their teachers (see `www.cgp.upenn.edu/ope_nation.html` for a list of states that have adopted VAMs). Discussions of the use of VAMs have appeared recently in the New York Times (Leonhardt, 2010) and the Los Angeles Times (Felch et al., 2010). The Los Angeles Times article describes a value-added analysis performed by the newspaper on Los Angeles public school data and publishes the resulting teacher rankings. The publication of the results created a strong reaction both from teachers whose names had been released with rankings and from parents of students.

In October of 2009, the National Academy of Sciences deemed VAMs to be a promising approach to estimating teacher effectiveness, but cautioned that they should not be used for “high stakes” decisions, such as tenure or dismissal, until they have been studied and developed further (The National Academies, 2009). One issue raised by Briggs and Domingue (2011) involves the sensitivity of teacher rankings to the inclusion of fixed-effect covariates in the VAM based analysis published by the Los Angeles Times (Felch et al., 2010). Despite such concerns, a number of states are now using VAMs for personnel decisions, including merit pay and tenure. For example, Louisiana recently began to use VAMs for 50 percent of each public school teacher’s evaluation (Louisiana Department of Education, 2010). With $4.35 billion in funding for Race to the Top, it seems that VAMs will play a major role in the future in judging teacher, school, district, and state performance (U.S. Department of Education, 2009). Given the magnitude of the decisions that the value-added education models will be used to make, it is extremely important that the models are thoroughly studied and their limitations explored.

Although the purpose of VAMs is to measure the teacher contributions, such causal inference can be misleading in the presence of non-random student assignment to classrooms (Draper, 1995). What VAMs attempt to measure as teacher effects may be more appropriately referred to as “unexplained heterogeneity at the classroom level” (Lockwood et al., 2007). The motivation in using VAMs is to try to compensate for this non-random assignment through effects in the model. This provides an improvement over simplistic gain score models that have been used in the past (Harris and McCaffrey, 2010).

Some VAMs are presented as standardized gain models (Reback, 2008), or as student growth percentile models, such as the Colorado Growth Model (Betebenner, 2009). The Colorado Growth Model uses quantile regression to measure student growth relative to peers who were at a similar achievement level in the previous year(s). Teachers are then ranked by the median student growth percentile from their class. However, there are concerns about the stability of the estimates from these nonparametric methods, and about their potential bias against teachers of students from poor socioeconomic backgrounds (Wright, 2010).

More complex VAMs are defined using linear mixed models (Sanders et al., 1997; Raudenbush and Bryk, 2002; Ballou et al., 2004; McCaffrey et al., 2004; Lockwood et al., 2007; Harris and McCaffrey, 2010; Wright et al., 2010). These models vary in their correlation structures, their reliance on covariates, and the structure for the effect of a teacher in year g on future year scores in years $g+1$, $g+2$, etc. In this context, estimation is done through either a nested or a multi-membership mixed model on longitudinal student data which includes random effects for teachers and students, as well as any desired fixed effects, such as socioeconomic status. The empirical best linear unbiased predictors (EBLUPs) of the teacher random effects serve as the estimates of teacher contributions to student learning. Rather than simply calculating the average test score for a classroom, value-added models control for information on the students' backgrounds, the students' individual test score histories, and contributions of previous teachers to the students' learning. We focus our attention on these mixed model based VAMs.

The generalized persistence (GP) VAM (Mariano et al., 2010) is one of the most general and flexible value-added models in the literature, containing most of the other linear mixed model based VAMs as special cases. Unlike other VAMs, the GP model allows the effects of teachers on future year scores to be imperfectly correlated. This contrasts with the perfect correlation assumed by variable persistence models (McCaffrey et al., 2004; Lockwood et al., 2007), and the equality of current and future year effects assumed by complete persistence models (Sanders et al., 1997). However, the flexibility afforded by the GP model comes at a heavy computational cost. No computational methods are currently available for maximum likelihood estimation of the GP model due to its multi-membership random effects structure and highly correlated random effects. SAS is able to estimate the model parameters for only very small data sets.

The first part of this dissertation develops a maximum likelihood procedure for estimating the GP VAM, which was estimated with Bayesian methods by Mariano et al. (2010). Mariano et al. (2010) find that even minimally-informative priors for the covariance matrix of the random effects were “quite informative”, meaning the results are sensitive to the choice of prior distributions. Estimation of the mixed model is done via an EM algorithm with custom-written code in R (R Development Core Team, 2012). We have built the estimation routine into the R package GPvam, and hope that the availability of software for the GP model will encourage further exploration of its properties. The GP VAM has the potential to be used in many settings outside of educational evaluation.

An often neglected aspect of value-added modeling is the effect of missing data on the results of the analysis (McCaffrey and Lockwood, 2011). Longitudinal education data often contain many incomplete student profiles. Students drop courses, change schools, move away, or may be absent on the day of an exam. Analysis of data where some observations are missing requires assumptions about the nature of the missing data. Of particular interest is a type of missing data that arises in the college setting. For example, students in calculus 2 who do not finish calculus 3 will have missing data for calculus 3. The missingness may be relevant to estimates of the calculus 2 teachers’ contributions to student learning. A student who is poorly prepared for calculus 3 may drop the class despite having received a high grade in calculus 2. Or, in the elementary-school setting, it is possible that low-performing students might be discouraged from taking a standardized exam (Ryan and Weinstein, 2009). In a simplistic example, suppose that students are randomized to one of several classrooms and teachers are evaluated based upon the average score for their classes on a standardized exam at the end of the year. If a teacher were to discourage her weakest students from taking the exam, she could inflate her class average and thus her ranking. Ignoring data that are missing in part due to the value that would have been observed may lead to biased estimates (Little and Rubin, 2002).

The assumptions about missing data made by VAMs are usually overly simplified and have been recognized as a potential problem for their use in teacher evaluation (McCaffrey et al., 2003; Braun, 2005). McCaffrey et al. (2005) and Wright (2004) explore the impact of the presence of missing data on VAMs, although they do not propose a model under a more relaxed assumption on the missing-data mechanism. Certain types of missing data require that the VAM be modeled jointly with the missingness process in order to obtain unbiased parameter estimates. To date, the only thorough investigation of the impact of missing data on VAMs by jointly modeling a missingness process comes from McCaffrey and Lockwood (2011). They use selection and pattern-mixture models, two particular types of joint models, for the missing data indicators with Bayesian inference, attributing missing data to intrinsic student – but not teacher – characteristics.

In the second part of this dissertation, we develop a new multi-response multi-membership mixed model that allows the missingness mechanism to depend on teachers as well as students. This model presents both theoretical and computational challenges due to 1) the need to jointly model a continuous and a binary response and 2) the non-nested, multi-membership random effects structure needed to account for student movement across classrooms. We obtain the maximum likelihood estimates with an EM algorithm, which provides a stable maximization procedure in light of the presence of highly-correlated random effects, as well as a dimensionality reduction feature. Steele (1996) proposes using a fully exponential Laplace approximation in the E-step of an EM algorithm to estimate generalized linear mixed models. Rizopoulos et al. (2009) use this approach to estimate a shared parameter model for a longitudinal and a time-to-dropout process. We generalize several aspects of their technique in order to estimate our model. Instead of using shared random effects, we use correlated random effects (Lin et al., 2009) between the GP VAM (Mariano et al., 2010) and a binary missing data mechanism. Furthermore, the multi-membership structure of the VAM leads to computational difficulties not faced in the nested model presented by Rizopoulos et al. (2009).

We use various structures that are available within our multi-response model to perform a sensitivity analysis (Xu and Blozis, 2011) on the teacher rankings produced when analyzing a data set containing semester calculus grades from a large public university. We find that the rankings of teacher effects may change depending on the assumptions made about the structure of the missing data mechanism. This is an important finding given the high-stakes decisions that these rankings may be used to make.

Chapter 2 provides background on VAMs and methods for computing estimates. The EM algorithm for the GP model appears in Chapter 3. A new, joint model for student outcomes and student attendance is presented in Chapter 4. The joint model allows for models to be fit to data under various assumptions about the nature of the missing data, providing a sensitivity analysis to the missing at random assumption of the GP model. Finally, Chapter 5 applies both the GP and the joint model to real and simulated data sets.

Chapter 2

LITERATURE REVIEW

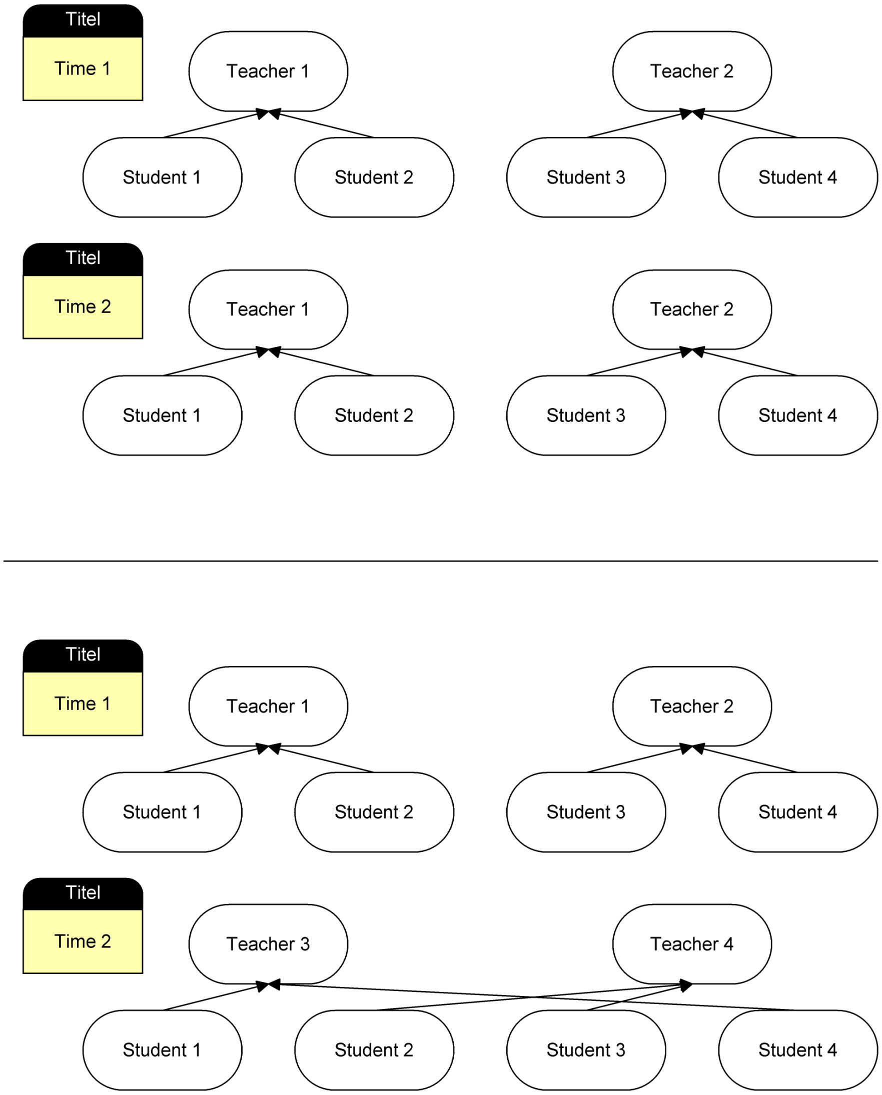
Over the past decade, value-added models have grown in popularity in the field of education as a tool for measuring teacher performance. The No Child Left Behind Act of 2001 (NCLB) mandates the use of state-wide standardized tests to evaluate school and district performance. NCLB focuses on identifying low-performing schools; however, the method of analysis for the standardized test scores is left up to the states. If an overly simplistic model is used to analyze the test scores, some schools may be identified as low-performing due to the composition of their student bodies. For example, schools in rural or impoverished areas may perform worse with respect to standardized tests simply because of the backgrounds of their students upon enrolling. Using simple score averages would provide a valid means of comparing schools if students were randomly assigned to schools across the state, but this is not the case. As a result, covariates such as free-lunch or migrant status are often included in the analyses in an attempt to account for the non-random assignment of students to schools and classes. The Race to the Top program places a greater emphasis on the use of student test scores to measure the performance of individual teachers than NCLB, resulting in a recent surge in the use of VAMs (U.S. Department of Education, 2009).

In an attempt to better control for student background characteristics, many VAMs use multiple years of student data and focus on student growth over time. A general VAM framework and related issues for education data are described by McCaffrey et al. (2004), and expanded upon by Lockwood et al. (2007). The nature of student movement through a school system leads to a complex relationship between students (level 1) and teachers (level 2) in a multi-level model. Since units at level 1 belong to multiple level 2 classes, the result is a multi-membership model (Browne et al., 2001). See Figure 2.1 for diagrams comparing nested and multi-membership structures. Some VAMs (Sanders et al., 1997; Rowan et al., 2002; McCaffrey et al., 2003, 2004, 2005; Lockwood et al., 2007; Mariano et al., 2010; Harris and McCaffrey, 2010) use mixed models for longitudinal student scores, modeling the score with random teacher intercepts. Under this scenario, the empirical best linear unbiased predictors (EBLUPs) for random teacher-intercepts serve as an estimate for the value-added to student learning by the individual teachers.

A heuristic summary of a mixed model VAM is in order. Suppose that students take a diagnostic test (e.g., in mathematics) at the end of each year in high school. In addition to keeping track of each student's test scores, we will also record who their teacher was each year. We wish to model the students' scores across the years. For fixed effects, we will include a yearly mean, and have the option of including student or teacher level covariates. Correlation between scores from the same student will be modeled with student-specific random intercepts. Likewise, scores from students who took a class together will be correlated (via the influence of the teacher and environmental factors from the shared classroom), and this will be accounted for by a classroom-membership random effect called the *teacher effect*, or *classroom effect*.

Teacher effects are measured using random-intercepts in these VAMs. Random effects for teachers are used instead of fixed effects in part because of the shrinkage properties of random effects. The magnitude of the effects for teachers with relatively few students are down-weighted towards 0, the mean of the random teacher effects. The amount of shrinkage depends on the amount of teacher-to-teacher variation present relative to the size of the error variance. The estimation of variance components for teacher-intercepts, student-intercepts, and residuals pro-

Figure 2.1: Diagrams of a nested structure (top) and a multi-membership structure (bottom)



vides insight, via the intra-class correlation coefficients, into the proportion of variation in test scores due to teacher differences. If this proportion is high, it indicates that classroom heterogeneity is largely responsible for the variance of students' test scores from their expected baseline score. By contrast, if the proportion of variation due to teachers is low, it indicates that classroom heterogeneity is not as influential as other factors, such as natural student-to-student variation.

In an ideal situation for statistical analysis, students would be randomly assigned to schools and teachers each year. Under this scenario, the EBLUPs for random teacher-intercepts would serve as an estimate for the value-added to student learning by the individual teachers. However, students are not randomly assigned to classrooms. Far from being a random sample, students often cluster at the classroom level based on significant predictors for educational performance, such as free-lunch and migrant status. If such clustering factors are not sufficiently modeled by fixed effects, they will be included in the estimated teacher effects. Note that, as observed by Lockwood et al. (2007), these effects (EBLUPs for teacher-intercepts) measure "unexplained heterogeneity at the classroom level," and not necessarily the causal effect of the teacher. Thus value-added models should be interpreted with caution, since students are not randomized to classrooms and the teacher effects may be influenced by classroom-level differences.

A relevant discussion of the utility and limitations of nested models (a specific form of VAM which might arise, say, from each student taking the same teacher every year) appears in Draper (1995). Complications for nested models also affect the more general VAMs. Draper (1995) urges a careful examination of the nature of the sampling in the study. The assumptions made by the sampling method or experimental design about 1) exchangeability of sampled and unsampled units in the target population and 2) the strong ignorability of treatment assignment (independence of outcome and experimental assignment given covariates) should be carefully examined. This is true, of course, for any method of statistical inference, but Draper (1995) sought to temper instances of overzealous interpretation of multilevel studies as they became more popular in educational settings.

The warning from Draper (1995) highlights a couple of restrictions on the interpretation of VAMs. Consider the example of college students taking a calculus course. Since the students taking calculus do not represent a random sample from the population of college students (let alone 18-22 year-olds in general), the results should not be generalized to serve as an estimation of teacher effectiveness *per se*, but rather should be understood as an estimate of teacher effectiveness on students "similar" to those that took the classes in the study.

The second concern of strong ignorability of treatment assignment relates to the non-random process by which students are assigned to classes. College students (or parents and counselors of elementary school students) select their classroom based on the time and location of the class, but more importantly based in part on their knowledge of the reputation of the instructor and in part on which class their friends pick. These influences on classroom selection may lead groups of low-achieving students to cluster and, likewise, groups of high-achieving students to cluster. The inclusion of student-specific random effects, referred to as the student's "general level of achievement" by McCaffrey and Lockwood (2011), should help ameliorate the effects of self-assignment, but it does not guarantee strong ignorability. This limits the extent to which teacher random effects may be considered as individual teacher contributions. It emphasizes the interpretation of the teacher random effects as "unexplained heterogeneity at the classroom level" (Lockwood et al., 2007). Thus it may be more accurate to refer to the classroom-membership random effects as "classroom effects" instead of "teacher effects." This is one of the reasons that the National Academy of Sciences cautioned that VAMs are not ready for high-stakes decisions. Nevertheless, these effects are being used to rank teachers, as in the Los Angeles Times article (Felch et al., 2010). If a data set contained multiple years of observations on teachers, it would be possible to separate the teacher and the classroom effects; however, we will leave this distinction aside for now and refer to the effects interchangeably as “teacher” and “classroom” effects.

2.1 Modeling Persistent Teacher Effects

After the first year of observations, we wish not only to model the effect of the current teacher on the student, but also to attribute their growth (or decline) to their past teachers. We would, however, expect the contribution of a teacher on a particular student to diminish over time. A challenge in the design of VAMs has been deciding how a student’s performance should be attributed to his or her current and prior teachers. The simplest VAMs examine each year separately and estimate the teacher effect by the average gain score of the teacher’s students. If the effects of good teaching persist, however, one would expect that students of a good teacher in year 1 would do well on the test in year 1 and would continue to do well in future years. One of the most popular VAMs in current use is the Educational Value-Added Assessment System (EVAAS) based on the model in Sanders et al. (1997). This model, sometimes called a “layered model,” assumes that the effect of a teacher persists undiminished over all subsequent years of his or her students’ achievement. That is, if a teacher is estimated to “add” 3 points to predicted student scores in year 1, those students continue to have those 3 points added to their predicted scores forever. Thus, a student’s predicted score in year 3 includes the teacher effects from the teachers in years 1, 2, and 3. The EVAAS model is implemented in SAS software (Wright et al., 2010) and through an R program (Lockwood et al., 2003), so it is readily usable by states and school districts.

The complete persistence assumption of the EVAAS model, however, limits its flexibility for modeling student achievement. A model proposed by McCaffrey et al. (2004) allows the effect of a teacher on students' scores to decay in future years, which Lockwood et al. (2007) termed variable persistence. Whereas the EVAAS model assumes that students of an excellent first grade teacher ought to carry that advantage, unabated, for the rest of their education, the variable persistence model recognizes that differences in exam content and distances in time may reduce the influence of former classroom membership on future performance. However, the variable persistence model restricts the way in which future effects decay. If a first grade teacher has a current year effect that is 2 standard deviations above the mean, their effect on third grade scores will also be 2 standard deviations above the mean (Mariano et al., 2010).

Recently, Mariano et al. (2010) introduced a generalized persistence (GP) model that allows a much more general structure for the effects of a current teacher on future test scores. The GP model estimates a different effect for each teacher in the current year and each future year of the study. Thus, a year-1 teacher might be estimated to have an effect of 3 points on the scores in year 1, and 2 points on the scores in years 2 and 3. While the EVAAS model requires the effects of a teacher on all years to be the same, the GP model allows a general correlation structure for the effects of a teacher in different years. The general correlation structure allows much more detailed exploration of the patterns of teacher effects, but greatly complicates the problem of computing estimates. In particular, neither the coefficient matrix for the random effects nor the overall covariance matrix can be written in block diagonal form, so computational methods that have been developed for nested hierarchical models and other special cases of linear mixed models do not apply.

2.2 A complex random effects structure

Many papers on mixed models consider a hierarchical (nested) random effects structure. In our model, this would mean that the students could not move between classrooms. This would be satisfied, for example, if the measurements were longitudinal observations on grade school students during a single year, and the each student belonged to only one teacher for the entire study.

For illustration, suppose that a set of observations $\boldsymbol{y}_{i(k)}$ are made on student $i(k)$ during the course of a year, during which student i belongs exclusively to classroom k, where $i = 1, \ldots, n_k$ and $k = 1 \ldots m$. Let $\boldsymbol{y}$ be a concatenation of the student scores, η_k represent a random intercept for classroom k, and assume that the intra-student correlation is modeled in the error covariance matrix. The likelihood may be factored as a product of one-dimensional integrals

$$f(\boldsymbol{y}) = \prod_{k=1}^{m} \int \prod_{i=1}^{n_k} f(\boldsymbol{y}_{i(k)} | \eta_k) f(\eta_k) \mathrm{d}\eta_k$$

Because of the correlation structure in multi-membership VAMs, however, this factorization cannot occur and we are left with a single, large integral. Let $\boldsymbol{\eta}$ be a concatenation of the random teacher effects. If students move between classrooms and their scores are modeled as a function of multiple teacher effects, the likelihood must be expressed as

$$f(\boldsymbol{y}) = \int \cdots \int f(\boldsymbol{y} | \boldsymbol{\eta}) f(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta} \tag{2.1}$$

Students in the same classroom share a teacher effect: as they move to other classes, they share teacher effects with their new classmates. When persistence effects are modeled, student scores in the second year are modeled as a function of multiple teacher effects (the current year effect of the second year teacher and the future year effect of the first year teacher), hence the name "multi-membership model" (Browne et al., 2001). Although a closed form solution exists for Equation (2.1) when continuous scores are modeled, this will not be the case when binary missing data indicators are included. We address this difficultly in Chapter 4. Even though a closed form solution exists when missing data are ignored, it is computationally expensive to compute since it is a function of large, dense matrices (due to the multi-membership structure). Only very small data sets may be fit with such models in SAS. We introduce a scalable, efficient method for these computations in Chapter 3.

Mariano et al. (2010) use Bayesian methods to estimate the parameters for the GP model using data from a large urban school district. To obtain a proper posterior distribution, however, a Bayesian approach to computations requires that an informative prior distribution be adopted for the covariance parameters, and different priors often result in different estimates of model parameters and teacher effects. A maximum likelihood (ML) approach avoids the need for priors, although ML estimation of even the variable persistence model, which is a subset of the GP model, has been "practically infeasible for all but small data sets" up to this point since "the multiple membership structure makes likelihood estimation difficult for realistically sized data sets" (Lockwood et al., 2007). In this paper we use the sparseness of the covariance and design matrices to develop an efficient EM algorithm for calculating ML estimates of parameters in the GP model. We implement the method in a user-friendly package in R statistical software (R Development Core Team, 2012) called GPvam. This development makes the GP model more accessible for use in practice, and provides an alternative to the Bayesian calculations implemented by Mariano et al. (2010). Application of the proposed methods to data from a large urban school district demonstrates the capabilities of the estimation procedure and software.

2.3 Nonignorable Missing Data

An often neglected aspect of value-added modeling is the effect of missing data on the results of the analysis (McCaffrey and Lockwood, 2011). Analysis of incomplete student profiles requires assumptions about the nature of the missing data. We use $\boldsymbol{y}$ to represent student test scores, and partition $\boldsymbol{y}$ into the set of scores $\boldsymbol{y}^o$ that were observed, and the measurements $\boldsymbol{y}^m$ that were planned but not observed. The vector $\boldsymbol{r}$ contains binary indicators for whether or not each planned observation was made, letting a value of 1 indicate a successful observation.

Data may be missing from a study for several reasons, and the cause of the missingness determines the degree to which the missing data affect the analysis. If data are missing completely at random (MCAR), then the joint likelihood of the longitudinal and missingness processes factors cleanly, and there is no need for joint modeling, since the longitudinal and missingness processes are independent. Likewise, if the data are missing at random (MAR) and the parameters for the longitudinal and missingness processes are distinct, then the missing data mechanism is said to be ignorable for likelihood inference (Little and Rubin, 2002, p. 119). However, if the missing data are missing not at random (MNAR) or if the parameter spaces of the longitudinal and missingness processes are not distinct, then the missing data are nonignorable and the longitudinal process, $f(\boldsymbol{y})$, and missingness processes, $f(\boldsymbol{r})$, must be modeled jointly to avoid bias in the estimation of the parameters of the longitudinal process. Models for nonignorable missing data require a factorization of the joint likelihood $f(\boldsymbol{y}, \boldsymbol{r})$ of the longitudinal and missingness mechanisms via one of three broad frameworks: selection, pattern-mixture, and shared-parameter models (Verbeke and Molenberghs, 2000).

- **Selection Models:** $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{r}|\boldsymbol{y})f(\boldsymbol{y})$

  Selection models require a marginal model for $\boldsymbol{y}$ and a conditional model $\boldsymbol{r}|\boldsymbol{y}$ to describe the factorization $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{r}|\boldsymbol{y})f(\boldsymbol{y})$ (Verbeke and Molenberghs, 2000, p. 234). The conditional model allows the probability of dropout to depend on the complete data (both observed and missing). For instance, McCaffrey and Lockwood (2011) modeled the number of observed scores to be dependent on each student's general level of achievement.

- **Pattern-Mixture Models:** $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{y}|\boldsymbol{r})f(\boldsymbol{r})$

  Pattern-mixture models begin with a marginal model for $\boldsymbol{r}$ and a conditional model for $\boldsymbol{y}|\boldsymbol{r}$ to describe the factorization $f(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{y}|\boldsymbol{r})f(\boldsymbol{r})$ (Verbeke and Molenberghs, 2000, p. 276). Pattern-mixture models posit a different test score model for each dropout pattern, or simply for each distinct number of dropouts. That is, the observed data model is stratified across the number of missing observations (Yuan and Little, 2009). Students with no missing observations are modeled together, students with one missing observation are modeled together, etc.

- **Shared-Parameter Models (SPM):** $f(\boldsymbol{y}, \boldsymbol{r}) = \int f(\boldsymbol{y}|\boldsymbol{\eta})f(\boldsymbol{r}|\boldsymbol{\eta})f(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}$

  Both the observed data and missingness models depend on the same random effects, and are independent, conditioned on these effects (Wu and Carroll, 1988). In the context of VAMs, the random effects are the student and teacher effects. We might expect students with higher scores to continue to stay in the program, and better teachers to be more likely to graduate students who continue their studies. The dependence of both the longitudinal and missing data mechanisms on the random effects, and thus the distribution of the random effects, means that the missingness process is not ignorable.

A discussion of the types of missing data appears in Little and Rubin (2002). These definitions fit conveniently in the selection model framework, but can be extended to the other model factorizations. Missing observations are missing completely at random (MCAR) if the missingness process is independent of the observed and the missing data. In selection models, MCAR implies $f(\boldsymbol{r}|\boldsymbol{y^o}, \boldsymbol{y^m}, \boldsymbol{\eta}) = f(\boldsymbol{r})$, where f denotes a density function. If the missingness process depends on the observed data but not the missing data, it is said to be missing at random (MAR). MAR implies $f(\boldsymbol{r}|\boldsymbol{y^o}, \boldsymbol{y^m}, \boldsymbol{\eta}) = f(\boldsymbol{r}|\boldsymbol{y^o})$. If, in addition to being MAR, the parameter spaces of $f(\boldsymbol{r}|\boldsymbol{y})$ and $f(\boldsymbol{y})$ are distinct, the missingness process is said to be ignorable. If the missingness process depends on the unobserved values, the missing observations are said to be missing not at random (MNAR). If observations are MNAR or if the parameter spaces are not distinct, the missingness process is nonignorable for maximum-likelihood based inference.

The GP model assumes that missing data are MAR. Inference is intended to be on $\boldsymbol{y} = (\boldsymbol{y}^o, \boldsymbol{y}^m)$, but only $\boldsymbol{y}^o$ have been observed. With non-ignorable missing data, $f(\boldsymbol{y}^o)$ is not the correct likelihood to maximize because $\boldsymbol{r}$ provides information about the distribution of $\boldsymbol{y}$. To obtain unbiased parameter estimates for the longitudinal process $\boldsymbol{y}$, the longitudinal and missingness processes must be modeled jointly and $f(\boldsymbol{y}^o, \boldsymbol{r})$ must be maximized. If we were to naively run an analysis on a data set with MNAR data in SAS `PROC MIXED`, SAS would ignore all observations for which the response was missing, yielding biased parameter estimates (Fitzmaurice et al., 2004; Verbeke and Molenberghs, 2000).

McCaffrey et al. (2005) and Wright (2004) explore the impact of the presence of missing data, including MNAR data, on VAMs. However, they do not attempt a joint analysis of the test-scores and missingness. To date, the most thorough investigation of the impact of non-ignorable missing data on VAMs by jointly modeling a missingness process comes from McCaffrey and Lockwood (2011). They use selection and pattern-mixture models to model the missing data in a Bayesian framework. When applied to a dataset from a large urban school district, they find that the MNAR analysis yielded approximately the same teacher effects as the MAR analysis. This, as McCaffrey and Lockwood (2011) admit, is a paradoxical result, since we would expect some teachers' effects to be influenced by the dropout of low-performing students. Naively, it seems that if a teacher's worst students all dropped out in the next year, that the teacher's estimated effect would be biased upward. Our new MNAR approach – using a variant of a shared-parameter model – will provide additional insight into the issue by allowing a student's propensity for missingness to depend on his or her teacher history.

Shared- and Correlated-Parameter Models

Instead of assuming that the full data $\boldsymbol{y}$ depend on the pattern of missingness or on the number of missing observations as in McCaffrey and Lockwood (2011), the SPM assumes that the longitudinal and missing mechanisms are conditionally independent, given a set of random effects. In our context, the assumption of conditional independence means that the student test scores and dropout patterns are independent given (1) the students' general levels of achievement and the propensity of each student to drop out and (2) the teacher effects and the propensity of teachers to pass students who do not drop out.

The original paper on SPM examines joint modeling of missing and observed data from a right-censored process on normally distributed data (Wu and Carroll, 1988). The name "shared-parameter model" first appears in Follmann and Wu (1995). A detailed, application-focused study that considers different models for continuous dropout data (survival times) as well as giving an implementation in SAS `PROC NLMIXED` appears in Vonesh et al. (2006). The SPM is useful both in joint modeling of observed and missing data and in joint modeling of observed and survival data. The difference between the two is in the structure of the density $f(\boldsymbol{r}|\boldsymbol{\eta})$. We will only consider the missing data modeling, but the relationship is noteworthy since some results we need later have been presented in the framework of survival time modeling.

A disadvantage of shared parameter models is that the random effects in the observed model and the random effects in the missingness model are perfectly correlated and are restricted to have the same variance in each model. This may not be realistic, depending on the units of measurement in each model. In addition, it may be unreasonable to expect the subject-specific effects to be the same in each model. An alternative to building "shared" random effects into the longitudinal and missingness models would be to build "correlated" random effects into the models, as done by Lin et al. (2009). Furthermore, under the correlated random effects factorization, the random teacher effects from the VAM enter the joint model in a linear fashion, which should improve the accuracy of the fully exponential Laplace approximation used in the E-step of an EM algorithm (see Chapter 4). For convenience, we will refer to the model using correlated random effects as a "correlated-parameter model" (CPM).

Sensitivity Analysis

When jointly modeling MNAR data, the missing data mechanism makes untestable assumptions about the nature of the relationship between the observed and missing data processes (Verbeke and Molenberghs, 2000). Molenberghs et al. (2008) show that it is not possible to perform an overall test of MNAR versus MAR since every MNAR model has an MAR counterpart that provides the same fit to the observed data but different predictions for the unobserved data. The plausibility of the assumed model cannot be tested empirically, and as a result it is necessary to fit several alternatives of the missing data mechanism to check the sensitivity of the inference to the choice of joint modeling structure (Xu and Blozis, 2011).

The likelihood based GP VAM assumes that observations are MAR. If any of the MNAR models were to produce substantially different results from the standard GP model, this would indicate that the conclusions of the VAM depend on assumptions made about the nature of the missing observations. However, the appropriate missing data process cannot be chosen by empirical investigation of the observed data (including examination of the log-likelihood) since the observed data do not provide information to support one particular MNAR model over another (Fitzmaurice et al., 2004; Xu and Blozis, 2011). As stated by Molenberghs and Kenward (2007), "ignoring MNAR models is no different an option than shifting to one particular MNAR model, it is just much more convenient." Consequently, we will fit different MNAR models with various assumptions about the structure of the missing data mechanism to the data sets in Chapter 5. These models all fit in the framework of the correlated random effects model we present in Chapter 4. We compare the teacher rankings produced by the MNAR models to those of the MAR GP value-added model to test the sensitivity of the rankings to assumptions about the relationship between the longitudinal and missingness processes.

2.4 High-Dimensional Integral Approximation

Much research has been done on the subject of approximating one-dimensional integrals involving probability densities, but the higher-dimensional case is not encountered as frequently, since many non-linear mixed models used in practice involve nested random effects which produce integrals that factor into a product of one-dimensional integrals. Even in the one-dimensional case, the appropriate method depends on the geometry of the integrand and other problem-specific details. In a review paper, Evans and Swartz (1995) outline four classes of integral approximation methods.

1. Asymptotic Methods (including the Laplace approximation and Penalized Quasi-likelihood)
2. (Adaptive) Importance Sampling
3. Multiple Quadrature (Including Gaussian Quadrature and Quasirandom Quadrature)
4. Markov Chain Methods (e.g. Gibbs Sampler)

Gaussian quadrature (GQ) and adaptive Gaussian quadrature (AGQ) are infeasible and we rule them out immediately due to the “curse of dimensionality.” The integral model in Equation (2.1) is k-dimensional, where k is the number of student and teacher random effects, so for large data sets the computational demands of using GQ are immense. In our applications, k is in the thousands. The quadrature methods approximate integrals with respect to a kernel by a weighted average of the integrand at certain quadrature points. The GQ abscissas are predetermined and centered around $\mathbf{0}$, the expected value of the random effects. The AGQ abscissas are centered at the current estimate of the mode of the integrand. The calculation of the appropriate quadrature points (and weighting) is computationally intensive. A rule of thumb suggests that either 20 GQ points or 6 AGQ points give a good approximation in one dimension. In order to maintain this level of coverage in each dimension, a k-dimensional integral requires 20^k GQ points, or 6^k AGQ points.

The (adaptive) importance samplers work well in one dimension but become complicated as the dimensionality rises due to the difficulty of determining an appropriate importance sampler (Evans and Swartz, 2000). “The analysis of convergence for adaptive importance sampling is more difficult than importance sampling because of the dependence between iterations” (Evans and Swartz, 1995, 2000), and “convergence will not take place with a poor choice of importance sampler” (Evans and Swartz, 1995). The importance sampler is “not the most efficient choice” if using normally distributed random effects and error terms (Pinheiro and Bates, 1995).

The Markov Chain methods mostly rely on the Gibbs sampler, which, in high-dimensional settings, may experience convergence “so slow as to be impractical” (Evans and Swartz, 1995). Assessing the achievement of stationarity and determining the strength of serial correlations complicates the analysis (Evans and Swartz, 1995; de Leeuw and Meijer, 2008). After attempting to fit a multi-level model, de Leeuw and Meijer (2008, p. 357) report, “In the end, the MCMC approach required extensive computation and judging convergence proved something of an arcane art form.” Gibbs sampling would have been our second choice for an approximation method. However, in addition to the concerns about rate of convergence, we would like to avoid relying on priors if possible. Along with the other model parameters, Bayesian methods would require a prior for the random effects covariance matrix $\boldsymbol{G}$. The estimated matrix $\hat{\boldsymbol{G}}$ and thus the EBLUPs $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{G}}\boldsymbol{S}'\hat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$ are sensitive to the choice of prior, and we would like to avoid introducing this subjective component into the calculation of the teacher effects. Mariano et al. (2010) found that even minimally informative priors for the covariance matrix of the random effects tended to exert a heavy influence on the results.

The asymptotic methods have a major computational advantage over the other methods. The work in these problems lies in calculating second derivatives of the integrand with respect to the random effects. However, once these derivatives are calculated or programmed, the calculation of the approximation is relatively fast, compared to Markov Chain methods. Under regularity conditions (Evans and Swartz, 1995, 2000), the Laplace formula approximates integrals of the form

$$I(h) = \int_{\mathrm{A}} h(\boldsymbol{t}) e^{-\lambda k(t)} \mathrm{d}\boldsymbol{t}$$

by

$$\hat{I}(h) = h\left(\hat{\boldsymbol{t}}\right)(2\pi)^{d/2} \left|\lambda \boldsymbol{K}\left(\hat{\boldsymbol{t}}\right)\right|^{-1/2} e^{-\lambda k\left(\hat{t}\right)}$$

where $\hat{\boldsymbol{t}}$ is the global minimum of k, $\boldsymbol{K}$ is the Hessian of k, and d is the dimension of A. There is an option in PROC GLIMMIX to approximate the marginal likelihood via a Laplace approximation, but this requires the error variance matrix $cov(\boldsymbol{y}|\boldsymbol{\eta}) = \boldsymbol{R}$ to be proportional to an identity matrix. By contrast, our estimation technique allows the inclusion of so-called "$\boldsymbol{R}$-side effects," including heterogeneous error variance-components in each year. In the framework of the EM algorithm, we do not directly approximate the marginal likelihood $f(\boldsymbol{y}^o, \boldsymbol{r})$ as SAS does, but rather the conditional expectation of the complete data likelihood in the E-step, as will be discussed in Chapter 3. Furthermore, SAS PROC GLIMMIX does not take into account the sparse structure of the design and covariance matrices, leading to memory deficiencies for even small data sets.

The default estimation method used by SAS `PROC GLIMMIX` is a pseudo-likelihood (PL) linearization method. The link function of the generalized linear mixed model is linearized, generating pseduo-data to which a linear mixed model is applied, updating model parameters. This doubly iterative process continues until convergence. This method is equivalent to penalized quasi-likelihood (PQL), as long as the PL overdispersion parameter is fixed at 1, which is the default behavior for the binary distribution in `PROC GLIMMIX` (Wolfinger and O'Connell, 1993; Littell et al., 2006). PQL is also equivalent to the Lindstrom-Bates method (Lindstrom and Bates, 1990; Wolfinger and Lin, 1997; Demidenko, 2004). PQL consist of the Laplace approximation, less one term: it relies on one further approximation beyond the Laplace approximation. Breslow and Lin (1995) and Lin and Breslow (1996) show that PQL tends to produce bias in parameter estimates, especially a downward bias in variance components. Their paper proposes a modified PQL method, but also shows that the first–, and especially second–, order Laplace approximations perform much better than PQL. This is also the case when the number of random effects increases with the sample size (Shun, 1997), such as with the famous salamander mating data (McCullagh and Nelder, 1989).

Pinheiro and Bates (1995) compare applications of the Lindstrom-Bates (PQL, PL) method, the first-order Laplace approximation, GQ, AGQ, and importance sampling, concluding that the Laplace approximation and AGQ approximations give the “best mix of efficiency and accuracy.” The Laplace approximation is a special case of AGQ, where just one abscissa is used.

2.5 EM Algorithm

Even when ignoring the missing data mechanism and assuming that the missing data are ignorable, the closed-form solution for the integrals in the likelihood in Equation (2.1) involves products of large, dense matrices, making direct maximization infeasible. The dimension of these matrices increases further with the inclusion of the missing data mechanism. The missing data mechanism also introduces nonlinear functions of the random effects into the integrand. Our model requires a maximization procedure that is capable of handling these difficulties.

The Expectation-Maximization (EM) algorithm may be used by treating the random effects as missing data (Dempster et al., 1977; McLachlan and Krishnan, 2008). It was one of the first methods used to estimate linear mixed models by treating latent random effects as missing data (Laird and Ware, 1982). The E-step calculates the expectation of the complete-data likelihood, given the observed data and current parameter estimates, and the M-step maximizes the conditional expectation of the complete data likelihood, given the observed data and the current parameter estimates. However, the E-step itself contains intractable integrals that require approximation. We handle the problem with a fully exponential Laplace approximation to approximate the E-step of the EM algorithm (Tierney et al., 1989). This approach was first proposed as a method of estimating generalized linear models by Steele (1996).

The fully exponential Laplace approximation was developed by Tierney and Kadane (1986) in order to reduce the approximation error of posterior means of positive functions. Roughly, it works by expressing the integral to be approximated as a fraction of two integrals and then applying the Laplace approximation separately to the numerator and denominator. The approximation originally applied only to strictly positive functions, but was extended to arbitrary functions by Tierney et al. (1989). The correlated-parameter model paper (Lin et al., 2009) also uses an improved Laplace approximation. However, Lin et al. (2009) apply the approximation directly to the integral yielding the marginal likelihood. Instead, we apply the approximation to the E-step of an EM algorithm.

The fully exponential Laplace approximation was used recently by Rizopoulos et al. (2009) to approximate integrals arising from joint modeling of longitudinal data with survival times in a shared-parameter model. Their results, however, do not apply directly to our situation. The paper assumes that the random effects are nested and does not face the computational complexity of a multi-membership model. In addition, we wish to model the probability of missingness for each observation, not survival time. Furthermore, we use a generalization of the SPM by modeling correlated- (instead of shared-) random effects. Our correlated random effects model appears in Chapter 4, where we further discuss the additional computational complexities faced by our model over those handled by Rizopoulos et al. (2009).

In the next chapter, we develop computational methods for the GP model under the assumption that missing data are MAR. Then, in Chapter 4, we present the correlated-parameter model and methods for computing estimates of the model parameters. Finally, Chapter 5 presents applications of the model.

Chapter 3

EFFICIENT MAXIMUM LIKELIHOOD ESTIMATION OF THE GENERALIZED PERSISTENCE VALUE-ADDED MODEL

The generalized persistence value-added model (GP VAM) (Mariano et al., 2010) models student scores using information about the history of observations on each student and each student's teacher-history. It would be possible to add a school-level to the multilevel model, though we do not consider this type of structure here. We present an efficient method for obtaining maximum likelihood estimates (MLEs) for the GP VAM, which was originally presented in a Bayesian framework. No scalable computational methods for ML estimation of the GP VAM currently exist. As we shall discuss, the model may be specified in SAS, but can only be estimated for very small data sets.

3.1 The Generalized Persistence Model

Suppose a data set tracks a cohort of n students over T years. The GP model assumes a linear mixed model as follows:

$$y_{ig} = \boldsymbol{x}'_{ig}\boldsymbol{\beta} + \boldsymbol{s}'_{ig}\boldsymbol{\eta} + \epsilon_{ig} \tag{3.1}$$

where y_{ig} denotes the score for student i during year g, for $i = 1, \ldots, n$, and $g \in A_i$; A_i is the set of years in which student i is observed. Students are taught by one of m_g teachers in each year g. We will also refer to the vector of concatenated student scores, $\boldsymbol{y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n)'$, where $\boldsymbol{y}_i = (y_{ig})$. The matrix $\boldsymbol{X}$, with rows $\boldsymbol{x}'_{ig}$, is the design matrix for the vector $\boldsymbol{\beta}$ of student and teacher level fixed-effect covariates such as demographic information or years of teaching experience. The matrix $\boldsymbol{S}$, with rows $\boldsymbol{s}'_{ig}$, is the design matrix for the random teacher effects. The structure of $\boldsymbol{S}$ determines whether the VAM models complete, variable, or generalized persistence.

The random effects vector $\boldsymbol{\eta}$ contains random teacher intercepts. The GP model estimates the effect of teachers on students in the year that they teach them, their lasting effect on the next year's score, and so on. Following the notation of Mariano et al. (2010), we let $\theta_{g[jt]}$ represent the effect for the j-th grade-g teacher on a student's grade t score, for $t \geq g$. A grade $g = 1, \ldots, T$ teacher has $K_g = T - g + 1$ effects. Thus $\boldsymbol{\theta}_{g[j\cdot]}$ gives the vector of current and future year effects of the j-th grade g teacher. The vector $\boldsymbol{\eta}$ concatenates the $\boldsymbol{\theta}_{g[j\cdot]}$ effects for all grades and teachers. The model is able to distinguish between the persistence effect of former teachers and the current effect of the present teacher because the students are not nested at the teacher level. If, for example, all of the students of a grade-1 teacher went on to have the same grade-2 teacher, it would not be possible for the model to separate the persistence effect of the grade-1 teacher from the current effect of the year-2 teacher.

We structure $\boldsymbol{\eta}$ in a way that leads to a block-diagonal random effects covariance matrix,

$$\boldsymbol{\eta} = (\boldsymbol{\theta}'_{1[1\cdot]}, \ldots, \boldsymbol{\theta}'_{1[m_1\cdot]}, \boldsymbol{\theta}'_{2[1\cdot]}, \ldots, \boldsymbol{\theta}'_{2[m_2\cdot]}, \ldots, \theta_{T[1\cdot]}, \ldots, \theta_{T[m_T\cdot]})'. \tag{3.2}$$

The vector $\boldsymbol{\eta}$ is distributed as $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{G})$ where

$$\boldsymbol{G} = \text{blockdiag}\left(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_T, \ldots, \boldsymbol{\Gamma}_T\right), \tag{3.3}$$

with m_g copies each of $\boldsymbol{\Gamma}_g$, where each $\boldsymbol{\Gamma}_g$ is unstructured. The matrix $\boldsymbol{\Gamma}_g$ is square with K_g rows and gives the covariance of current and future year effects for teachers of grade g. The design matrix $\boldsymbol{S}$ of the random effects has rows s'_{ig}, which contain 1's in entries corresponding to teachers who could affect each response.

After ordering the data by student and then by year, the error terms $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_n')'$ are distributed as $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{R})$ where $\boldsymbol{R}$ is a block diagonal matrix with blocks

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{T1} \\ \vdots & \ddots & \vdots \\ \sigma_{T1} & \cdots & \sigma_{TT} \end{pmatrix}. \tag{3.4}$$

If student i is missing an observation, then $\boldsymbol{R}_i$ omits the corresponding row and column corresponding to the year in which the observation is missing. $\boldsymbol{R}_i$ depends on i only through the dimension. In addition, we assume $\operatorname{cov}(\boldsymbol{\eta}, \boldsymbol{\epsilon}) = \mathbf{0}$.

Based on the GP model (3.1), the log-likelihood based on the observed data $\boldsymbol{y}$ is

$$l(\boldsymbol{\Psi}; \boldsymbol{y}) \propto -\frac{1}{2} \log |\boldsymbol{V}| - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.5}$$

where $\boldsymbol{V} = \boldsymbol{SGS'} + \boldsymbol{R}$, and $\boldsymbol{\Psi}$ is a vector of the unique model parameters from $\boldsymbol{\beta}$, $\boldsymbol{G}$, and $\boldsymbol{R}$. Due to the multi-membership data structure, $\boldsymbol{V}$ has no patterned structure (see Section 3.2 for an example), and its dimension is equal to the number of observations in the data set. As a result, a direct maximization of the likelihood function (3.5) is either highly inefficient or impossible for large data sets. For example, the application in Section 5.3 contains 26019 observations, which would result in a $\boldsymbol{V}$ matrix requiring over 5 GB of RAM to store in R. In order to develop a scalable method of estimation for the model, we must 1) use a method that requires numerical inversion of a matrix of reduced dimension from that of $\boldsymbol{V}$ and 2) utilize the sparseness of $\boldsymbol{S}, \boldsymbol{G}$, and $\boldsymbol{R}$. In Section 3.4 we consider maximum likelihood estimation of model (1) using an efficient EM algorithm based on the augmented data $(\boldsymbol{y}, \boldsymbol{\eta})$.

3.2 Alternative Model Specification

In the case when the scores from each year are measured on the same scale, an alternative model specification is available. This alternative model is used in the joint model in Chapter 4. Using a variable persistence structure for the teacher effects, McCaffrey and Lockwood (2011) model the intra-student correlation with random effects instead of in the error covariance matrix $\boldsymbol{R}$. We implement their model here, except we use the generalized persistence structure for teacher effects.

$$y_{ig} = \boldsymbol{x}'_{ig}\boldsymbol{\beta} + \boldsymbol{s}'_{ig}\boldsymbol{\eta} + \delta_i + \epsilon_{ig} \tag{3.6}$$

The terms in Equation (3.6) are defined the same as they were in Equation (3.1), with the exception of ϵ_{ig} and the new term δ_i. Instead of modeling $\boldsymbol{\epsilon}_i$ with an unstructured covariance matrix, we model a separate error variance in each year $\epsilon_{ig} \sim N(0, \sigma^2_g)$. As a result, $\boldsymbol{R}$ is diagonal with entries from the set $\{\sigma^2_1, \ldots, \sigma^2_T\}$, corresponding to the year of the observation. This is a new expression for $\boldsymbol{R}$, which was originally defined in Equation (3.4). We do not introduce a new notation, because several of the steps in the EM algorithm result in the same operation on $\boldsymbol{R}$, regardless of its definition. The appropriate version of $\boldsymbol{R}$ depends on whether the original GP or alternative model is being used. We likewise offer new definitions for $\boldsymbol{G}, \boldsymbol{S}$ and $\boldsymbol{\eta}$.

The δ_i are random student intercepts, with $\delta_i \sim N(0, \Gamma_{stu})$ and $\text{cov}(\epsilon_{ig}, \delta_i) = 0$. We may express Equation (3.6) by including the δ_i in the random effects vector $\boldsymbol{\eta}$,

$$\boldsymbol{\eta} = (\delta_1, \ldots, \delta_n, \boldsymbol{\theta}'_{1[1\cdot]}, \ldots, \boldsymbol{\theta}'_{1[m_1\cdot]}, \boldsymbol{\theta}'_{2[1\cdot]}, \ldots, \boldsymbol{\theta}'_{2[m_2\cdot]}, \ldots, \theta_{T[1\cdot]}, \ldots, \theta_{T[m_T\cdot]})'. \tag{3.7}$$

The vector $\boldsymbol{\eta}$ is then distributed as $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{G})$ where

$$\boldsymbol{G} = \text{blockdiag}\left(\Gamma_{stu}\mathbf{I}_n, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_T, \ldots, \boldsymbol{\Gamma}_T\right), \tag{3.8}$$

with m_g copies each of $\boldsymbol{\Gamma}_g$, where each $\boldsymbol{\Gamma}_g$ is unstructured. To accommodate the new $\boldsymbol{\eta}$, the design matrix $\boldsymbol{S}$ is composed of the blocks $[\boldsymbol{S}_1|\boldsymbol{S}_2]$, where $\boldsymbol{S}_1$ is the design matrix for the student effects and $\boldsymbol{S}_2$ is the design matrix for the teacher effects.

3.3 Examples of GP Structure

To illustrate the type of covariance structure induced by the generalized persistence model, we present an example. We first consider the alternative model of section 3.2. Table 3.1 provides an example with four students observed over two years, with two teachers in each year. The corresponding $\boldsymbol{S}$ matrix appears in Table 3.1, the $\boldsymbol{G}$ matrix appears in Equation (3.9), and the $\boldsymbol{R}$ matrix appears in Equation (3.10). In the $\boldsymbol{G}$ matrix, γ_{A1}^2 denotes the variance of the proximal effects of the year 1 teachers, γ_{A2}^2 is the variance of the future effects of the year 1 teachers, and γ_{A12}^2 is the covariance of the two effects. The variance of the proximal effect of year 2 teachers is denoted by γ_{B2}^2.

Table 3.1: Example Student Data with Teacher Links

Obs.	Year	Student	Teacher (by year)
1	1	S1	1
2	1	S2	1
3	1	S3	2
4	1	S4	2
5	2	S1	1
6	2	S2	2
7	2	S3	1
8	2	S4	2

Figure 3.1: Example Generalized Persistence Random Effects Design Matrix, $\boldsymbol{S}$, for Alternative Model

Obs.	δ_1	δ_2	δ_3	δ_4	$\theta_{1[11]}$	$\theta_{1[12]}$	$\theta_{1[21]}$	$\theta_{1[22]}$	$\theta_{2[12]}$	$\theta_{2[22]}$
1	1				1					
2		1			1					
3			1				1			
4				1			1			
5	1					1			1	
6		1				1				1
7			1					1	1	
8				1				1		1

$$\boldsymbol{G} = \begin{pmatrix} \gamma_S^2 & & & & & & & & & \\ & \gamma_S^2 & & & & & & & & \\ & & \gamma_S^2 & & & & & & & \\ & & & \gamma_S^2 & & & & & & \\ & & & & \gamma_{A1}^2 & \gamma_{A12} & & & & \\ & & & & \gamma_{A12} & \gamma_{A2}^2 & & & & \\ & & & & & & \gamma_{A1}^2 & \gamma_{A12} & & \\ & & & & & & \gamma_{A12} & \gamma_{A2}^2 & & \\ & & & & & & & & \gamma_{B2}^2 & \\ & & & & & & & & & \gamma_{B2}^2 \end{pmatrix} \tag{3.9}$$

$$\boldsymbol{R} = diag\left(\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_2^2\right) \tag{3.10}$$

The irregular structure of the design matrix for the random effects, as demonstrated in Figure 3.1, causes difficulties in the estimation of the model. In nested designs, it is possible to factor the likelihood into a product over the subjects, but that is not possible with the multi-membership structure of VAMs. Even though we have constructed $\boldsymbol{G}$ to be block-diagonal, the marginal variance $\boldsymbol{V}$ will not have a uniform, simplified structure as is the case for nested models. The $\boldsymbol{V} = [\boldsymbol{V}_1\ \boldsymbol{V}_2]$ matrix appears in two parts, in Equations (3.11) and (3.12).

$$\boldsymbol{V}_1 = \begin{pmatrix} \gamma_{\text{A1}}^2 + \gamma_S^2 + \sigma_1^2 & \gamma_{\text{A1}}^2 & 0 & 0 \\ \gamma_{\text{A1}}^2 & \gamma_{\text{A1}}^2 + \gamma_S^2 + \sigma_1^2 & 0 & 0 \\ 0 & 0 & \gamma_{\text{A1}}^2 + \gamma_S^2 + \sigma_1^2 & \gamma_{\text{A1}}^2 \\ 0 & 0 & \gamma_{\text{A1}}^2 & \gamma_{\text{A1}}^2 + \gamma_S^2 + \sigma_1^2 \\ \gamma_{\text{A12}} + \gamma_S^2 & \gamma_{\text{A12}} & 0 & 0 \\ \gamma_{\text{A12}} & \gamma_{\text{A12}} + \gamma_S^2 & 0 & 0 \\ 0 & 0 & \gamma_{\text{A12}} + \gamma_S^2 & \gamma_{\text{A12}} \\ 0 & 0 & \gamma_{\text{A12}} & \gamma_{\text{A12}} + \gamma_S^2 \end{pmatrix} \tag{3.11}$$

$$\boldsymbol{V}_2 = \begin{pmatrix} \gamma_{\text{A12}} + \gamma_S^2 & \gamma_{\text{A12}} & 0 & 0 \\ \gamma_{\text{A12}} & \gamma_{\text{A12}} + \gamma_S^2 & 0 & 0 \\ 0 & 0 & \gamma_{\text{A12}} + \gamma_S^2 & \gamma_{\text{A12}} \\ 0 & 0 & \gamma_{\text{A12}} & \gamma_{\text{A12}} + \gamma_S^2 \\ \gamma_{\text{A2}}^2 + \gamma_{\text{B2}}^2 + \gamma_S^2 + \sigma_2^2 & \gamma_{\text{A2}}^2 & \gamma_{\text{B2}}^2 & 0 \\ \gamma_{\text{A2}}^2 & \gamma_{\text{A2}}^2 + \gamma_{\text{B2}}^2 + \gamma_S^2 + \sigma_2^2 & 0 & \gamma_{\text{B2}}^2 \\ \gamma_{\text{B2}}^2 & 0 & \gamma_{\text{A2}}^2 + \gamma_{\text{B2}}^2 + \gamma_S^2 + \sigma_2^2 & \gamma_{\text{A2}}^2 \\ 0 & \gamma_{\text{B2}}^2 & \gamma_{\text{A2}}^2 & \gamma_{\text{A2}}^2 + \gamma_{\text{B2}}^2 + \gamma_S^2 + \sigma_2^2 \end{pmatrix} \tag{3.12}$$

It is interesting to examine the correlations between the different observations. The last column of matrix $\boldsymbol{V}_2$ corresponds to the last observation in Table 3.1. The component at bottom-right of $\boldsymbol{V}_2$, $\gamma_{\text{A2}}^2 + \gamma_{\text{B2}}^2 + \gamma_S^2 + \sigma_2^2$, gives the variance of the second year observation on student S4. It contains contributions from the year 1 teacher, the year 2 teacher, and the year 2 error variance. Moving up the last column of $\boldsymbol{V}_2$ reveals which observations are correlated with observation 8 (see Table 3.1). The covariance of γ_{A2}^2 with observation 7 is due the fact that students S3 and

S4 shared the same teacher in year 1. The covariance of γ^2_{B2} with observation 6 is due to the fact that S2 and S4 are classmates in year 2. The covariance $\gamma_{\text{A12}} + \gamma^2_S$ with observation 4 is due to the fact that that these are both observations on the same student (hence the γ^2_S term) and that that the current and future year effects of teachers from year 1 are correlated (hence the γ_{A12} term). This also explains the covariance with observation 3, which was a measurement on a year 1 classmate of student S4.

We next consider the structure of the original formulation of the GP model as described in Section 3.1. Unlike the alternative model, the student effects are modeled in the error covariance matrix instead of modeling them as random effects. The design matrix, $\boldsymbol{S}$ for the random effects appears in Figure 3.2, and the covariance matrix, $\boldsymbol{G}$, for the random effects appears in Equation 3.13.

Figure 3.2: Example Generalized Persistence Random Effects Design Matrix, $\boldsymbol{S}$

Obs.	$\theta_{1[11]}$	$\theta_{1[12]}$	$\theta_{1[21]}$	$\theta_{1[22]}$	$\theta_{2[12]}$	$\theta_{2[22]}$
1	1					
2	1					
3			1			
4			1			
5		1			1	
6		1				1
7				1	1	
8				1		1

$$\boldsymbol{G} = \begin{pmatrix} \gamma_{A1}^2 & \gamma_{A12} & & & & \\ \gamma_{A12} & \gamma_{A2}^2 & & & & \\ & & \gamma_{A1}^2 & \gamma_{A12} & & \\ & & \gamma_{A12} & \gamma_{A2}^2 & & \\ & & & & \gamma_{B2}^2 & \\ & & & & & \gamma_{B2}^2 \end{pmatrix} \tag{3.13}$$

This example illustrates why we sort the observations by student and then by year in our program to make the error covariance matrix, $\boldsymbol{R}$, block-diagonal. Table 3.1 is not sorted, and as a result the $\boldsymbol{R}$ matrix in Equation (3.14) is not block-diagonal. The term σ_{12} represents the covariance of year 1 and year 2 observations on the same student.

$$\boldsymbol{R} = \begin{pmatrix} \sigma_1^2 & & & & \sigma_{12} & & & \\ & \sigma_1^2 & & & & \sigma_{12} & & \\ & & \sigma_1^2 & & & & \sigma_{12} & \\ & & & \sigma_1^2 & & & & \sigma_{12} \\ \sigma_{12} & & & & \sigma_2^2 & & & \\ & \sigma_{12} & & & & \sigma_2^2 & & \\ & & \sigma_{12} & & & & \sigma_2^2 & \\ & & & \sigma_{12} & & & & \sigma_2^2 \end{pmatrix} \tag{3.14}$$

The resulting marginal covariance matrix $\boldsymbol{V} = [\boldsymbol{V}_1 \; \boldsymbol{V}_2]$ for this model formulation appears in two parts, in Equations (3.15) and (3.16).

$$\boldsymbol{V}_1 = \begin{pmatrix} \gamma_{\mathsf{A1}}^2 + \sigma_1^2 & \gamma_{\mathsf{A1}}^2 & 0 & 0 \\ \gamma_{\mathsf{A1}}^2 & \gamma_{\mathsf{A1}}^2 + \sigma_1^2 & 0 & 0 \\ 0 & 0 & \gamma_{\mathsf{A1}}^2 + \sigma_1^2 & \gamma_{\mathsf{A1}}^2 \\ 0 & 0 & \gamma_{\mathsf{A1}}^2 & \gamma_{\mathsf{A1}}^2 + \sigma_1^2 \\ \gamma_{\mathsf{A12}} + \sigma_{12} & \gamma_{\mathsf{A12}} & 0 & 0 \\ \gamma_{\mathsf{A12}} & \gamma_{\mathsf{A12}} + \sigma_{12} & 0 & 0 \\ 0 & 0 & \gamma_{\mathsf{A12}} + \sigma_{12} & \gamma_{\mathsf{A12}} \\ 0 & 0 & \gamma_{\mathsf{A12}} & \gamma_{\mathsf{A12}} + \sigma_{12} \end{pmatrix} \tag{3.15}$$

$$\boldsymbol{V}_2 = \begin{pmatrix} \gamma_{\mathsf{A12}} + \sigma_{12} & \gamma_{\mathsf{A12}} & 0 & 0 \\ \gamma_{\mathsf{A12}} & \gamma_{\mathsf{A12}} + \sigma_{12} & 0 & 0 \\ 0 & 0 & \gamma_{\mathsf{A12}} + \sigma_{12} & \gamma_{\mathsf{A12}} \\ 0 & 0 & \gamma_{\mathsf{A12}} & \gamma_{\mathsf{A12}} + \sigma_{12} \\ \gamma_{\mathsf{A2}}^2 + \gamma_{\mathsf{B2}}^2 + \sigma_2^2 & \gamma_{\mathsf{A2}}^2 & \gamma_{\mathsf{B2}}^2 & 0 \\ \gamma_{\mathsf{A2}}^2 & \gamma_{\mathsf{A2}}^2 + \gamma_{\mathsf{B2}}^2 + \sigma_2^2 & 0 & \gamma_{\mathsf{B2}}^2 \\ \gamma_{\mathsf{B2}}^2 & 0 & \gamma_{\mathsf{A2}}^2 + \gamma_{\mathsf{B2}}^2 + \sigma_2^2 & \gamma_{\mathsf{A2}}^2 \\ 0 & \gamma_{\mathsf{B2}}^2 & \gamma_{\mathsf{A2}}^2 & \gamma_{\mathsf{A2}}^2 + \gamma_{\mathsf{B2}}^2 + \sigma_2^2 \end{pmatrix} \tag{3.16}$$

When only two years are included in the data set, both the GP model of Section 3.1 and the alternative model of Section 3.2 yield the same fit. The terms σ_1^2 and σ_2^2 in Equations (3.15) and (3.16) account for both the student-to-student variation and the error variance. The covariance between observations on the same student is modeled by σ_{12}. When more than two years are included in the study, the model formulation from Section 3.1 models a different covariance between observations on the same student in different pairs of years, e.g. σ_{ij} between years i and j. On the other hand, the alternative model formulation in Section 3.2 imposes a compound-symmetric structure by modeling the same covariance σ_S between any pair of observations on the same student.

3.4 The EM Algorithm

The EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) provides a broad framework for maximum likelihood estimation in the presence of missing data. It was one of the first methods used to estimate linear mixed models by treating latent random effects as missing data (Laird and Ware, 1982). Assuming that the students are independent conditional on the effects of their current and previous teachers, estimation of the GP model (1) based on the augmented data $(\boldsymbol{y}, \boldsymbol{\eta})$ then becomes a much more scalable optimization problem.

Unbalanced observations on students are common in longitudinal studies. In this chapter we assume that students with incomplete profiles have observations missing at random and that the parameters governing the outcome process are distinct from those characterizing the missingness process, yielding a valid likelihood-based analysis under the specified model (Little and Rubin, 2002). Our VAM for incomplete data with observations missing not at random appears in Chapter 4.

We will refer to $f(\boldsymbol{y}; \boldsymbol{\Psi})$ as the observed data density function and

$$f(\boldsymbol{y}, \boldsymbol{\eta}; \boldsymbol{\Psi}) = f(\boldsymbol{y}|\boldsymbol{\eta}; \boldsymbol{\Psi}) f(\boldsymbol{\eta}; \boldsymbol{\Psi})$$

as the complete data density function, where

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\eta}; \boldsymbol{\Psi}) &\propto |\boldsymbol{R}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{S\eta}\right)' \boldsymbol{R}^{-1} \left(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{S\eta}\right)\right\} \\ f(\boldsymbol{\eta}; \boldsymbol{\Psi}) &\propto |\boldsymbol{G}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\eta}' \boldsymbol{G}^{-1} \boldsymbol{\eta}\right\} \end{aligned}$$

Given initial values for the parameters and the random effects, the EM algorithm alternates between an expectation (E) step and a maximization (M) step. At iteration (k + 1), the E step calculates the conditional expectation of the complete data log-likelihood, given the observed data, $\boldsymbol{y}$, and parameter estimates obtained in the k-th step, $\boldsymbol{\Psi}^{(k)}$. That is, the E step computes

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = \int \left\{\log f\left(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}\right) + \log f\left(\boldsymbol{\eta};\boldsymbol{\Psi}\right)\right\} f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}^{(k)})\mathrm{d}\boldsymbol{\eta}.$$

The M step then maximizes $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$ as if $\boldsymbol{\eta}$ were known, resulting in the updated parameter vector $\boldsymbol{\Psi}^{(k+1)}$ satisfying

$$\int \frac{\partial}{\partial \boldsymbol{\Psi}} \left\{\log f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) + \log f(\boldsymbol{\eta};\boldsymbol{\Psi})\right\} f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}^{(k)})\mathrm{d}\boldsymbol{\eta}\Big|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k+1)}} = \mathbf{0}, \qquad (3.17)$$

provided that differentiation and integration are interchangeable, which is valid because the complete data likelihood $f(\boldsymbol{y},\boldsymbol{\eta};\boldsymbol{\Psi})$ is a member of the exponential family (Lehmann and Romano, 2010). Note that the expression on the left side of Equation (3.17) equals the observed data score vector $S(\boldsymbol{\Psi};\boldsymbol{y}) = (\partial/\partial\boldsymbol{\Psi})\, l(\boldsymbol{\Psi};\boldsymbol{y})$ (Louis, 1982), as demonstrated by the following steps (McLachlan and Krishnan, 2008).

$$\begin{aligned}
S(\boldsymbol{\Psi}) =& \frac{\partial}{\partial \boldsymbol{\Psi}} \log f(\boldsymbol{y};\boldsymbol{\Psi})\\
=& \frac{\partial}{\partial \boldsymbol{\Psi}} \log \int f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) f(\boldsymbol{\eta};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta}\\
=& \frac{1}{f(\boldsymbol{y};\boldsymbol{\Psi})}\frac{\partial}{\partial \boldsymbol{\Psi}} \int f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) f(\boldsymbol{\eta};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta}\\
=& \frac{1}{f(\boldsymbol{y};\boldsymbol{\Psi})}\int \frac{\partial}{\partial \boldsymbol{\Psi}} \left\{f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) f(\boldsymbol{\eta};\boldsymbol{\Psi})\right\}\mathrm{d}\boldsymbol{\eta}\\
=& \int \left[\frac{\partial}{\partial \boldsymbol{\Psi}} \log\left\{f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) f(\boldsymbol{\eta};\boldsymbol{\Psi})\right\}\right] \frac{f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) f(\boldsymbol{\eta};\boldsymbol{\Psi})}{f(\boldsymbol{y};\boldsymbol{\Psi})}\mathrm{d}\boldsymbol{\eta}\\
=& \int \frac{\partial}{\partial \boldsymbol{\Psi}} \left\{\log f(\boldsymbol{y}|\boldsymbol{\eta};\boldsymbol{\Psi}) + \log f(\boldsymbol{\eta};\boldsymbol{\Psi})\right\} f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \qquad (3.18)
\end{aligned}$$

We next present details on the M-step and the E-step that computes the conditional expectations required in the M-step. A method on obtaining the asymptotic standard errors for parameter estimates is discussed and descriptions of the convergence and initial values of the EM algorithm are provided.

3.5 The M-Step for the GP Model

Using the definition of $\boldsymbol{G}$ in Equation (3.3), we may write the density of $\boldsymbol{\eta}$ as

$$
\begin{aligned}
f(\boldsymbol{\eta};\boldsymbol{\Psi}) &\propto \det(\boldsymbol{G})^{-1/2}\exp\left(-\frac{\boldsymbol{\eta}'\boldsymbol{G}^{-1}\boldsymbol{\eta}}{2}\right) \\
&= \left[\prod_{g=1}^{T}\det(\boldsymbol{\Gamma}_{\mathrm{g}})^{-m_g/2}\right]\exp\left(-\sum_{g=1}^{T}\sum_{j=1}^{m_g}\frac{\boldsymbol{\theta}'_{g[j\cdot]}\boldsymbol{\Gamma}_g^{-1}\boldsymbol{\theta}_{g[j\cdot]}}{2}\right)
\end{aligned}
$$

We use Petersen and Pedersen (2008) and Harville (2008) for matrix differentiation, and note that each $\boldsymbol{\Gamma}_g$ is symmetric. Referring to Equation (3.17), the score vector with respect to $\boldsymbol{\Gamma}_g$ is

$$
\begin{aligned}
S(\boldsymbol{\Gamma}_{\boldsymbol{g}}) &= \int \frac{\partial}{\partial\boldsymbol{\Gamma}_{\boldsymbol{g}}}\log\left[\det(\boldsymbol{G})^{-1/2}\exp\left(-\frac{\boldsymbol{\eta}'\boldsymbol{G}^{-1}\boldsymbol{\eta}}{2}\right)\right] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \\
&= \int \frac{\partial}{\partial\boldsymbol{\Gamma}_{\boldsymbol{g}}}\left[-\frac{m_g}{2}\log\left(\det\left(\boldsymbol{\Gamma}_{\boldsymbol{g}}\right)\right)\right] \\
&\quad + \frac{\partial}{\partial\boldsymbol{\Gamma}_{\boldsymbol{g}}}\left[-\frac{1}{2}\sum_{j=1}^{m_g}\boldsymbol{\theta}'_{g[j\cdot]}\boldsymbol{\Gamma}_g^{-1}\boldsymbol{\theta}_{g[j\cdot]}\right] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \\
&= \text{matrix with components } \begin{cases} d_{ij} & \text{if } i=j \\ 2d_{ij} & \text{if } i\neq j \end{cases}
\end{aligned}
$$

where d_{ij} is the ij-th component of the matrix

$$
\boldsymbol{D} = -\frac{1}{2}\left\{m_g\boldsymbol{\Gamma}_{\boldsymbol{g}}{}^{-1} - \boldsymbol{\Gamma}_{\boldsymbol{g}}{}^{-1}\left(\sum_{j=1}^{m_g}\mathrm{E}\left[\boldsymbol{\theta}_{g[j\cdot]}\boldsymbol{\theta}'_{g[j\cdot]}|\boldsymbol{y};\boldsymbol{\Psi}\right]\right)\boldsymbol{\Gamma}_{\boldsymbol{g}}{}^{-1}\right\}
$$

Let

$$
\widetilde{\boldsymbol{\eta}} = \mathrm{E}[\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}] \tag{3.19}
$$

$$
\widetilde{\boldsymbol{v}} = \mathrm{var}[\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}] \tag{3.20}
$$

represent the conditional expectation and variance, respectively, of $\boldsymbol{\eta}$. These quantities are calculated in the E-step and remain fixed during the M-step. Likewise, let the sub-vector of $\widetilde{\boldsymbol{\eta}}$ corresponding to $\mathrm{E}[\boldsymbol{\theta}_{g[j\cdot]}|\boldsymbol{y};\boldsymbol{\Psi}]$ be denoted $\widetilde{\boldsymbol{\theta}}_{g[j\cdot]}$, and the block of the matrix $\widetilde{\boldsymbol{v}}$ corresponding to $\mathrm{E}[\boldsymbol{\theta}_{g[j\cdot]}\boldsymbol{\theta}'_{g[j\cdot]}|\boldsymbol{y};\boldsymbol{\Psi}]$ be denoted $\widetilde{\boldsymbol{v}}_{g[j\cdot]}$. Now, since $\widetilde{\boldsymbol{v}} = \mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}'|\boldsymbol{y};\boldsymbol{\Psi}] - \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$, setting $S(\boldsymbol{\Gamma}_g) = \mathbf{0}$ implies

$$m_g \boldsymbol{\Gamma_g}^{-1} = \boldsymbol{\Gamma_g}^{-1} \sum_{j=1}^{m_g} \left(\widetilde{\boldsymbol{v}}_{g[j\cdot]} + \widetilde{\boldsymbol{\theta}}_{g[j\cdot]} \widetilde{\boldsymbol{\theta}}'_{g[j\cdot]} \right) \boldsymbol{\Gamma_g}^{-1}$$

Thus the M-step update for $\boldsymbol{\Gamma_g}$ is

$$\widehat{\boldsymbol{\Gamma}}_g = \frac{1}{m_g} \sum_{j=1}^{m_g} \left(\widetilde{\boldsymbol{v}}_{g[j\cdot]} + \widetilde{\boldsymbol{\theta}}_{g[j\cdot]} \widetilde{\boldsymbol{\theta}}'_{g[j\cdot]} \right) \tag{3.21}$$

Equation (3.21) calculates an average of the blocks of $\widetilde{\boldsymbol{v}} + \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$ that correspond to teachers who taught in year g.

The M-step update for $\boldsymbol{\beta}$ is the value that solves $S(\boldsymbol{\beta}) = 0$, where

$$S(\boldsymbol{\beta}) = \int \frac{\partial}{\partial \boldsymbol{\beta}} \left[-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{S}\boldsymbol{\eta})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{S}\boldsymbol{\eta}) \right] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}) \mathrm{d}\boldsymbol{\eta} \tag{3.22}$$

$$= \boldsymbol{X}'\boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{S}\widetilde{\boldsymbol{\eta}}), \tag{3.23}$$

namely,

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{S}\widetilde{\boldsymbol{\eta}}) \tag{3.24}$$

Equation (3.24) provides some insight into the relationship between the fixed and random effects. Compare Equation (3.24) to the generalized least-square (GLS) estimator of $\boldsymbol{\beta}$, namely

$$(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{3.25}$$

where $\boldsymbol{V} = \boldsymbol{S}\boldsymbol{G}\boldsymbol{S}' + \boldsymbol{R}$. The predicted random effects, $\widetilde{\boldsymbol{\eta}}$, are assumed to be known and fixed in the M-step. Heuristically, we may think of $\widetilde{\boldsymbol{\eta}}$ as having no variance in the M-step, meaning $\boldsymbol{G} = \mathbf{0}$ in the GLS Equation (3.25). This would reduce $\boldsymbol{V}^{-1}$ to $\boldsymbol{R}^{-1}$. The known values of $\widetilde{\boldsymbol{\eta}}$ are used to account for the subject-specific effects via the term $\boldsymbol{y}_i - \boldsymbol{s}_i'\widetilde{\boldsymbol{\eta}}$ in Equation (3.24). This provides an interpretation of the fixed effects of a linear mixed model as a summary of the "residual" information remaining after sweeping out subject-specific effects.

Table 3.2: Parameterizing the attendance patterns for example with 3 years

Attendance indicators	Pattern
001	1
010	2
011	3
100	4
101	5
110	6
111	7

The calculation of the M-step update for $\boldsymbol{R}$ from Equation (3.4) is complicated by the fact that the structure of $\boldsymbol{R}$ changes in the presence of unbalanced data. The M-step update for the component σ_{kl} of $\boldsymbol{R}$ is the value that solves $S(\sigma_{kl}) = 0$, where

$$\begin{aligned} S\left(\sigma_{kl}\right) = \int \frac{\partial}{\partial \sigma_{kl}} \Bigg[& \log\left(|\boldsymbol{R}|^{-1/2}\right) \\ & -\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{S\eta}\right)' \boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{S\eta}\right)\Bigg] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta}. \end{aligned}$$

If the observations are sorted by students and then by year, $\boldsymbol{R}$ is block-diagonal with block sizes depending on the number of observations on each student. For T years, there are $2^T - 1$ possible combinations of years in which a student may be observed, although not all of these patterns may appear in a given data set. To parameterize these combinations, we treat the ordered, binary attendance-indicators for each student as a number in base-2. So in a study over three years, each student will have an attendance pattern from the first column of Table 3.2.

For example, a student with observations in each year has pattern 7, with the corresponding block of $\boldsymbol{R}$ given by

$$\begin{pmatrix} \sigma_{11} & \sigma_{21} & \sigma_{31} \\ \sigma_{21} & \sigma_{22} & \sigma_{32} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}.$$

The matrices corresponding to the other patterns are subsets of this matrix, using the rows and columns suggested by the attendance indicator. A student who is missing an observation in year 2 has pattern 5 and corresponding error covariance matrix

$$\begin{pmatrix} \sigma_{11} & \sigma_{31} \\ \sigma_{31} & \sigma_{33} \end{pmatrix}.$$

Let p denote the attendance pattern, n_p be the number of students with that pattern, and $\boldsymbol{R}_{(p)}$ represent the covariance matrix corresponding to the p-th pattern. In addition, let P_{kl} denote the set of patterns p whose covariance matrix $\boldsymbol{R}_{(p)}$ contains σ_{kl}. Furthermore, let $b(p)$ denote the b-th student with pattern p. We may write

$$|\boldsymbol{R}| = \prod_p \left|\boldsymbol{R}_{(p)}\right|^{n_p}.$$

Thus the score function may be expressed as

$$\begin{aligned} S\left(\sigma_{kl}\right) = -\frac{1}{2} \int \frac{\partial}{\partial \sigma_{kl}} \Bigg\{ \sum_p n_p \log\left|\boldsymbol{R}_{(p)}\right| + \sum_p \sum_b \Big[& \\ \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta} - \boldsymbol{S}_{b(p)}\boldsymbol{\eta}\right)' \boldsymbol{R}_{(p)}^{-1} \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta} - \boldsymbol{S}_{b(p)}\boldsymbol{\eta}\right) \Big] \Bigg\} & \\ \times f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} & \end{aligned}$$

where $\boldsymbol{y}_{b(p)}$ is the vector of observations from student $b(p)$, with corresponding design matrices for fixed and random effects $\boldsymbol{X}_{b(p)}$ and $\boldsymbol{S}_{b(p)}$. The derivative will be 0 for all terms that do not contain the parameter σ_{kl}. This includes observations on students who do not have observations in both years i and j. Then, taking the derivative and letting $1_{\{\mathcal{C}\}}$ be the indicator function that takes the value 1 if condition $\mathcal{C}$ is true and 0 otherwise,

$$\begin{aligned} S\left(\sigma_{kl}\right) = &- \left(1_{\{k \neq l\}} + \frac{1}{2} \times 1_{\{k=l\}}\right) \sum_{p \in P_{kl}} \left\{ n_p \left(\boldsymbol{R}_{(p)}^{-1}\right)_{\{kl\}} - \right. \\ &\int \sum_{b} \left[\boldsymbol{R}_{(p)}^{-1} \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta} - \boldsymbol{S}_{b(p)}\boldsymbol{\eta}\right) \right. \\ &\left.\left. \times \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta} - \boldsymbol{S}_{b(p)}\boldsymbol{\eta}\right)' \boldsymbol{R}_{(p)}^{-1} \right]_{\{kl\}} f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \right\}. \end{aligned}$$

The notation $\{kl\}$ indicates the matrix component corresponding to the position of the parameter σ_{kl} in $R_{(p)}$. Again using the relationship $\widetilde{\boldsymbol{v}} = \mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}'|\boldsymbol{y};\boldsymbol{\Psi}] - \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$,

$$\begin{aligned} S\left(\sigma_{kl}\right) = &- \left(1_{\{k \neq l\}} + \frac{1}{2} \times 1_{\{k=l\}}\right) \sum_{p \in P_{kl}} \left\{ n_p \boldsymbol{R}_{(p)}^{-1} \right. \\ &- \boldsymbol{R}_{(p)}^{-1} \sum_{b} \left[\left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta}\right) \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta}\right)' \right. \\ &- \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta}\right) \left(\boldsymbol{S}_{b(p)}\widetilde{\boldsymbol{\eta}}\right)' - \boldsymbol{S}_{b(p)}\widetilde{\boldsymbol{\eta}} \left(\boldsymbol{y}_{b(p)} - \boldsymbol{X}_{b(p)}\boldsymbol{\beta}\right)' \\ &\left.\left. + \boldsymbol{S}_{b(p)} \left(\widetilde{\boldsymbol{v}} + \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'\right) \boldsymbol{S}_{b(p)}' \right] \boldsymbol{R}_{(p)}^{-1} \right\}_{\{kl\}}. \end{aligned} \tag{3.26}$$

If there were no missing observations then there would only be one attendance pattern and the calculation of the M-step update for $\boldsymbol{R}$ would have a solution that followed the same pattern as the M-step update for $\boldsymbol{G}$. However, the presence of unbalanced student profiles disrupts the structure of $\boldsymbol{R}$, and score functions must be calculated for each of the unique model parameters in $\boldsymbol{R}$. The closed form solution for $S(\sigma_{kl}) = 0$ depends on the number of years and on the attendance patterns that are present in the data set. There is not a simple, general solution and so we use the Newton-Raphson (NR) algorithm to calculate the M-step update. During the first two M-step updates for $\boldsymbol{R}$, we modify the appropriate Hessian by adding a scaled diagonal matrix to improve the stability of the NR update of $\boldsymbol{R}$. This results in a hybrid of a Newton and a gradient descent method that produces more reliable convergence when the initial value is far away from the critical point (Nocedal and Wright, 1999).

3.6 The M-step for the Alternative Model

The M-step update for $\boldsymbol{\beta}$ in the alternative model is the same as the update for the GP model appearing in Equation (3.24), given the appropriate definition of $\boldsymbol{R}$. Likewise, the M-step updates for the $\boldsymbol{\Gamma}_g$ appearing in Equation (3.21) are unchanged.

The new work required for the alternative model is the calculation of the M-step update for the student variance component Γ_{stu} and the yearly error variances σ_g^2 from Equation (3.10), for $g = 1, \ldots, T$. The M-step update for Γ_{stu} is derived in the same way way as the update for $\boldsymbol{\Gamma}_g$, and is equal to the mean of the first n diagonal elements of $\widetilde{\boldsymbol{v}} + \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$. For the purpose of calculating $\widehat{\sigma}_g^2$, let B_g be the set students that are observed in year g.

$$
\begin{aligned}
S(\sigma_g^2) &= \int \frac{\partial}{\partial \sigma_g^2}\left[\log\left(\prod_{j=1}^{T}\prod_{i\in B_g}\sigma_j^{-1}\exp\left[-\frac{\left(y_{ij}-\boldsymbol{x}_{ij}'\boldsymbol{\beta}-\boldsymbol{s}_{ij}'\boldsymbol{\eta}\right)^2}{2\sigma_j^2}\right]\right)\right] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \\
&= \int \frac{\partial}{\partial \sigma_g^2}\left[\sum_{j=1}^{T}\sum_{i\in B_j}\left(-\frac{1}{2}\log\left(\sigma_j^2\right)-\frac{\left(y_{ij}-\boldsymbol{x}_{ij}'\boldsymbol{\beta}-\boldsymbol{s}_{ij}'\boldsymbol{\eta}\right)^2}{2\sigma_j^2}\right)\right] f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \\
&= -\frac{1}{2}\int n_g\left(\frac{\partial}{\partial \sigma_g^2}\log\left(\sigma_g^2\right)\right)+\frac{\partial}{\partial \sigma_g^2}\left(\sum_{i\in B_g}\frac{\left(y_{ig}-\boldsymbol{x}_{ig}'\boldsymbol{\beta}-\boldsymbol{s}_{ig}'\boldsymbol{\eta}\right)^2}{\sigma_g^2}\right) f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta} \\
&= -\frac{1}{2}\int \frac{n_g}{\sigma_g^2}-\sum_{i\in B_g}\frac{\left(y_{ig}-\boldsymbol{x}_{ig}'\boldsymbol{\beta}-\boldsymbol{s}_{ig}'\boldsymbol{\eta}\right)^2}{\left(\sigma_g^2\right)^2} f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta}
\end{aligned}
$$

Using the fact that if $\mathrm{E}[Y] = \mu$ and $\mathrm{cov}(Y) = \Sigma$ then for an $n \times n$ symmetric matrix A, $\mathrm{E}\left[Y'AY\right] = \mathrm{tr}(A\Sigma) + \mu' A \mu$, we obtain

$$\mathrm{E}\left[\boldsymbol{\eta}' \boldsymbol{s}_{ig} \boldsymbol{s}_{ig}' \boldsymbol{\eta} | \boldsymbol{y}; \boldsymbol{\Psi}\right] = \mathrm{tr}\left(\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}' \widetilde{\boldsymbol{v}}\right) + \widetilde{\boldsymbol{\eta}}' \boldsymbol{s}_{ig} \boldsymbol{s}_{ig}' \widetilde{\boldsymbol{\eta}}$$

Setting the score equation equal to zero,

$$\begin{aligned} \widehat{\sigma}_g^2 =& \frac{1}{n_g} \sum_{i \in B_g} \int \left(y_{ig} - \boldsymbol{x}_{ig}'\boldsymbol{\beta} - \boldsymbol{s}_{ig}'\boldsymbol{\eta}\right)' \left(y_{ig} - \boldsymbol{x}_{ig}'\boldsymbol{\beta} - \boldsymbol{s}_{ig}'\boldsymbol{\eta}\right) f(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}) \mathrm{d}\boldsymbol{\eta} \\ =& \frac{1}{n_g} \sum_{i \in B_g} \left[\left(y_{ig}^o - \boldsymbol{x}_{ig}'\boldsymbol{\beta}\right)\left(y_{ig}^o - \boldsymbol{x}_{ig}'\boldsymbol{\beta} - 2\boldsymbol{s}_{ig}'\widetilde{\boldsymbol{\eta}}\right) + \boldsymbol{s}_{ig}'\widetilde{\boldsymbol{v}}\boldsymbol{s}_{ig} + \widetilde{\boldsymbol{\eta}}'\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}'\widetilde{\boldsymbol{\eta}}\right] \end{aligned} \quad (3.27)$$

3.7 The E-step

Calculation of the components of observed data score vector requires the first two moments, $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{v}}$, of $f\left(\boldsymbol{\eta}|\boldsymbol{y};\boldsymbol{\Psi}\right)$. Using the method of Henderson (1950, 1975), the moments are obtained from the gradient and Hessian of $f(\boldsymbol{y}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$. The resulting estimates are

$$\widetilde{\boldsymbol{v}} = \left(\boldsymbol{S}'\boldsymbol{R}^{-1}\boldsymbol{S} + \boldsymbol{G}^{-1}\right)^{-1} \quad (3.28)$$

$$\widetilde{\boldsymbol{\eta}} = \widetilde{\boldsymbol{v}}\boldsymbol{S}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad (3.29)$$

The expression for the EBLUP in Equation (3.29) is equivalent, via a matrix identity (Petersen and Pedersen, 2008, Eq. 147), to the perhaps more familiar expression

$$\widetilde{\boldsymbol{\eta}} = \boldsymbol{G}\boldsymbol{S}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad (3.30)$$

However, from a computational standpoint, (3.29) is much more efficient than (3.30) since it does not require calculation of the full marginal covariance matrix $\boldsymbol{V}$. The calculation of $\widetilde{\boldsymbol{v}}$ requires inversion of a dense matrix of dimension equal to the number of teacher effects, whereas $\boldsymbol{V}$ is dense with dimension equal to the number of observations in the data set. The calculation of (3.29) is relatively fast despite the large dimension of $\boldsymbol{R}$ because both $\boldsymbol{S}'$ and $\boldsymbol{R}^{-1}$ are sparse.

The E-step updates for the alternative model are the same as those appearing in Equations (3.28) and (3.29), using the appropriate definitions of $\boldsymbol{S}, \boldsymbol{G}, \boldsymbol{\eta}$, and $\boldsymbol{R}$.

3.8 EM Standard Errors

One criticism of the EM algorithm is that it does not produce the Hessian of the MLE $\widehat{\boldsymbol{\Psi}}$ as a byproduct. The work we have already done, however, makes it possible for us to compute the observed data information matrix directly without working through a correction to the complete-data information matrix, as done by Louis (1982). Equation (3.17) expresses the observed data score vector $S(\boldsymbol{\Psi})$ as the conditional expectation of the complete data likelihood. We derived the components of the observed data score vector in order to calculate the M-step equations. Together with the values $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{v}}$ from the E-step, our expression for the score vector allows us to calculate the observed information matrix,

$$-\left.\partial S(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}\right|_{\boldsymbol{\Psi}=\widehat{\boldsymbol{\Psi}}} . \tag{3.31}$$

with a central difference approximation at the MLE $\widehat{\boldsymbol{\Psi}}$. This method was suggested by Jamshidian and Jennrich (2000), who proposed using either a forward or central difference approximation, or a Richardson extrapolation (Lindfield and Penny, 1988). Our experience has been that the Richardson extrapolation greatly increases the computation time over the central difference approximation without providing a noticeable increase in precision. However, our code does offer an option to use the Richardson extrapolation instead of the central difference approximation. Regardless, it is important to remember that $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{v}}$ are functions of $\boldsymbol{\Psi}$ and must be re-calculated for each perturbation of the parameter vector for the central difference approximation.

Likelihood theory shows the asymptotic covariance matrix of the model parameters is the inverse of Fisher information matrix

$$cov\left(\hat{\mathbf{\Psi}}\right) = -E\{\partial S(\mathbf{\Psi})/\partial\mathbf{\Psi}\}^{-1},$$

and that that the observed information matrix is a consistent estimator for the Fisher information matrix. The standard errors are obtained by taking the square root of the diagonal elements of $cov\left(\hat{\mathbf{\Psi}}\right)$.

It is also useful to calculate standard errors for the predicted random effects. The matrix $\widetilde{\boldsymbol{v}}$ provides the covariance matrix for $\boldsymbol{\eta}$; however, since $\boldsymbol{\eta}$ is random, $\widetilde{\boldsymbol{v}}$ underestimates the prediction variance of $\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}$ (Littell et al., 2006). As demonstrated by McLean et al. (1991), the prediction variance matrix of the random effects appears in block $\boldsymbol{C}_{22}$ of

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11}\,\boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\,\boldsymbol{C}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{S}'\boldsymbol{R}^{-1}\boldsymbol{S} + \boldsymbol{G}^{-1} \end{pmatrix}^{-1}$$

This procedure also yields the standard errors for $\widehat{\boldsymbol{\beta}}$ in block $\boldsymbol{C}_{11}$. The standard errors obtained by this method for $\widehat{\boldsymbol{\beta}}$ are the same as those obtained by the central difference approximation: the central difference approximation is needed only for the standard errors of the covariance parameters.

3.9 Convergence and Initial Values of the EM Algorithm

Dempster et al. (1977) demonstrate that the observed data likelihood is monotonically increasing in each EM iteration. That is,

$$f\left(\boldsymbol{y}; \mathbf{\Psi}^{(k+1)}\right) \geq f\left(\boldsymbol{y}; \mathbf{\Psi}^{(k)}\right).$$

Wu (1983) presents conditions under which the EM algorithm converges. Beyond regularity conditions which are satisfied (Demidenko, 2004), a sufficient condition to guarantee that our EM algorithm converges to a stationary value of the observed data likelihood is that $Q(\mathbf{\Psi};\mathbf{\Psi}')$ is continuous in both $\mathbf{\Psi}$ and $\mathbf{\Psi}'$. The existence of the derivatives $\partial Q(\mathbf{\Psi};\mathbf{\Psi}')/\partial\mathbf{\Psi}$ that yield the score functions in the M-step guarantees that $Q(\mathbf{\Psi};\mathbf{\Psi}')$ is continuous with respect to $\mathbf{\Psi}$. Furthermore, $Q(\mathbf{\Psi};\mathbf{\Psi}')$ is continuous with respect to $\mathbf{\Psi}'$ since the E-step update of $\widetilde{\boldsymbol{\eta}}$ is a continuous function of $(\boldsymbol{\beta},\boldsymbol{G},\boldsymbol{R})$ in Equations (3.28) and (3.29).

The default convergence criterion stops the algorithm when the relative change in the log-likelihood at iteration k, $l(\mathbf{\Psi}^{(k)})$, is less than a fixed tolerance. Letting $\alpha = 10^{-7}$, we say the algorithm has converged when

$$\frac{l(\mathbf{\Psi}^{(k)}) - l(\mathbf{\Psi}^{(k-1)})}{l(\mathbf{\Psi}^{(k)})} < \alpha.$$

Verification that the EM algorithm has converged to a local maximum of the likelihood function is possible by checking that the Hessian of the observed data likelihood is negative definite.

For several different datasets, we consistently obtained the same estimates from different starting values, including letting the covariance matrices start from nearly singular states. An advantage of the EM algorithm is that no restrictions need to be placed on the $\boldsymbol{G}$ matrix to ensure that it is positive definite (see Section 4.11). This is a valuable characteristic because the future year effects in the GP model may be highly correlated, placing $\boldsymbol{G}$ near the boundary of the parameter space. The default initialization sets $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{G} = 100 * \mathbf{I}$. $\boldsymbol{R}$ is set to a diagonal matrix with each diagonal entry equal to half of the variance of the scores from the corresponding year, and $\boldsymbol{\beta}$ is set to the ordinary least squares estimate of the fixed effects.

3.10 Efficient Implementation of the EM Algorithm

Our EM algorithm is implemented and made available in the R (R Development Core Team, 2012) package GPvam. The program takes advantage of the sparseness of the design and certain covariance matrices, and handles large data sets relatively well. Because the program was custom-designed for the GP model, it requires minimal input. The user must supply a data frame with columns for test scores, year of observation, student ID, and teacher ID. Optionally, other columns may be included for fixed effects, which are declared to the program through an R `formula` statement. Sparse matrices are constructed and handled via the R package Matrix (Bates and Maechler, 2011).

The GP model requires specification of a complex random effects structure. The paper by Doran and Lockwood (2006) provides a tutorial to the implementation of VAMs in R using the functions `lme` and `lmer`. However, Lockwood et al. (2003) explain that, for a less complicated multi-membership model, data sets with more than 200 teachers require several tricks to program with `lme`, and often fail to converge. For our program, the sparse design matrix for the random effects is built automatically from the data, and the model performs well in the application in Section 5.3 which contains 4781 teacher effects.

GPvam implements both the GP model specified in Equation (3.1) and the alternative model in Equation (3.6). Both the log-likelihood and AIC are reported, allowing comparison of the fit of the two models. Note that the models are equivalent when $T = 2$. In GPvam, model (3.6) is faster than model (3.1), although model (3.1) has better scalability properties with respect to memory. For the application in Section 5.3, model (3.1) requires a total of about 2.5 GB of RAM, while model (3.6) requires around 6.0 GB. The alternative model requires more memory because the student effects are modeled in the G matrix, increasing the dimension of the dense matrix $\widetilde{v}$ by the number of students.

It is possible to fit the GP model in SAS software (SAS Institute Inc., 2011) for small data sets, but as noted by Broatch and Lohr (2011) implementations of

VAMs in SAS do not have good scalability properties. The GP model requires the EFFECT statement of PROC GLIMMIX, as well as specifying a user-defined covariance structure for the random effects. Dimension-reduction techniques exist for the Newton-Raphson (NR) methods used by SAS to avoid calculation of the marginal covariance matrix $\boldsymbol{V}$ (Wolfinger et al., 1994), but NR requires that special steps be taken to ensure that the covariance matrix of the random effects remains positive definite (Demidenko, 2004).

Table 3.3 gives the results of the comparison of the run-times for SAS and GPvam for a data set 6236 observations on 2834 students over 3 years, with 102, 104, and 98 teachers in each year, respectively. We use both the standard specification of the GP model given in Equation 3.1, denoted by the suffix "-gp", as well as the alternate model specification of the GP VAM, denoted by the suffix "-alt". SAS-gp failed after encountering a negative-definite covariance matrix after 125 minutes, while SAS-alt ran out of memory after a few minutes. Note that the application in Section 5.3 involves a much larger data set than the one used in this example. The usual advantages of NR over EM (Lindstrom and Bates, 1988) are negated because of the size of the matrices produced by the multi-membership structure and the high correlations between future year effects that are typical in the GP model.

The efficiency of our estimation of the GP model stems from utilization of methods for manipulating sparse matrices, the stability of the EM algorithm when $\boldsymbol{G}$ is nearly singular, and the reduced dimension of the covariance matrix that needs to be inverted (see Section 3.7). In some cases, R offers different options for the same procedure. For each step of the program, the chosen methods were tested for speed against alternatives. For some of the M-step calculations for the $\boldsymbol{R}$ matrix in GPvam-gp, it is faster to work with a dense version of $\boldsymbol{S}$ since subsets of the matrix are required. In this case, we take advantage of the fact that $\boldsymbol{S}$ is a binary matrix,

Table 3.3: Run times in minutes

GPvam-gp	SAS-gp	GPvam-alt	SAS-alt
114	>125 (Failed)	12	Failed

meaning that multiplication of a vector by S is equivalent to taking sums of subsets of the vector. It may be possible to improve the speed of GPvam-gp, bringing it closer to the performance of GPvam-alt by executing some of the required loops and matrix operations in C via R. However, the current performance of the program has exceeded our needs and we have not pursued this option.

3.11 Discussion

The GP model provides a flexible framework for modeling education data without making the same assumptions about vertical test design and scaling as made by previous VAMs. We have developed a method for computing maximum-likelihood estimates for the generalized persistence model (Mariano et al., 2010). The EM algorithm offers an efficient method of computation, taking advantage of matrix sparsity and requiring inversion of a matrix whose dimension depends on the number of teachers, which is typically much smaller than the total number of observations as would be used in routine implementation. The algorithm produces stable behavior in the presence of a nearly-singular covariance matrix for the random effects that results from highly correlated teacher effects across years. We have implemented the proposed methods in the R package GPvam. The availability of maximum-likelihood estimates should be useful for those preferring Bayesian estimation as well, providing a sensitivity analysis to their choice of priors. We hope that this user-friendly implementation of the model will facilitate further empirical study of the model's properties.

In the next chapter, we combine the GP VAM with a model for the missing data mechanism. The joint model provides the ability to test for the sensitivity of teacher rankings from the GP model to the presence of nonignorable missing data under various structures for the missing data mechanism. In Chapter 5, the joint model is used to perform a sensitivity analysis on the rankings produced by the GP model when applied to a data set of elementary school standardized test scores and a data set containing college calculus grades.

Chapter 4

A CORRELATED RANDOM EFFECTS MODEL

The generalized persistence (GP) model presented in Chapter 3 assumes that missing data may be ignored. However, if the observations are missing systematically, due to any of the effects being measured by the model, then the model will yield biased results. This chapter develops a model to incorporate information about the missingness process. Notation is introduced, the GP model is presented in the framework of this notation, a model for the missing data indicators is proposed, and the two models are combined into a joint model. The joint model is applied to real data sets in Chapter 5, where estimates from the new model are compared to those from the unmodified GP model.

Let y_{ig} be the potential response of student i at time g, with $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{iT})'$ and $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$. The indicator variable

$$r_{ig} = \begin{cases} 1 \text{ if } y_{ig} \text{ is observed} \\ 0 \text{ otherwise} \end{cases}$$

tracks whether the planned measurement on student i at time g is observed or missing. Let $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{iT})'$ and $\boldsymbol{r} = (\boldsymbol{r}_1', \ldots, \boldsymbol{r}_n')'$. The complete data $\boldsymbol{y} = \{\boldsymbol{y}^o, \boldsymbol{y}^m\}$ consists of both the observed data $\boldsymbol{y}^o$ and the missing data $\boldsymbol{y}^m$. The vector $\boldsymbol{y}^o$ consists of the values y_{ig} such that $r_{ig} = 1$, and $\boldsymbol{y}^m$ consists of the values y_{ig} that would have been observed if the observations were not missing. Note that by missing data we are referring to missing observations on the response variable, not missing covariates. We assume that we do not have any missing covariates. For our model this assumption amounts to knowing which teachers taught each student at each time, as well as the values of the fixed effects covariates for each observation.

4.1 The Observed Data Model

We wish to model student scores $\boldsymbol{y}$ using information about the history of observations on each student and each student's teacher-history. To be precise, suppose a data set tracks a cohort of n students over T years. The GP model assumes a linear mixed model as follows:

$$y_{ig}^{o} = \boldsymbol{x}_{ig}'\boldsymbol{\beta}_{obs} + \boldsymbol{s}_{ig}'\boldsymbol{\eta}_{obs} + \epsilon_{ig} \tag{4.1}$$

where y_{ig}^{o} denotes the score for student i during year g, for $i = 1, \ldots, n$, and $g \in A_i$; A_i is the set of years in which student i is observed. Students are taught by one of m_g teachers in each year g. We will also refer to the vector of concatenated student scores, $\boldsymbol{y^o} = (\boldsymbol{y}_1^{o\prime}, \ldots, \boldsymbol{y}_n^{o\prime})'$, where $\boldsymbol{y}_i^{o} = (y_{ig}^{o})$. The matrix $\boldsymbol{X}$, with rows $\boldsymbol{x}_{ig}'$, is the design matrix for the vector $\boldsymbol{\beta}$ of student and teacher level covariates such as demographic information or years of teaching experience. The matrix $\boldsymbol{S}$, with rows $\boldsymbol{s}_{ig}'$, is the design matrix for the random student and teacher effects.

The random effects vector $\boldsymbol{\eta}_{obs} = [\boldsymbol{\delta}_{obs}' \ \boldsymbol{\theta}_{obs}']'$ has two components. Student i has a latent effect δ_i that represents an underlying level of achievement not explained by the fixed covariates, and $\boldsymbol{\delta}_{obs}' = (\delta_1, \ldots, \delta_n)'$. As in the "alternate model" of Chapter 3, we assume that $\delta_1, \ldots, \delta_n$ are i.i.d. $N(0, \Gamma_{stu})$. This represents a slight departure from Mariano et al. (2010), who model the intra-student correlation in an unstructured error covariance matrix. However, that structure is not as amenable to the joint model for missingness because it precludes the possibility of including student effects in the missing data mechanism. As a result, we model the intra-student correlation with random effects, similar to the VAM used by McCaffrey and Lockwood (2011). When the annual responses y_{ig} have the same scale, this leads to a compound-symmetry covariance structure for the students.

The generalized persistence model estimates the effect of teachers on students in the year that they teach them, their lasting effect on the next year's score, and so on. Following the notation of Mariano et al. (2010), we let $\theta_{g[jt]}$ represent the effect for the j-th grade-g teacher on a student's grade t score. A grade $g = 1, \ldots, T$ teacher has $K_g = T - g + 1$ effects. Thus $\boldsymbol{\theta}_{g[j\cdot]}$ gives the vector of current and future year effects of the j-th grade g teacher. The vector $\boldsymbol{\theta}_{obs}$ concatenates the $\boldsymbol{\theta}_{g[j\cdot]}$ effects for all grades and teachers. The model is able to distinguish between the persistence effect of former teachers and the current effect of the present teacher because the students are not nested at the teacher level.

We structure $\boldsymbol{\eta}_{obs}$ in a way that leads to a block-diagonal random effects covariance matrix. The first n elements of $\boldsymbol{\eta}_{obs}$ correspond to the student random effects, and the last $\sum_{g=1}^{T} K_g m_g$ elements correspond to the classroom random effects. This yields

$$\boldsymbol{\eta}_{obs} = (\delta_1, \ldots, \delta_n, \boldsymbol{\theta}'_{1[1\cdot]}, \ldots, \boldsymbol{\theta}'_{1[m_1\cdot]}, \boldsymbol{\theta}'_{2[1\cdot]}, \ldots, \boldsymbol{\theta}'_{2[m_2\cdot]}, \ldots, \theta_{T[1\cdot]}, \ldots, \theta_{T[m_T\cdot]})'. \quad (4.2)$$

The vector $\boldsymbol{\eta}_{obs}$ is distributed as $\boldsymbol{\eta}_{obs} \sim N(\mathbf{0}, \boldsymbol{G})$ where

$$\boldsymbol{G} = \text{blockdiag}\left(\Gamma_{stu}\mathbf{I}_n, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_T, \ldots, \boldsymbol{\Gamma}_T\right),$$

with m_g copies each of $\boldsymbol{\Gamma}_g$, where the $\boldsymbol{\Gamma}_g$ are unstructured. The design matrix $\boldsymbol{S}$ of the random effects has rows s'_{ig}, and may be partitioned into two blocks $\boldsymbol{S} = [\boldsymbol{S}_1 \; \boldsymbol{S}_2]$. $\boldsymbol{S}_1$ contains a 1 in column i if the observation is for student i, and $\boldsymbol{S}_2$ contains 1's in entries corresponding to teachers who could affect that response.

The error terms are distributed as $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{R})$ where $\boldsymbol{R}$ is a diagonal matrix with entries coming from the set $\{\sigma_1^2, \ldots, \sigma_T^2\}$, depending on the year of the observation. In addition, we assume $\text{cov}(\boldsymbol{\eta}_{obs}, \boldsymbol{\epsilon}) = \mathbf{0}$. The log-likelihood of the GP model is thus

$$l(\boldsymbol{\Psi}) = \log \int \left\{ \prod_{i=1}^{n} \prod_{g \in A_i} f(y_{ig}^{o} | \boldsymbol{\eta}_{obs}) \right\} f(\boldsymbol{\eta}_{obs}) \mathrm{d}\boldsymbol{\eta}_{obs} \tag{4.3}$$

$$\propto -\frac{1}{2} \log |\boldsymbol{V}| - \frac{1}{2} \left(\boldsymbol{y}^{o} - \boldsymbol{X\beta} \right)' \boldsymbol{V}^{-1} \left(\boldsymbol{y}^{o} - \boldsymbol{X\beta} \right) \tag{4.4}$$

where

$$\boldsymbol{V} = \boldsymbol{SGS'} + \boldsymbol{R}$$

$$f(y_{ig}^{o} | \boldsymbol{\eta}_{obs}) \propto \left(\sigma_g^2 \right)^{-1/2} \exp \left[- \left(y_{ig}^{o} - \boldsymbol{x}_{ig}' \boldsymbol{\beta} - \boldsymbol{s}_{ig}' \boldsymbol{\eta}_{obs} \right)^2 / (2\sigma_g^2) \right]$$

$$f(\boldsymbol{\eta}) \propto \det \left(\boldsymbol{G} \right)^{-1/2} \exp \left[-(\boldsymbol{\eta}_{obs}' \boldsymbol{G}^{-1} \boldsymbol{\eta}_{obs})/2 \right]$$

and $\boldsymbol{\Psi}$ is a vector of the model parameters. The integral in Equation (4.3) has a closed form solution, but this will not be the case for the joint model. Our joint model will approximate the intractable integral with a fully exponential Laplace approximation. The Laplace approximation is exact in Equation (4.3) since the data are assumed to be normally distributed and an identity link is used, meaning the random effects enter the model linearly (Pinheiro and Bates, 1995).

4.2 The Missing Data Model

To model p_{ig}, the conditional probability that $r_{ig} = 1$, we use a threshold model originally developed by McCulloch (1994) and discussed in McCulloch et al. (2008) by defining

$$p_{ij} = P\{ \boldsymbol{w}_{ig}' \boldsymbol{\beta}_{mis} + \boldsymbol{z_{ig}'} \boldsymbol{\eta}_{mis} + \epsilon_{mis} > 0 \}$$

where $\boldsymbol{w}_{ig}'$ and $\boldsymbol{z}_{ig}'$ are the rows of the design matrices $\boldsymbol{W}$ and $\boldsymbol{Z}$ of the fixed and random effects in the missing data model, and ϵ_{mis} is distributed as $N(0, 1)$. This results in a generalized linear mixed model with a probit link:

$$r_{ig} | \boldsymbol{\eta}_{mis} \sim \mathrm{Bin}(1, p_{ig})$$

$$\Phi^{-1}(p_{ig}) = \boldsymbol{w}_{ig}' \boldsymbol{\beta}_{mis} + \boldsymbol{z}_{ig}' \boldsymbol{\eta}_{mis}$$

The vectors $\boldsymbol{w}'_{ig}$ and $\boldsymbol{z}'_{ig}$ describe which fixed and random effects are thought to be related to the response mechanism. The vector of fixed effects $\boldsymbol{\beta}_{mis}$ of the missing model will be different from the $\boldsymbol{\beta}_{obs}$ of the observed model. It will represent a baseline propensity for missingness at each level of the fixed effects. The missing data model should not include stochastic time-varying covariates in $\boldsymbol{\beta}_{mis}$, since these would be missing along with the test score. Furthermore, the missing data model requires that there is at least one missing observation at each level of each categorical fixed effect in the missing data mechanism. Otherwise, the data suffer from quasi-complete separation (Allison, 2008; Agresti, 2002). In this case, the maximum likelihood estimate for the particular fixed effect does not exist.

We may include either random teacher effects, random student effects, or both in $\boldsymbol{\eta}_{mis}$. The structure of the random effects is flexible, and may be modified depending on the goals of the study. This flexibility provides the means for performing a sensitivity analysis. When jointly modeling MNAR data, the missing data mechanism makes untestable assumptions about the nature of the relationship between the observed and missing data processes. Molenberghs et al. (2008) show that it is not possible to perform an overall test of MNAR versus MAR since every MNAR model has an MAR counterpart that provides the same fit to the observed data but different predictions for the unobserved data. The plausibility of the assumed model cannot be tested empirically, and as a result it is necessary to fit several alternatives of the missing data mechanism to check the sensitivity of the inference to the choice of joint modeling structure (Xu and Blozis, 2011).

The student effects in the missingness model, if included, will be denoted δ_i^{mis}. The teacher effects in the missing data model will be denoted by $\Lambda_{g[j]}$. These effects may be structured in a number of different ways. In our applications in Chapter 5, $\Lambda_{g[j]}$ represents the effect that the j-th grade g teacher has on the probability of his or her students being measured in year $g+1$. This effect measures how likely it is that students are observed in the year after studying under a particular teacher. This effect is not calculated for teachers in the last year of observations (year T) because no information is available on the future dropout patterns of students of those teachers. This feature of the model would detect instructors whose students drop out (of the school or sequence of courses) at a relatively high rate. We refer to these effects as the "completion effects" of the grade g teachers, since they measure the rate with which students complete year $g+1$. The conditional density of r_{ig} given the random effects is

$$\begin{aligned} f(r_{ig}|\boldsymbol{\eta}_{mis}) &= \Phi\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right)^{r_{ig}}\left[1-\Phi\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right)\right]^{1-r_{ig}} \\ &= \Phi\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right)^{r_{ig}}\Phi\left(-\left[\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right]\right)^{1-r_{ig}} \\ &= \Phi\left((-1)^{1-r_{ig}}\left[\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right]\right) \end{aligned}$$

As with the y_{ig}, we assume the r_{ig} are conditionally independent given the random effects, yielding

$$\begin{aligned} f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) &= \prod_{i=1}^{n}\prod_{g=1}^{T} f(r_{ig}|\boldsymbol{\eta}_{mis}) \\ &= \prod_{i=1}^{n}\prod_{g=1}^{T} \Phi\left((-1)^{1-r_{ig}}\left[\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right]\right) \qquad (4.5) \end{aligned}$$

4.3 The Joint Model

Instead of assuming that the full data depend on the pattern of missingness or on the number of missing observations as in McCaffrey and Lockwood (2011), we will use a correlated-parameter model (CPM) (Lin et al., 2009), a generalization of a shared-parameter model (Wu and Carroll, 1988). The CPM proposed in this chapter allows the missing data mechanism to depend on the effect of students and teachers. This should give more flexibility in detecting sensitivity to missing data than models that only consider student effects, since it is plausible that the attendance trajectory of students depends on their current and former teachers.

The GP model assumes that missing data are MAR. Inference is intended to be on $\boldsymbol{y} = (\boldsymbol{y}^o, \boldsymbol{y}^m)$, but only the $\boldsymbol{y}^o$ have been observed. With MNAR data, $f(\boldsymbol{y}^o)$ is not the correct likelihood to maximize because $\boldsymbol{r}$ provides information about the distribution of $\boldsymbol{y}$. To obtain unbiased parameter estimates for the longitudinal process $\boldsymbol{y}$, the longitudinal and missingness processes must be modeled jointly and $f(\boldsymbol{y}^o, \boldsymbol{r})$ must be maximized. The joint model specifies $f(\boldsymbol{y}, \boldsymbol{r})$, meaning the missing data $\boldsymbol{y}^m$ must be integrated out of the joint density $f(\boldsymbol{y}, \boldsymbol{r})$ to yield the appropriate likelihood function.

$$f(\boldsymbol{y^o}, \boldsymbol{r}) = \iiint f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis}) \mathrm{d}\boldsymbol{y}^m \mathrm{d}\boldsymbol{\eta}_{obs} \mathrm{d}\boldsymbol{\eta}_{mis}$$

In addition, we must factor the joint likelihood $f(\boldsymbol{y}^o, \boldsymbol{r})$. In the CPM, we assume that the missing data mechanism, $f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})$ and the longitudinal process, $f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})$ are conditionally independent, given a set of correlated random effects, $(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis})$. Since neither the longitudinal process nor the missing data mechanism condition on the the $\boldsymbol{y^m}$, the integral of $f(\boldsymbol{y^o}, \boldsymbol{y^m}|\boldsymbol{\eta})$ over the $\boldsymbol{y^m}$ produces the marginal density $f(\boldsymbol{y^o}|\boldsymbol{\eta})$ (this would not be the case in the framework of selection models). This results in the observed data likelihood

$$f(\boldsymbol{y^o}, \boldsymbol{r}) = \iint f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs}) f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis}) \mathrm{d}\boldsymbol{\eta}_{obs} \mathrm{d}\boldsymbol{\eta}_{mis} \tag{4.6}$$

where $f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis})$ is the density of a multivariate normal distribution. The vector $\boldsymbol{\eta}_{obs}$ contains the students' general levels of achievement as well as the teacher effects on test scores, while the effects in $\boldsymbol{\eta}_{mis}$ measure some combination of students' attendance probabilities and/or the relative frequency of teachers' former students completing the next year.

CPMs make different assumptions on the joint model than selection and pattern-mixture models (e.g. conditional independence) and present a different approach for missing data modeling. A major benefit of using CPMs is that they allow us to use the teacher history in the modeling of the dropout mechanism. The EBLUPs of the classroom effects in the missingness model (part of the $\boldsymbol{\eta}_{mis}$ vector) provide a direct method of evaluating the frequency with which teachers' former students drop out. Since the missing data model estimates the probability that a given observation would be observed, a larger EBLUP for a classroom effect in the missingness model indicates that students who took that particular class are more likely to complete the next year than students who took another class that year (i.e. with another teacher). It would, however, be unrealistic to expect the effect of a teacher on student learning to be identical to the effect of the teacher on the future student attendance.

The CPM is constructed with the observed and missing data mechanisms in Equation (4.6). We concatenate the random effects vectors $\boldsymbol{\eta}_{obs}$ and $\boldsymbol{\eta}_{mis}$ into a single random effects vector, $\boldsymbol{\eta}$. To ensure that the $\mathrm{cov}(\boldsymbol{\eta}) = \boldsymbol{G}$ matrix is block-diagonal, we structure the $\boldsymbol{\eta}$ vector as

$$\boldsymbol{\eta} = \left(\delta_1, \delta_1^{\text{mis}}, \ldots, \delta_n, \delta_n^{\text{mis}}, \boldsymbol{\theta_{1[1\cdot]}}, \Lambda_{1[1]}, \ldots, \boldsymbol{\theta_{1[m_1\cdot]}}, \Lambda_{1[m_1]}, \boldsymbol{\theta_{2[1\cdot]}}, \Lambda_{2[1]}, \ldots, \right. \\ \left. \boldsymbol{\theta_{2[m_2\cdot]}}, \Lambda_{2[m_2]}, \ldots, \boldsymbol{\theta_{T[m_T\cdot]}}\right)' \quad (4.7)$$

We model the random student effects and their counterparts for the missing data mechanism, if they are included, as $(\delta_i, \delta_i^{mis})' \sim N_2(\mathbf{0}, \boldsymbol{\Gamma}_{\text{stu}})$ where $\boldsymbol{\Gamma}_{\text{stu}}$ is a 2×2 unstructured covariance matrix. If the random student effects are not included in the missing data model, simply omit the δ_i^{mis} from $\boldsymbol{\eta}$ and model $\delta_i \sim N_1(0, \Gamma_{\text{stu}})$. The teacher effects are independent of the student effects and distributed as

$$\begin{cases} \left(\boldsymbol{\theta}'_{g[j\cdot]}, \Lambda'_{g[j]}\right)' \sim N_{K_g+1}(\mathbf{0}, \Gamma_g) \text{ if } g \neq T \\ \left(\boldsymbol{\theta}'_{g[j\cdot]}\right)' \sim N_{K_g}(\mathbf{0}, \Gamma_g) \qquad\quad \text{if } g = T \end{cases}$$

where Γ_g is an unstructured $(K_g+1)\times(K_g+1)$ covariance matrix if $g \neq T$, or $K_g \times K_g$ if $g = T$. Then

$$\boldsymbol{G} = \operatorname{cov}(\boldsymbol{\eta}) = \operatorname{blockdiag}\left(\boldsymbol{\Gamma}_{\text{stu}}, \ldots, \boldsymbol{\Gamma}_{\text{stu}}, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_T, \ldots, \boldsymbol{\Gamma}_T\right) \tag{4.8}$$

where there are n copies of $\boldsymbol{\Gamma}_{\text{stu}}$, and for each $g = 1, \ldots, T$ there are m_g copies of $\boldsymbol{\Gamma}_g$ in $\boldsymbol{G}$. In general, $\boldsymbol{G}$ will have

$$3 + \sum_{g=1}^{T-1} \frac{(K_g+1)(K_g+2)}{2} + K_T = 1 + \frac{1}{6}\left(T^3 + 6T^2 + 11T\right)$$

free parameters, and $\boldsymbol{G}$ is a square matrix of dimension $2n + \sum_{g=1}^{T-1} m_g(K_g+1) + m_T K_T$. The $\boldsymbol{R}$ matrix for $f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})$ is diagonal, with a separate error variance estimated for each year. We will use σ_g^2 to denote the variance of the error terms for year g in the observed data mechanism. Thus $\boldsymbol{R}$ is diagonal with diagonal entries coming from the set $\{\sigma_1^2, \ldots, \sigma_T^2\}$, with their order of appearance in the matrix depending on the ordering of the data. $\boldsymbol{R}$ introduces T parameters into the model.

Adding up the parameters introduced by the covariance matrices $\boldsymbol{G}$ an $\boldsymbol{R}$ and assuming for the moment that the only fixed effects included in the sub-models are the yearly means, we find that the model has a total of $1 + \frac{1}{6}\left(T^3 + 6T^2 + 11T\right) + 3T$ parameters that must be estimated. See Table 4.1. Notice that this is independent of both the number of students and the number of teachers: it depends only on the number of years measured. For applications to datasets containing many years of measurements, it may be reasonable to assume that the future year effects decay to zero after several years. This would reduce the number of parameters that need to be estimated; however, we do not consider such a modification here.

Table 4.1: Number of parameters to be estimated

Number of Years	Free Parameters
2	16
3	29
4	47
5	71

The log-likelihood for the joint model is

$$l(\boldsymbol{\Psi}) = \log \iint \prod_{i=1}^{n} \left\{ \prod_{g \in A_i} f(y_{ig}|\boldsymbol{\eta}_{obs}) \prod_{g=1}^{T} f(r_{ig}|\boldsymbol{\eta}_{mis}) \right\} f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis}) \mathrm{d}\boldsymbol{\eta}_{obs} \mathrm{d}\boldsymbol{\eta}_{mis} \tag{4.9}$$

where

$$\begin{aligned} f(y_{ig}|\boldsymbol{\eta}_{obs}) &\propto \left(\sigma_g^2\right)^{-1/2} \exp\left[- \left(y_{ig}^o - \boldsymbol{x}_{ig}' \boldsymbol{\beta}_{obs} - \boldsymbol{s}_{ig}' \boldsymbol{\eta}_{obs} \right)^2 / (2\sigma_g^2) \right], \\ f(r_{ig}|\boldsymbol{\eta}_{mis}) &= \Phi\left[(-1)^{1-r_{ig}} \left(\boldsymbol{w}_{ig}' \boldsymbol{\beta}_{mis} + \boldsymbol{z}_{ig}' \boldsymbol{\eta}_{mis} \right) \right], \\ f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis}) &= f(\boldsymbol{\eta}) \propto \det\left(\boldsymbol{G}\right)^{-1/2} \exp\left[-(\boldsymbol{\eta}' \boldsymbol{G}^{-1} \boldsymbol{\eta})/2 \right], \end{aligned}$$

and A_i is the set of years in which student i has an observation.

4.4 Estimation Procedure

The joint model presents a high-dimensional integration problem when calculating the marginal distribution of the observed data in Equation (4.9). The source of the problem is twofold, due to the presence of a nonlinear link in the integrand for the modeling of the binary missingness process and the multi-membership structure of VAMs. The random effects' correlation structure is not nested, which means that the integral over the random effects cannot be factored into a product of low-dimensional integrals (e.g. one- or two-dimensional integrals).

The Expectation-Maximization (EM) algorithm may be used by treating the random effects as missing data (Dempster et al., 1977). The E-step calculates the conditional expectation of the complete-data likelihood, given the observed data and current parameter estimates, and the M-step maximizes the conditional expectation of the complete data likelihood. The EM algorithm is used to develop an efficient routine for estimating the GP model (Mariano et al., 2010) under an assumption of MAR in Chapter 3, and is available via the package GPvam. We extend that work to estimate the parameters of the CPM. We implement a fully exponential Laplace approximation to approximate the intractable integral in the E-step of the EM algorithm (Steele, 1996). Rizopoulos et al. (2009) and Rizopoulos (2010) propose an EM algorithm with Laplace approximations to estimate the joint model for longitudinal outcomes and survival with hierarchical data. We modify their work to allow 1) a multi-membership dependence structure and 2) a missingness process that depends on random teacher and/or student effects. As will be demonstrated, the added computational challenges are tremendous in order to make the estimation possible for practical use with large data sets, and the extensions in theory and computation are not straightforward.

4.5 The M-step

The M-step of the EM algorithm maximizes the conditional expectation of the complete data likelihood. Often, expressions in the M-step have a closed form solution, providing part of the motivation for using the EM algorithm. However, the fixed effects for the missing data mechanism enter the model nonlinearly and their M-step update requires numerical optimization.

Let $\widetilde{\boldsymbol{\eta}} = \mathrm{E}[\boldsymbol{\eta}|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi}]$ and $\widetilde{\boldsymbol{v}} = \mathrm{var}[\boldsymbol{\eta}|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi}]$ represent the conditional expectation and variance, respectively, of $\boldsymbol{\eta}$. These quantities are calculated in the E-step and remain fixed during the M-step. Likewise, let the sub-vector of $\widetilde{\boldsymbol{\eta}}$ corresponding to $\mathrm{E}[\boldsymbol{\theta}_{g[j\cdot]}|\boldsymbol{y}^o; \boldsymbol{\Psi}]$ be denoted $\widetilde{\boldsymbol{\theta}}_{g[j\cdot]}$, and the block of the matrix $\widetilde{\boldsymbol{v}}$ corresponding to $\mathrm{E}[\boldsymbol{\theta}_{g[j\cdot]}\boldsymbol{\theta}'_{g[j\cdot]}|\boldsymbol{y}^o; \boldsymbol{\Psi}]$ be denoted $\widetilde{\boldsymbol{v}}_{g[j\cdot]}$. The M-step for the parameters of $\boldsymbol{\beta}_{obs}, \boldsymbol{G}, \boldsymbol{R}$ are unchanged from the equations derived in the previous chapter.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{obs} &= \left(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\left(\boldsymbol{y}^{o} - \boldsymbol{S}\widetilde{\boldsymbol{\eta}}_{obs}\right) \\
\widehat{\boldsymbol{\Gamma}}_{g} &= \frac{1}{m_g}\sum_{j=1}^{m_g}\left(\widetilde{\boldsymbol{v}}_{g[j\cdot]} + \widetilde{\boldsymbol{\theta}}_{g[j\cdot]}\widetilde{\boldsymbol{\theta}}'_{g[j\cdot]}\right) \qquad (4.10) \\
\widehat{\sigma}^{2}_{g} &= \frac{1}{n_g}\sum_{i\in B_g}\left[\left(y^{o}_{ig} - \boldsymbol{x}'_{ig}\boldsymbol{\beta}_{obs}\right)\left(y^{o}_{ig} - \boldsymbol{x}'_{ig}\boldsymbol{\beta}_{obs} - 2\boldsymbol{s}'_{ig}\widetilde{\boldsymbol{\eta}}_{obs}\right)\right. \\
&\qquad \left.\boldsymbol{s}'_{ig}\widetilde{\boldsymbol{v}}\boldsymbol{s}_{ig} + \widetilde{\boldsymbol{\eta}}'_{obs}\boldsymbol{s}_{ig}\boldsymbol{s}'_{ig}\widetilde{\boldsymbol{\eta}}_{obs}\right] \qquad (4.11)
\end{aligned}$$

The M-step update for $\boldsymbol{\beta}_{mis}$ does not have a closed form solution. The observed data score function for the fixed effects of the missing data mechanism is

$$\begin{aligned}
&S(\boldsymbol{\beta}_{mis}) \\
=&\int \frac{\partial}{\partial\boldsymbol{\beta}_{mis}}\left[\sum_{i=1}^{n}\sum_{g=1}^{T}\log\left(\Phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]\right)\right] f(\boldsymbol{\eta}|\boldsymbol{y}^{o},\boldsymbol{r})\mathrm{d}\boldsymbol{\eta} \\
=&\sum_{i=1}^{n}\sum_{g=1}^{T}\boldsymbol{w}_{ig}\int(-1)^{1-r_{ig}}\frac{\phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]}{\Phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]} f(\boldsymbol{\eta}|\boldsymbol{y}^{o},\boldsymbol{r})\mathrm{d}\boldsymbol{\eta} \qquad (4.12)
\end{aligned}$$

where $\phi(\cdot)$ is the density of a standard normal random variable. We solve Equation (4.12) via Newton-Raphson. The Hessian $\mathcal{H}(\widehat{\boldsymbol{\beta}}_{mis})$ is calculated by applying a central difference approximation to $S(\boldsymbol{\beta}_{mis})$ at $\widehat{\boldsymbol{\beta}}_{mis}$. Iteration $p+1$ of the approximation yields

$$\widehat{\boldsymbol{\beta}}^{p+1}_{mis} = \widehat{\boldsymbol{\beta}}^{p}_{mis} - \mathcal{H}(\widehat{\boldsymbol{\beta}}^{p}_{mis})^{-1}S(\widehat{\boldsymbol{\beta}}^{p}_{mis}).$$

Iterations continue until

$$S(\widehat{\boldsymbol{\beta}}^{p}_{mis})\,'\mathcal{H}(\widehat{\boldsymbol{\beta}}^{p}_{mis})^{-1}S(\widehat{\boldsymbol{\beta}}^{p}_{mis}) < \alpha. \qquad (4.13)$$

We recommend $\alpha = 10^{-8}$.

4.6 The E-step with a Fully Exponential Laplace Approximation

Calculation of the components of observed data score vector requires the first two moments, $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{v}}$, of $f\left(\boldsymbol{\eta}|\boldsymbol{y}^{o},\boldsymbol{r};\boldsymbol{\Psi}\right)$, as well as the conditional expectations appearing in Equation (4.12). Letting $\{E\left[H(\boldsymbol{\eta})\right]\}_k$ denote the (k)-th component of the vector (or scalar) $E\left[H(\boldsymbol{\eta})\right]$, the M-step updates require

$$\{\mathrm{E}\left[H(\boldsymbol{\eta})|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi}\right]\}_k = \int \{H(\boldsymbol{\eta})\}_k \, f(\boldsymbol{\eta}|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi})\mathrm{d}\boldsymbol{\eta}$$

where $H(\cdot)$ is a function of the random effects, and $\boldsymbol{\Psi}$ is fixed at its value from the previous iteration. For the M-step updates $\boldsymbol{\beta}_{obs}, \boldsymbol{G}$ and $\boldsymbol{R}$, we need $H(\boldsymbol{\eta}) = \boldsymbol{\eta}$ and $\widetilde{\boldsymbol{v}} = \mathrm{var}[\boldsymbol{\eta}|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi}]$. For the M-step update of the fixed effects, $\boldsymbol{\beta}_{mis}$, of the missing data mechanism, we need

$$H(\boldsymbol{\eta}) = (-1)^{1-r_{ig}} \frac{\phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]}{\Phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]}.$$

To solve these high-dimensional integration problems, we follow the examples of Steele (1996) and Rizopoulos et al. (2009) and use the fully exponential Laplace approximation of Tierney et al. (1989), approximating the cumulant-generating function $\log\left\{E\left[\exp\left(\boldsymbol{c}'H(\boldsymbol{\eta})\right)\right]\right\}$ at the mode $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}}^{(\mathbf{0})}$, where

$$\widehat{\boldsymbol{\eta}}^{(\boldsymbol{c})} = \mathrm{argmax}_{\boldsymbol{\eta}} \left\{\log\left[f\left(\boldsymbol{y}^o, \boldsymbol{r}, \boldsymbol{\eta}\right) + \boldsymbol{c}'H\left(\boldsymbol{\eta}\right)\right]\right\}.$$

The mode $\widehat{\boldsymbol{\eta}}$ is obtained by Newton-Raphson, using the same convergence criterion as in Equation (4.13):

$$\widehat{\boldsymbol{\eta}}^{\mathrm{p}+1} = \widehat{\boldsymbol{\eta}}^{\mathrm{p}} - \left(\boldsymbol{\Sigma}^{\mathrm{p}}\right)^{-1}\boldsymbol{L}\left(\widehat{\boldsymbol{\eta}}^{\mathrm{p}}\right)$$

where “p” is the iteration counter. Using properties of matrix differentiation (Magnus and Neudecker, 1999; Harville, 2008),

$$\begin{aligned}
\boldsymbol{L}(\boldsymbol{\eta}) =& -\frac{\partial}{\partial\boldsymbol{\eta}}\left[\log\left\{f\left(\boldsymbol{y}^o|\boldsymbol{\eta}\right)\right\} + \log\left\{f\left(\boldsymbol{r}|\boldsymbol{\eta}\right)\right\} + \log\left\{f\left(\boldsymbol{\eta}\right)\right\} + \boldsymbol{c}'H(\boldsymbol{\eta})\right]|_{\boldsymbol{c}=\mathbf{0}} \\
=& -\sum_{g=1}^{T}\sum_{i\in A_j}\left(\frac{y^o_{ig} - \boldsymbol{x}'_{ig}\boldsymbol{\beta}_{obs} - \boldsymbol{s}'_{ig}\boldsymbol{\eta}}{\sigma^2_j}\right)\boldsymbol{s}_{ig} \\
& -\sum_{i=1}^{n}\sum_{g=1}^{T}\left((-1)^{1-r_{ig}}\frac{\phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]}{\Phi\left[(-1)^{1-r_{ig}}\left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}\right)\right]}\right)\boldsymbol{z}_{ig} \\
& +\boldsymbol{G}^{-1}\boldsymbol{\eta} + \qquad (4.14)
\end{aligned}$$

and $\boldsymbol{\Sigma}^w = \boldsymbol{\Sigma}^{(\boldsymbol{c})}|_{(\boldsymbol{c},\boldsymbol{\eta})=(\boldsymbol{0},\widehat{\boldsymbol{\eta}}^{\mathrm{w}})}$, with

$$\begin{aligned}
\boldsymbol{\Sigma}^{(c)} &= -\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\left[\log\left\{f\left(\boldsymbol{y}^o|\boldsymbol{\eta}\right)\right\} + \log\left\{\left(\boldsymbol{r}|\boldsymbol{\eta}\right)\right\} + \log\left\{f\left(\boldsymbol{\eta}\right)\right\} + \boldsymbol{c}'H(\boldsymbol{\eta})\right] \\
&= \sum_{g=1}^{T}\sum_{i\in A_g}\frac{\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}'}{\sigma_j^2} - \sum_{i=1}^{n}\sum_{g=1}^{T}\left[\frac{\frac{\partial\phi(\lambda_{ig})}{\partial\lambda}\Phi\left(\lambda_{ig}\right) - \phi^2\left(\lambda_{ig}\right)}{\Phi^2\left(\lambda_{ig}\right)}\right]\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}' \\
&\quad + \boldsymbol{G}^{-1} - \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\left[\boldsymbol{c}'H(\boldsymbol{\eta})\right] \qquad (4.15)
\end{aligned}$$

where $\lambda_{ig} = (-1)^{1-r_{ig}}\left(\boldsymbol{w}_{ig}'\boldsymbol{\beta}_{mis} + \boldsymbol{z}_{ig}'\boldsymbol{\eta}\right)$ and $\partial\phi\left(\lambda_{ig}\right)/\partial\lambda = -\lambda_{ig}/\sqrt{2\pi}\exp(-\lambda_{ig}^2/2)$ is the derivative of the standard normal density function. Once the Newton-Raphson algorithm converges to an estimate $\widehat{\boldsymbol{\eta}}$, the next step is to apply a fully exponential Laplace approximation to $E[\exp\{\boldsymbol{c}'H(\boldsymbol{\eta})\}]$. We apply the result of Theorem 2 of Tierney et al. (1989). Using properties of the cumulant-generating function,

$$\begin{aligned}
\left\{E\left[H(\boldsymbol{\eta})|\boldsymbol{y}^o,\boldsymbol{r};\boldsymbol{\Psi}\right]\right\}_k &= \frac{\partial}{\partial c_k}\log\left\{E\left[\exp\left(\boldsymbol{c}'H(\boldsymbol{\eta})|\boldsymbol{y},\boldsymbol{r};\boldsymbol{\Psi}\right)\right]\right\}|_{\boldsymbol{c}=\boldsymbol{0}} \\
&\approx \frac{\partial}{\partial c_k}\left\{\boldsymbol{c}'H\left(\widehat{\boldsymbol{\eta}}^{(\boldsymbol{c})}\right) + \log\left[\det\left(\boldsymbol{\Sigma}^{(\boldsymbol{c})}\right)^{-1/2}\right]\right\}\Bigg|_{\boldsymbol{c}=\boldsymbol{0}} \\
&= \boldsymbol{e}_k'H(\widehat{\boldsymbol{\eta}}) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left\{\frac{\partial\boldsymbol{\Sigma}^{(c)}}{\partial c_k}\Bigg|_{(\boldsymbol{c},\boldsymbol{\eta})=(\boldsymbol{0},\widehat{\boldsymbol{\eta}})}\right\}\right) \qquad (4.16)
\end{aligned}$$

where $\boldsymbol{e}_k$ is the vector of zeros with a 1 in the k-th component. The kl-th component of $\mathrm{var}(\boldsymbol{\eta})$, evaluated at $(\boldsymbol{c},\boldsymbol{\eta}) = (\boldsymbol{0},\widehat{\boldsymbol{\eta}})$, is

$$\begin{aligned}
\left\{\mathrm{var}\left(\boldsymbol{\eta}|\boldsymbol{y}^o,\boldsymbol{r};\boldsymbol{\Psi}\right)\right\}_{kl} &= \frac{\partial^2}{\partial c_k\partial c_l}\log\left\{E\left[\exp\left(\boldsymbol{c}'\boldsymbol{\eta}|\boldsymbol{y},\boldsymbol{r};\boldsymbol{\Psi}\right)\right]\right\}|_{\boldsymbol{c}=\boldsymbol{0}} \\
&\approx \boldsymbol{e}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_l - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\Sigma}}{\partial c_k\partial c_l} - \boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial c_l}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial c_k}\right). \qquad (4.17)
\end{aligned}$$

The first-order Laplace approximation consists of only the first terms of Equations (4.16) and (4.17) (Kass and Steffey, 1989). The terms involving the trace function in both equations are the fully exponential corrections to the first-order Laplace approximation. Calculation of these terms is furnished in Section 4.7. These are similar to the expressions presented by Rizopoulos et al. (2009); however, their application is to a nested design where the likelihood factored over the subjects, where eight random effects are shared between a measurement and a time-to-dropout model due to a cubic spline with seven degrees of freedom. The multi-membership structure of our model affects the terms derived inside of the trace functions in Section 4.7 for Equations (4.16) and (4.17), as well as the computational complexity of the model.

4.7 Derivation of Terms for the E-step

The E-step calculations requires the terms $\partial\boldsymbol{\Sigma}^{(c)}/\partial c_k$, $\partial^2\boldsymbol{\Sigma}^{(c)}/\partial c_k\partial c_l$, and $\partial H(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$. Furthermore, the calculation of the first two of these terms requires calculation of $\partial\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k$ and $\partial^2\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k\partial c_l$, both evaluated at $\boldsymbol{c} = 0$. Rizopoulos, Verbeke, and Lesaffre (2009) performed these calculations for their model. We must calculate these quantities for our model, which differs from those of Rizopoulos et al. (2009) due to its multi-membership structure and use of a binary attendance indicator instead of a continuous time-to-event outcome. Furthermore, our model makes use of correlated instead of shared random effects. We will use the notational convention that, for example,

$$\frac{\partial\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}} = \frac{\partial\kappa(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\Big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^{(c)}}.$$

Let the scalars C_{ig} and D_{ig} be defined as

$$C_{ig} = (-1)^{1-r_{ig}}\,\frac{\frac{\partial^2\phi(\lambda_{ig})}{\partial\lambda^2}\Phi^2(\lambda_{ig}) - 3\frac{\partial\phi(\lambda_{ig})}{\partial\lambda}\Phi(\lambda_{ig})\phi(\lambda_{ig}) + 2\phi^3\left(\lambda_{ig}\right)}{\Phi^3\left(\lambda_{ig}\right)}$$

and

$$\begin{aligned} D_{ig} = & \frac{\frac{\partial^3\phi(\lambda_{ig})}{\partial\lambda^3}\Phi^3\left(\lambda_{ig}\right) - 4\frac{\partial^2\phi(\lambda_{ig})}{\partial\lambda^2}\Phi^2(\lambda_{ig})\phi(\lambda_{ig})}{\Phi^4\left(\lambda_{ig}\right)} \\ & + \frac{12\frac{\partial\phi(\lambda_{ig})}{\partial\lambda}\Phi(\lambda_{ig})\phi^2(\lambda_{ig}) - 3\left(\frac{\partial\phi(\lambda_{ig})}{\partial\lambda}\right)^2\Phi^2(\lambda_{ig}) - 6\phi^4(\lambda_{ig})}{\Phi^4\left(\lambda_{ig}\right)} \end{aligned}$$

with $\lambda_{ig} = (-1)^{1-r_{ig}} \left(\boldsymbol{w}'_{ig}\boldsymbol{\beta}_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}\right)$, where

$$\begin{aligned}
\frac{\partial\phi\left(\lambda_{ig}\right)}{\partial\lambda} &= -\frac{\lambda_{ig}}{\sqrt{2\pi}}\exp(-\frac{\lambda_{ig}^2}{2}) \\
\frac{\partial\phi^2\left(\lambda_{ig}\right)}{\partial\lambda^2} &= \frac{\lambda_{ig}^2 - 1}{\sqrt{2\pi}}\exp(-\frac{\lambda_{ig}^2}{2}) \\
\frac{\partial\phi^3\left(\lambda_{ig}\right)}{\partial\lambda^3} &= -\frac{\lambda_{ig}\left(\lambda_{ig}^2 - 3\right)}{\sqrt{2\pi}}\exp(-\frac{\lambda_{ig}^2}{2})
\end{aligned}$$

are the first, second, and third derivatives, respectively, of the standard normal density function. We first calculate $\partial\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k$. Let

$$\kappa(\boldsymbol{\eta}) = \log\left\{f\left(\boldsymbol{y}|\boldsymbol{\eta}\right)\right\} + \log\left\{f\left(\boldsymbol{r}|\boldsymbol{\eta}\right)\right\} + \log\left\{f\left(\boldsymbol{\eta}\right)\right\}.$$

Since, by definition, $\widehat{\boldsymbol{\eta}}^{(c)} = \mathrm{argmax}_{\boldsymbol{\eta}}[\log\left\{f(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\eta})\right\} + \boldsymbol{c}'H(\boldsymbol{\eta})]$, we have

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\boldsymbol{\eta}}\left\{\kappa(\boldsymbol{\eta}) + \boldsymbol{c}'H(\boldsymbol{\eta})\right\}_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^{(c)}} \\
&= \frac{\partial\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}} + \frac{\partial\boldsymbol{c}'H(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}}
\end{aligned}$$

Taking the derivative with respect to c_k yields

$$\frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k} + \frac{\partial\boldsymbol{e}'_k H(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}} + \left(\frac{\partial}{\partial c_k}\left\{\frac{\partial H(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}}\right\}\right)'\boldsymbol{c} = \mathbf{0}$$

Solving for $\partial\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k$ and evaluating at $c = 0$ gives

$$\begin{aligned}
\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k}|_{\boldsymbol{c}=0} &= \left(-\frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\right)^{-1}\left(\frac{\partial\boldsymbol{e}'_k H\left(\widehat{\boldsymbol{\eta}}\right)}{\partial\boldsymbol{\eta}}\right) \\
&= \boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{e}'_k H\left(\widehat{\boldsymbol{\eta}}\right)}{\partial\boldsymbol{\eta}}\right)
\end{aligned}$$

We only need the terms $\partial^2\boldsymbol{\Sigma}^{(c)}/\partial c_k\partial c_l$ and $\partial^2\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k\partial c_l$ for the case $H(\boldsymbol{\eta}) = \boldsymbol{\eta}$. These terms are used in the calculation of $\mathrm{var}\left(\boldsymbol{\eta}\right)$. To find $\partial^2\widehat{\boldsymbol{\eta}}^{(c)}/\partial c_k\partial c_l$ where

$H(\boldsymbol{\eta}) = \boldsymbol{\eta}$, note

$$
\begin{aligned}
&\frac{\partial}{\partial c_l}\left\{\frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k} + \frac{\partial \boldsymbol{e}_k' H(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}} + \left(\frac{\partial}{\partial c_k}\left\{\frac{\partial H(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}}\right\}\right)'\boldsymbol{c}\right\} = \mathbf{0}\\
&\Rightarrow \frac{\partial}{\partial c_l}\left\{\frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k} + \boldsymbol{e}_k\right\} = \mathbf{0}\\
&\Rightarrow \frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\frac{\partial^2\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k\partial c_l} + \frac{\partial^3\kappa(\widehat{\boldsymbol{\eta}}^{(c)})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'\partial\boldsymbol{\eta}'}\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_l}\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k} = \mathbf{0}\\
&\overset{\boldsymbol{c}=\mathbf{0}}{\Rightarrow} \frac{\partial^2\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k\partial c_l}|_{\boldsymbol{c}=\mathbf{0}} = \left(\frac{\partial^2\kappa(\widehat{\boldsymbol{\eta}})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\right)^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\boldsymbol{\eta}'}\frac{\partial\widehat{\boldsymbol{\eta}}}{\partial c_l}\frac{\partial\widehat{\boldsymbol{\eta}}}{\partial c_k}\\
&\Rightarrow \frac{\partial^2\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k\partial c_l}|_{\boldsymbol{c}=\mathbf{0}} = \boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}\sum_{g=1}^{T} C_{ig}\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}'\left(\boldsymbol{z}_{ig}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_l\right)\right]\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_k.
\end{aligned}
$$

Using the terms we have found, we may now calculate $\partial\boldsymbol{\Sigma}^{(c)}/\partial c_k$ and $\partial^2\boldsymbol{\Sigma}^{(c)}/\partial c_k\partial c_l$ at $(\boldsymbol{c},\boldsymbol{\eta}) = (\mathbf{0},\widehat{\boldsymbol{\eta}})$.

$$
\begin{aligned}
&\frac{\partial\boldsymbol{\Sigma}^{(\boldsymbol{c})}}{\partial c_k}|_{\boldsymbol{c}=\mathbf{0}}\\
=&\frac{\partial}{\partial c_k}\left[-\sum_{i=1}^{n}\sum_{g=1}^{T}\left[\frac{\frac{\partial\phi(\lambda)}{\partial\lambda}\Phi\left(\lambda\right)-\phi^2\left(\lambda\right)}{\Phi^2\left(\lambda\right)}\right]\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}' - \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\left[\boldsymbol{c}'H(\widehat{\boldsymbol{\eta}}^{(\boldsymbol{c})})\right]\right]_{\boldsymbol{c}=\mathbf{0}}\\
=&-\sum_{i=1}^{n}\sum_{g=1}^{T} C_{ig}\left(\boldsymbol{z}_{ig}'\frac{\partial\widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k}|_{\boldsymbol{c}=0}\right)\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}' - \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\left[\boldsymbol{e}_k'H(\widehat{\boldsymbol{\eta}})\right]\\
=&-\sum_{i=1}^{n}\sum_{g=1}^{T} C_{ig}\left[\boldsymbol{z}_{ig}'\boldsymbol{\Sigma}^{-1}\left(\frac{\partial\boldsymbol{e}_k'H\left(\widehat{\boldsymbol{\eta}}\right)}{\partial\boldsymbol{\eta}}\right)\right]\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}' - \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\left[\boldsymbol{e}_k'H(\widehat{\boldsymbol{\eta}})\right]
\end{aligned}
$$

For $H(\boldsymbol{\eta}) = \boldsymbol{\eta}$ and evaluating at $(\boldsymbol{c}, \boldsymbol{\eta}) = (\mathbf{0}, \widehat{\boldsymbol{\eta}})$

$$
\begin{aligned}
&\frac{\partial^2 \boldsymbol{\Sigma}^{(\boldsymbol{c})}}{\partial c_k \partial c_l} \\
=&\frac{\partial^2}{\partial c_k \partial c_l}\left[-\sum_{i=1}^{n}\sum_{g=1}^{T}\left[\frac{\frac{\partial \phi(\lambda)}{\partial \lambda}\Phi\left(\lambda\right) - \phi^2\left(\lambda\right)}{\Phi^2\left(\lambda\right)}\right]\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}' - \frac{\partial^2}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'}\left[\boldsymbol{c}' H(\widehat{\boldsymbol{\eta}}^{(\boldsymbol{c})})\right]\right]_{\boldsymbol{c}=\mathbf{0}} \\
=&\frac{\partial}{\partial c_l}\left[-\sum_{i=1}^{n}\sum_{g=1}^{T} C_{ig}\left(\boldsymbol{z}_{ig}'\frac{\partial \widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k}|_{\boldsymbol{c}=0}\right)\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}'\right]_{\boldsymbol{c}=\mathbf{0}} \\
=&-\sum_{i=1}^{n}\sum_{g=1}^{T}\left[\boldsymbol{z}_{ig}'\frac{\partial^2 \widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k \partial c_l}|_{\boldsymbol{c}=0} C_{ig}\right. \\
&\left. +\boldsymbol{z}_{ig}'\left(\frac{\partial \widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_k}|_{\boldsymbol{c}=0}\right)\boldsymbol{z}_{ig}'\left(\frac{\partial \widehat{\boldsymbol{\eta}}^{(c)}}{\partial c_l}|_{\boldsymbol{c}=0}\right) D_{ig}\right]\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}'
\end{aligned}
$$

Finally, for the two cases of $H(\boldsymbol{\eta})$, the required terms $\partial H(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$ and $\partial^2 H(\boldsymbol{\eta})/\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'$ are

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}}\left(\boldsymbol{e}_k'\boldsymbol{\eta}\right) &= \boldsymbol{e}_k \\
\frac{\partial^2}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'}\left(\boldsymbol{e}_k'\boldsymbol{\eta}\right) &= \mathbf{0} \\
\frac{\partial}{\partial \boldsymbol{\eta}}\left[(-1)^{1-r_{ig}}\frac{\phi(\lambda_{ig})}{\Phi(\lambda_{ig})}\right] &= \left(\frac{\frac{\partial \phi(\lambda_{ig})}{\partial \lambda}\Phi\left(\lambda_{ig}\right) - \phi^2\left(\lambda_{ig}\right)}{\Phi^2\left(\lambda_{ig}\right)}\right)\boldsymbol{z}_{ig} \\
\frac{\partial^2}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'}\left[(-1)^{1-r_{ij}}\frac{\phi(\lambda_{ig})}{\Phi(\lambda_{ig})}\right] &= C_{ig}\boldsymbol{z}_{ig}\boldsymbol{z}_{ig}'
\end{aligned}
$$

4.8 Convergence and EM Standard Errors

The EM algorithm converges to a stationary value of the approximated observed data likelihood as long as the E- and M-step updates are continuous in the model parameters, $\boldsymbol{\Psi}$ and the parameter space is compact (Wu, 1983). Although the parameter space for $\boldsymbol{\Psi}$ is not compact for our model, this regularity condition is satisfied by a truncation of the parameter space (McCulloch, 1994; Demidenko, 2004). The existence of the derivatives that yield the score functions in the M-step guarantees that the M-step is continuous with respect to $\boldsymbol{\Psi}$. Finally, the E-step functions appearing in Equations (4.16) and (4.17) are continuous functions of the elements of $\boldsymbol{\Psi}$. Due to the approximation used in the E-step, the limit point lies on the approximated likelihood function: the quality of this approximation is discussed in Section 4.10.

The observed data score vector $S(\boldsymbol{\Psi})$ used in the M-step is equal to the the conditional expectation of the complete data score vector (Louis, 1982). As proposed by Jamshidian and Jennrich (2000), we calculate the observed information matrix using a central difference approximation to the Hessian in Equation (4.18) at the MLE $\widehat{\boldsymbol{\Psi}}$. They also suggest using a Richardson extrapolation (Lindfield and Penny, 1988), though our experience has shown that this greatly increases the run-time without providing a noticeable change in the approximation.

$$- \partial S(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}|_{\boldsymbol{\Psi}=\widehat{\boldsymbol{\Psi}}} . \tag{4.18}$$

4.9 Verification of a Regularity Condition

The introduction of the missing data mechanism requires an extra step to verify that the exchange of differentiation and integration in Equation (3.17) is valid. The exchange involving the derivatives with respect to $\boldsymbol{\beta}_{obs}, \boldsymbol{G}$, and $\boldsymbol{R}$ are still valid by the result of Lehmann and Romano (2010), but we must check that $\partial/\partial\boldsymbol{\beta}_{mis}$ may be moved under the integral. For simplicity in exposition, we assume that $\boldsymbol{\beta}_{mis}$ is a scalar. According to Corollary 2.4.4 of Casella and Berger (2001), the exchange

$$\frac{\partial}{\partial \beta_{mis}} \int f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{mis}) \mathrm{d}\boldsymbol{\eta}_{mis} = \int \frac{\partial}{\partial \beta_{mis}} f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{mis}) \mathrm{d}\boldsymbol{\eta}_{mis}$$

is valid if

1. $f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})f(\boldsymbol{\eta}_{mis})$ is differentiable in β_{mis}
2. $\exists\ g(\boldsymbol{\eta}_{mis}, \beta_{mis})$ such that $\left| \frac{\partial}{\partial \beta_{mis}} f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{mis})|_{\beta_{mis}=\beta'_{mis}} \right| \leq g(\boldsymbol{\eta}_{mis}, \beta_{mis})$ for all β'_{mis} such that $|\beta'_{mis} - \beta_{mis}| \leq \delta_0$, where $\delta_0 > 0$ is a constant in $\boldsymbol{\eta}_{mis}$
3. $\int g(\boldsymbol{\eta}_{mis}, \beta_{mis}) \mathrm{d}\boldsymbol{\eta}_{mis} < \infty$

We now verify that these conditions are satisfied.

Proof. Letting $t_{ig} = (-1)^{1-r_{ig}}$, condition (1) is satisfied by the existence of the derivative

$$\begin{aligned}
&\left|\frac{\partial}{\partial \beta_{mis}} f(\boldsymbol{r}|\boldsymbol{\eta}_{mis}) f(\boldsymbol{\eta}_{mis})\right| \\
=&\left| f(\boldsymbol{\eta}_{mis}) \sum_{i=1}^{n}\sum_{g=1}^{T}\left[t_{ig} w_{ig} \phi(t_{ig}[w_{ig}\beta_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}]) \prod_{(k,l)\neq(i,g)} \Phi(t_{kl}[w_{kl}\beta_{mis} + \boldsymbol{z}'_{kl}\boldsymbol{\eta}_{mis}])\right]\right| \\
&\hfill (4.19) \\
\leq& f(\boldsymbol{\eta}_{mis}) \sum_{i=1}^{n}\sum_{g=1}^{T}\left[|w_{ig}| \phi(w_{ig}\beta_{mis} + \boldsymbol{z}'_{ig}\boldsymbol{\eta}_{mis}) \prod_{(k,l)\neq(i,g)} \Phi(w_{kl}\beta_{mis} + \boldsymbol{z}'_{kl}\boldsymbol{\eta}_{mis})\right] \qquad (4.20)
\end{aligned}$$

where Equation (4.19) results from the product rule and Equation (4.20) is a result of the triangle inequality, the monotonicity of the cumulative distribution function, and the non-negativity of probability density functions. Since there are a finite number of summands in Equation (4.20), there exist a maximum summand with indices (i^*, j^*). Thus we may continue by writing

$$\begin{aligned}
&\leq nTf(\boldsymbol{\eta}_{mis})\left[|w_{i^*j^*}| \phi(w_{i^*j^*}\beta_{mis} + \boldsymbol{z}'_{i^*j^*}\boldsymbol{\eta}_{mis}) \prod_{(k,l)\neq(i^*,j^*)} \Phi(w_{kl}\beta_{mis} + \boldsymbol{z}'_{kl}\boldsymbol{\eta}_{mis})\right] \\
&\leq nTf(\boldsymbol{\eta}_{mis})|w_{i^*j^*}|\phi(w_{i^*j^*}\beta_{mis} + \boldsymbol{z}'_{i^*j^*}\boldsymbol{\eta}_{mis}) \qquad (4.21) \\
&\leq nT|w_{i^*j^*}|f(\boldsymbol{\eta}_{mis}) \qquad (4.22)
\end{aligned}$$

where Equation (4.21) holds because the product of distribution functions is bounded above by 1 and Equation (4.22) holds because the standard normal density is bounded above by 1. The expression in Equation (4.22) does not depend on β_{mis}, meaning condition (2) is satisfied for any $\delta_0 > 0$ by

$$g(\boldsymbol{\eta}_{mis}, \beta_{mis}) = nT|w_{i^*j^*}|f(\boldsymbol{\eta}_{mis}).$$

Finally, condition (3) is satisfied because

$$\int nT|w_{i^*j^*}|f(\boldsymbol{\eta}_{mis})\mathrm{d}\boldsymbol{\eta}_{mis} = nT|w_{i^*j^*}| < \infty \ \square$$

4.10 Approximation Error

Theorem 1 of Tierney et al. (1989) demonstrates that, for the approximations appearing in Equations (4.16) and (4.17),

$$\begin{aligned} E\left[H\right] &= \hat{E}\left[H\right] + O(\lambda^{-2}) \\ V\left[H\right] &= \hat{V}\left[H\right] + O(\lambda^{-3}) \end{aligned}$$

where the hat denotes the fully exponential Laplace approximation, and $\lambda = \lambda(\boldsymbol{y}^o, \boldsymbol{r})$ is a measure of the size of the data set $(\boldsymbol{y}^o, \boldsymbol{r})$ such that $\lambda \to \infty$ as the size of the data set grows (Evans and Swartz, 1995). In a nested model where the Laplace approximation is applied separately to each cluster, λ is equal to the smallest number of observations in each cluster. However, the joint model presents a multi-membership random effects structure in $f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})$ and a choice of random effects structures for $f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})$. Nevertheless, λ may be expressed as a function of $\boldsymbol{r}$ alone in the CPM. Consider the application of the Laplace approximation to the marginal likelihood,

$$\begin{aligned} &\iint f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis})\mathrm{d}\boldsymbol{\eta}_{obs}\mathrm{d}\boldsymbol{\eta}_{mis} \\ =&\int f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})\left\{\int f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis})\mathrm{d}\boldsymbol{\eta}_{obs}\right\}\mathrm{d}\boldsymbol{\eta}_{mis} \\ =&\int f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})I(\boldsymbol{\eta}_{mis};\boldsymbol{y}^o)\mathrm{d}\boldsymbol{\eta}_{mis} \end{aligned} \tag{4.23}$$

The Laplace approximation is exact for $\int f(\boldsymbol{y}^o|\boldsymbol{\eta}_{obs})f(\boldsymbol{\eta}_{obs}, \boldsymbol{\eta}_{mis})\mathrm{d}\boldsymbol{\eta}_{obs}$, since the $\boldsymbol{\eta}_{obs}$ enter the model linearly (Pinheiro and Bates, 1995). The result, $I(\boldsymbol{\eta}_{mis};\boldsymbol{y}^o)$, is normally distributed with mean and variance depending on the covariance between $\boldsymbol{\eta}_{obs}$ and $\boldsymbol{\eta}_{mis}$. For example, under an assumption of MAR, these effects are uncorrelated – and hence independent, due to their joint normality – and $I(\boldsymbol{\eta}_{mis}; \boldsymbol{y}^o) = f(\boldsymbol{\eta}_{mis})f(\boldsymbol{y}^o)$.

Applying the Laplace approximation to Equation (4.23) will result in approximation error because the $\boldsymbol{\eta}_{mis}$ enter the integrand non-linearly through the non-linear link in $f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})$. The approximation error depends on the amount of information in $\boldsymbol{r}$ and the random effects structure of $f(\boldsymbol{r}|\boldsymbol{\eta}_{mis})$. If the missing data mechanism depends only on random teacher effects, then λ equals the minimum classroom size (out of every classroom included in the data set), since the number of students in a teacher's class determines the amount of information associated with that teacher's effect in the missing data model, along the lines of the work by Vonesh (1996). Likewise, if only student effects are included in the missing data mechanism, then λ is the number of years T in the study. Note that this does not assume that the data are balanced, since each student will have an attendance indicator recorded regardless of whether or not they have a score recorded.

If both student and teacher effects are included in the missing data model, then the dimension of the integral in Equation (4.23) increases with the sample size. This property is not typical for applications of the Laplace approximation; Shun and McCullagh (1995) describe how a modification to the first-order Laplace approximation is needed to retain its usual order of convergence, although they did not study the behavior of the fully exponential approximation, and note that it may not suffer the same extent of deterioration in this setting as the first-order approximation. The results of Shun (1997) show, in an application to the salamander mating data, that the uncorrected first-order Laplace approximation outperforms penalized quasi-likelihood (PQL), which is equivalent to the pseudo-likelihood method used by SAS PROC GLIMMIX (Wolfinger and O'Connell, 1993). This is reasonable since PQL makes an additional approximation to one of the terms in the first-order Laplace approximation. At worst, the (fully exponential) Laplace approximation presents an improvement over PQL in this setting, even if consistency is not guaranteed. PQL is widely used – such as in SAS PROC GLIMMIX – despite the the fact that it produces potentially biased estimates, because PQL makes some problems computationally feasible that otherwise would not be (Broatch and Lohr, 2011). Even the bias-corrected PQL of Lin and Breslow (1996) is inconsistent (Jiang, 2007). We rely on the fully exponential Laplace approximation, despite its potential inconsistency when both student and teacher effects are included in the missing data mechanism, because it relies on fewer approximations than PQL, which we would have used if SAS were capable of fitting the model. The only other alternative would have been to use MCMC methods, whose drawbacks for high-dimensional integrals in multi-level models were discussed in Section 2.4.

4.11 Computation

We extend the work used to develop R package GPvam in Chapter 3 to accommodate the joint model, relying heavily on the sparse matrix methods of the Matrix package (Bates and Maechler, 2011). The program is computationally demanding – despite the use of the Laplace approximation instead of Monte Carlo methods for the E-step – due to its multi-membership structure. In fact, the GP model itself is sufficiently complex that no scalable maximum-likelihood estimation routine has been available for it until GPvam.

The multimembership random effects structure causes difficulties, but at least the design matrix for the random effects, $\boldsymbol{S}$, is sparse. In an example using calculus grades in Chapter 5, the length of $\boldsymbol{\eta}$ is 4257, and $\boldsymbol{S}$ is a 9271 by 4257 matrix. Only 0.0434% of the components are nonzero. Storing the sparse version of $\boldsymbol{S}$ takes 223 KB instead of the 316 MB needed for the dense matrix. Taking advantage of this sparsity with the R package Matrix (Bates and Maechler, 2011) greatly improves the performance of the program, since many of the calculations in the estimation of the model involve products with $\boldsymbol{S}$.

The fixed effects for both the score and missing data models are conveniently specified with R formula objects, and the $\boldsymbol{X}$ and $\boldsymbol{W}$ matrices are constructed via calls to the function `sparse.model.matrix`. The design matrices for the random effects, $\boldsymbol{S}$ and $\boldsymbol{Z}$, however, have irregular structures, and we build them with custom functions. The identity matrix is used as the initial value for $\boldsymbol{G}$. To test the of the algorithm for sensitivity to this choice, we analyzed the examples of Chapter 5 with 10 different choices for initial values for $\boldsymbol{G}$, obtained by taking the nearest positive-definite matrix to the symmetric part of an appropriately sized matrix populated with random values from the standard normal distribution. All 10 analyses converged to the same parameter estimates, and they were mostly in agreement after only a few EM iterations. The yearly error variances that appear in the diagonal of the $\boldsymbol{R}$ matrix are initially set to 1. The initial values for the fixed effects are obtained by performing the M-step update using the initial $\boldsymbol{G}$ matrix and setting $\boldsymbol{\eta} = \mathbf{0}$. This produces the ordinary least-squares estimates of the fixed effects.

The convergence criterion for the joint model involves the maximum relative change in parameters. Namely, we declare convergence if the maximum change in each parameter is less than a certain tolerance. The criterion signals convergence if for each component Ψ_k of the vector of parameters $\boldsymbol{\Psi}$,

$$\begin{cases} \left|\Psi_k^{p+1} - \Psi_k^p\right| < 0.0001 & \text{if } |\Psi_k^p| \leq 0.01 \\ \left|\frac{\Psi_k^{p+1} - \Psi_k^p}{\Psi_k^p}\right| < 0.0001 & \text{if } |\Psi_k^p| > 0.01 \end{cases}$$

An advantage of the EM algorithm over Newton type algorithms is that no restrictions need to be placed on the $\boldsymbol{G}$ matrix to ensure that it is positive-definite after each iteration. This is an advantage because the future year effects tend to be highly correlated, placing $\boldsymbol{G}$ near the boundary of the parameter space. To see why this is true, notice that $\boldsymbol{G}$ is a block-diagonal portion of $\widetilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$. The matrix $\widetilde{\boldsymbol{v}}$ is the inverse of $\boldsymbol{\Sigma}$ (plus some correction terms for the joint model), which is defined by Equation (4.15). Thus, $\widetilde{\boldsymbol{v}}$ is positive definite as long as $\boldsymbol{\Sigma}$ is. Looking at Equation (4.15), it is clear that $\boldsymbol{\Sigma}$ is positive definite as long as the initial $\boldsymbol{G}$ is positive definite, because the first two terms in the equation are positive semi-definite (the last term in the equation is irrelevant because $c = 0$). Furthermore, $\widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$ is positive semi-definite, and the sum of a positive definite and a positive semi-definite matrix is positive definite. A similar discussion appears in Demidenko (2004). It is possible, however, that the fully exponential corrections for the joint model in Equation (4.17) disrupt the positive-definiteness of $\widetilde{\boldsymbol{v}}$.

The conditional variance $\widetilde{\boldsymbol{v}}$ from Rizopoulos et al. (2009) is an 8×8 matrix. By contrast, our joint model for university calculus data called MNAR-t in Chapter 5 produces a $\widetilde{\boldsymbol{v}}$ of dimension 4265×4265. Thus the computational burden of calculating the fully exponential corrections in Equation (4.17) is tremendous. Each of the matrices $\boldsymbol{\Sigma}^{-1}$, $\frac{\partial \boldsymbol{\Sigma}}{\partial c_k}$, and $\frac{\partial^2 \boldsymbol{\Sigma}}{\partial c_k \partial c_l}$ have dimension equal to the number of random effects. Even ignoring the calculations required to obtain the derivatives $\frac{\partial \boldsymbol{\Sigma}}{\partial c_k}$ and $\frac{\partial^2 \boldsymbol{\Sigma}}{\partial c_k \partial c_l}$, the term inside the trace function of Equation (4.17) requires approximately $4(2l^3 - l^2)$ calculations, where l is the length of $\boldsymbol{\eta}$ (since multiplication of two $l \times l$ matrices requires $2l^3 - l^2$ arithmetic operations). For example, the calculus data example in Chapter 5 contains 4265 random effects in the joint model. This results in a requirement of just over 6.2×10^{11} operations to calculate the fully exponential correction for a single component of $\widetilde{\boldsymbol{v}}$, if implemented naively. Since, in this example, there are $4256 * 4257/2 = 9058896$ components in the upper-triangle of the symmetric $\widetilde{\boldsymbol{v}}$ matrix, this implementation would result in a need for around 5.6×10^{18} operations per iteration, excluding the operations needed to calculate the derivatives. The author's computer runs at 7.2 billion floating point operations per second, and would require around 25 years to execute one iteration of a naively implemented fully exponential correction to the full $\widetilde{\boldsymbol{v}}$ matrix.

In order to compute the fully exponential approximation in a reasonable amount of time, we use a few key facts. First, $\widetilde{\boldsymbol{v}}$ is symmetric and only the upper triangle needs to be calculated. Secondly, the operations inside of the trace function in Equation (4.17) contain some elements that are common to all of the components of $\widetilde{\boldsymbol{v}}$ and do not need to be re-calculated for each component. Thirdly, some of the needed computations involve multiplication by sparse matrices, greatly reducing the number of needed arithmetic operations (this is where our true gain in performance lies). Finally, it turns out that not all of the components of $\mathrm{var}\,(\boldsymbol{\eta}|\boldsymbol{y}^o, \boldsymbol{r}; \boldsymbol{\Psi})$ will be used in the M-step. The unused components may be ignored.

To see why only some of the components of $\widetilde{\boldsymbol{v}}$ are needed, observe how the covariance matrix $\widetilde{\boldsymbol{v}}$ is used in the M-step updates of $\boldsymbol{\Gamma}_g$ and σ_g^2 in Equations (4.10) and (4.11), respectively. The M-step update of $\boldsymbol{\Gamma}_g$ requires only relatively small block-diagonal portions of $\widetilde{\boldsymbol{v}}$, while the M-step update of σ_g^2 requires $\sum_{i \in B_g} \mathrm{tr}\left(\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}'\widetilde{\boldsymbol{v}}\right) = \sum_{row}\sum_{column}\left(\boldsymbol{S}'\boldsymbol{S} \circ \widetilde{\boldsymbol{v}}\right)$, where $\circ$ represents the Hadamard product. Thus, the only components of $\widetilde{\boldsymbol{v}}$ that are needed in addition to the aforementioned block-diagonal elements are those that have the same indices as non-zero components of the sparse matrix $\boldsymbol{S}'\boldsymbol{S}$. In the calculus data example, only 14132 (1.5%) of the upper-triangular components of $\widetilde{\boldsymbol{v}}$ need to be calculated. Combined with the methods described above, we are able to calculate an entire EM iteration, using the fully exponential corrections, in a mere 30 hours for this example. To speed our program even further, we use the following result.

From experience, we have found that using the trace corrections to $\widetilde{\boldsymbol{\eta}}$ found in Equation (4.16), but not the trace corrections to $\widetilde{\boldsymbol{v}}$ in Equation (4.17) produces results that are extremely close to those obtained by using all of the trace corrections. To see why this works, recall that the M-step update for the $\boldsymbol{G}$ matrix is composed of diagonal blocks from the matrix $\widetilde{\boldsymbol{v}} + \widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}'$. Hence the $\boldsymbol{G}$ matrix does experience at least part of the benefit of the fully exponential trace corrections when the corrections to $\widetilde{\boldsymbol{\eta}}$ are included. Breslow and Lin (1995) showed that using the first-order Laplace approximation results in downward bias in the variance of random effects in a GLMM, but that using a second-order approximation substantially reduces the bias. Likewise, our simulations have shown that the fully-exponential approximation results in larger estimates for the variance of the teacher missingness effect than the first-order approximation. Furthermore, the fully-exponential corrections to $\boldsymbol{\eta}$ account for approximately 90% of the increase observed when the corrections to both $\widetilde{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{v}}$ are included. This is explored further in Chapter 5. We take advantage of this by first running the algorithm to convergence using only the corrections to $\widetilde{\boldsymbol{\eta}}$, and then treating the resulting parameter estimates as initial values for the fully exponential approximation, which may only require a few further iterations. In the calculus data example, we are able to calculate an entire EM iteration with a fully exponential correction for $\widetilde{\boldsymbol{\eta}}$ in just under 2.5 minutes.

The E-step maximizes the conditional complete data likelihood with respect to the random effects $\boldsymbol{\eta}$. The required NR algorithm runs reasonably fast because the multiplications involve sparse matrices. The gradient used by the Newton-Raphson algorithm is $\boldsymbol{L}(\boldsymbol{\eta})$ from Equation (4.14) and the Hessian $\boldsymbol{\Sigma}^{(0)}$ from Equation (4.15). However, programming the M-step equations as they are written is sometimes inefficient and may cause computational difficulties. In several instances, we improved the performance of the program by using linear algebra to express quantities in equivalent, but more computationally efficient forms. For example, the evaluation $\log(\det(\widetilde{\boldsymbol{v}}))$ will fail because in several applications $\det(\widetilde{\boldsymbol{v}})$ is larger than the word size of the computer. Instead, we obtain the log of this determinant by taking 2 times the sum of the logarithm of the diagonal elements of the Cholesky decomposition of $\widetilde{v}$. We also take advantage of the vectorized data structure of R, avoiding the use of loops wherever possible.

Chapter 5

APPLICATIONS

In this chapter, we run simulations to examine the performance of the EM algorithms for estimating the GP VAM and the correlated random-effects model, obtain maximum likelihood estimates of the GP VAM using the data analyzed by Mariano et al. (2010), and apply our joint model to two real data sets. We first test the behavior of our EM estimation of the GP model with no missing data. Afterward, we examine results from the application of the joint model to simulated data sets. Sections 5.4 and 5.5 consider elementary school and university data sets, respectively.

5.1 Simulation for the GP VAM

We generate and analyze 500 datasets, simulating three years of observations with 25 teachers in each grade and 30 students per teacher. For the first simulation (GP1), we generate yearly means $\mu_1 = \mu_2 = \mu_3 = 0$, yearly error variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.5$, student variance $\sigma_S^2 = 1$, year 1 teacher variance

$$\mathbf{\Gamma}_1 = \begin{pmatrix} g_1^2 & g_{12} & g_{13} \\ g_{12} & g_2^2 & g_{23} \\ g_{13} & g_{23} & g_3^2 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.7 & 0.49 \\ 0.7 & 1.0 & 0.7 \\ 0.49 & 0.7 & 1.0 \end{pmatrix},$$

year 2 teacher variance

$$\mathbf{\Gamma}_2 = \begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix},$$

and year 3 teacher variance $k^2 = 1$.

Table 5.1 summarizes the mean of the parameter estimates from the simulations, as well as a p-value from a two-tailed t-test for equality of the mean of the estimates and the true parameter values. The off-diagonal terms of $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are

Table 5.1: Results of Simulation GP1

	True Value	Mean	p-value
μ_1	0	-0.0087	0.312
μ_2	0	-0.0008	0.953
μ_3	0	0.0130	0.437
σ_1^2	0.5	0.4979	0.228
σ_2^2	0.5	0.5012	0.488
σ_3^2	0.5	0.5017	0.320
σ_S^2	1	1.0045	0.106
g_1^2	1	0.961	0.006
g_2^2	1	0.9553	<0.001
g_3^2	1	0.9625	0.004
h_1^2	1	0.9554	0.001
h_2^2	1	0.9698	0.028
k^2	1	0.9958	0.753

not listed to prevent cluttering the table: estimates for those parameters were unbiased. There is evidence of some downward bias in the variance components of the teacher effects. It is well known that the maximum-likelihood (ML) estimates for variance components in mixed models are subject to a downward bias, as stated in Demidenko (2004). In a linear regression model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, the ML estimate for σ^2 is SS/n, where SS is the residual sum of squares, and n is the sample size. The unbiased estimate is $SS/(n-p)$, where p is the number of fixed effects, meaning the ML estimate $\widehat{\sigma}^2_{ML}$ is subject to a downward bias, $E[\widehat{\sigma}^2_{ML}] = \sigma^2(n-p)/n$. While this particular bias factor is not the same for the variance components of a mixed model, it is interesting to note that our example includes 25 teachers and one fixed effect in each year. This roughly corresponds to estimating the variance of 25 observations, which would yield a biased estimate of $24/25 = 0.96$ times the true value. This is exactly what we observed for the variance components of the year one and two teachers in our simulation.

We ran this simulation using the package GPvam, which is developed in Chapter 3. Each data set requires an average of 15 iterations to converge, and around 0.9 seconds per iteration. In this example, the design matrix, $\boldsymbol{S}$, for the random effects is 2250×900. Of the 2025000 components of $\boldsymbol{S}$, only 6750 (0.003%) are nonzero. As a simple illustration of how our program benefits from considering the sparseness of the matrices, the calculation $\boldsymbol{S}'\boldsymbol{S}$ using the dense representation of $\boldsymbol{S}$ takes an average of 2.22 seconds in R when using the `crossprod` function: this is twice the amount of time taken by our program to execute an entire iteration. By contrast, using the sparse representation of $\boldsymbol{S}$, the operation takes less than 0.01 seconds. Furthermore, since $\boldsymbol{S}'\boldsymbol{S}$ and other functions of the design matrices are constant for each iteration, we calculate these values at the beginning of the program and store them.

We ran two additional simulations for the GP model, being careful to re-seed our random number generator before each simulation. Simulation GP2 uses the same parameters as GP1, except the off-diagonal terms in the matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are set to 0, meaning that the current and future year teacher effects are assumed to be uncorrelated with each other. The results appear in Table 5.2. Again, the estimates of the off-diagonal terms of $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are unbiased.

Simulation GP3 sets the off-diagonal terms to .9, which mimics a complete persistence VAM. We cannot set these values to 1 because the covariance matrix for the random effects needs to be positive definite. We would need to make programming changes to actually impose the restriction of complete persistence. The results of this simulation appear in Table 5.3. There is some downward bias in the off diagonal terms of $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, but this is not surprising given the downward bias in the diagonal terms of those matrices. Interestingly, the estimates for the error variance terms and the student effect variance appear to be unbiased. One implication

Table 5.2: Results of Simulation GP2

	True Value	Mean	p-value
μ_1	0	0.0021	0.818
μ_2	0	0.0023	0.866
μ_3	0	0.0038	0.811
σ_1^2	0.5	0.4975	0.155
σ_2^2	0.5	0.4996	0.811
σ_3^2	0.5	0.5015	0.375
σ_S^2	1	1.0050	0.078
g_1^2	1	0.9484	<0.001
g_2^2	1	0.9856	0.279
g_3^2	1	0.9883	0.403
h_1^2	1	0.9732	0.045
h_2^2	1	0.9857	0.273
k^2	1	0.9697	0.020

of this finding is that the percent of variation in test scores that is due to teacher effects will be slightly underestimated. In this simulation, 40% of the variance in year 1 scores is attributed to teachers during the simulation. For comparison, the estimated percentage of variation due to the teacher effects is 39%. At least in this simulation, the downward bias in the estimated variance of teacher effects does not lead to a practically significant difference in the estimated percentage of variation due to teachers.

5.2 Simulation for the Joint Model

We next perform a simulation study for the joint model with missing data. For the joint model, we simulate 150 data sets containing two years of observations on 750 students, with 30 students assigned to each of 25 teachers in each year. The parameters for the test score model are generated according to the joint model with teacher effects in the missing data mechanism, using the following parameters. The yearly means are $\mu_1 = \mu_2 = 0$, yearly error variances are $\sigma_1^2 = \sigma_2^2 = 0.5$, student

Table 5.3: Results of Simulation GP3

	True Value	Mean	p-value
μ_1	0	-0.0041	0.675
μ_2	0	0.0125	0.345
μ_3	0	-0.0156	0.323
σ_1^2	0.5	0.5013	0.466
σ_2^2	0.5	0.4982	0.282
σ_3^2	0.5	0.4983	0.310
σ_S^2	1	0.9990	0.726
g_1^2	1	0.9564	0.001
g_2^2	1	0.9574	0.001
g_3^2	1	0.9639	0.008
h_1^2	1	0.9616	0.002
h_2^2	1	0.9733	0.032
k^2	1	0.9794	0.109

variance $\sigma_S^2 = 1$, year 1 teacher variance

$$\mathbf{\Gamma}_1 = \begin{pmatrix} g_1^2 & g_{12} & g_{13} \\ g_{12} & g_2^2 & g_{23} \\ g_{13} & g_{23} & g_3^2 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.7 & 0.5 \\ 0.7 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{pmatrix},$$

and year 2 teacher variance equal to 1. The mean for the missing data mechanism, which models the year 2 attendance, is 0.2, which removes observations from the second year so that on average 58% of the students from year 1 completed year 2. The probability of students completing year 2 was allowed to depend on random teacher effects corresponding to their year 1 teachers via the missing data mechanism specified by Equation (4.5), where the $\boldsymbol{\eta}_{mis}$ in this simulation are correlated with the $\boldsymbol{\eta}_{obs}$ via the last row/column of $\mathbf{\Gamma}_1$. We present the results of the simulation using these particular parameters for illustration. It would be possible to run several other parameter settings as determined by a designed experiment. However, the joint model is extremely computationally demanding, despite the computational efficiency achieved in Chapter 4. Analyzing the number of data sets required for a simulation analysis is time consuming. Recall that the calculus data example requires around 30 hours for each iteration that applies the fully exponential corrections to the Laplace approximations of both the conditional mean and conditional variance of the random effects in the E step.

Each simulated data set is analyzed with the GP model under an assumption of MAR, as well as with the joint model. We would expect the joint model to provide a better fit for the data since the observations were removed according to the structure assumed by the joint model: the simulation allows us to see which parameters experience the greatest improvement in their estimates as a result of using the joint model. Table 5.4 presents the mean squared error (MSE) for the 150 parameter estimates from the analyses of the simulated data using both the MAR and joint models. For the parameters used in this particular simulation, the fixed effects experienced the greatest improvement in MSE and most of the teacher effects recorded some improvement; however, the estimates for the error variances, student variance, and the current year effect of the first grade teachers were unaffected. Since missing data were introduced only in year 2, it seems reasonable that the current year effects of the first grade teachers did not change. It is not clear, however, why the MAR estimates for the other variance components were robust in this simulation.

Table 5.4: MSE for 150 Simulations

Parameter	MAR	Joint
Grade 1 mean score	0.0381	0.0371
Grade 2 mean score	0.0659	0.0593
σ_1^2	0.0037	0.0037
σ_2^2	0.0039	0.0039
Student variance	0.0068	0.0068
Grade 1: (1,1)	0.0884	0.0884
Grade 1: (2,1)	0.0358	0.0351
Grade 1: (2,2)	0.0248	0.0242
Grade 2: (1,1)	0.0793	0.0790

We also tested the behavior of the model under misspecification of the missing data mechanism. In one example, we remove data according to student effects, but only model the simulated data with random teacher effects in the missing data mechanism. In another example, we remove second year observations according to teacher effects, but only model random student effects in the missing data mechanism. In both cases, the joint model produces estimates with the same MSE as the estimates from the GP model when ignoring the missing data. The joint model does not detect the nonignorable missingness in this case, because it does not match the form of dropout modeled by the missing data mechanism. It is reassuring to see that the inclusion of the misspecified missing data mechanism does not negatively impact the parameter estimates of the VAM.

Another interesting aspect of the joint model worth inspecting is the comparative performance of the fully exponential approximation to the first order Laplace approximation. It turns out that using the fully exponential correction for the conditional mean $\widetilde{\boldsymbol{\eta}}$, but not the conditional variance $\widetilde{\boldsymbol{v}}$ improves the accuracy of the calculation in a fraction of the time. For the simulated 150 data sets, the first order Laplace approximation yields smaller estimates of the random teacher effect variance component in the missing data model than the fully exponential approximation, consistent with the downward bias reported by Breslow and Lin (1995) and Lin and Breslow (1996). Calculating the fully exponential correction only for $\widetilde{\boldsymbol{\eta}}$ accounts for an average of 91% (Q1=90%, Q3=92%) of the difference between the first order and fully exponential estimates of the variance component, while requiring only 0.2% of the additional computational time needed to compute the complete fully exponential approximation. For this simulation, the mean time per iteration is 6.8 seconds for the first order Laplace approximation, 7.8 seconds for the fully exponential corrections to $\widetilde{\boldsymbol{\eta}}$ only, and 560 seconds for the fully exponential approximation. We take advantage of this discovery in the implementation of the joint model by first running the algorithm to convergence using only the corrections to $\widetilde{\boldsymbol{\eta}}$, and then treating the resulting parameter estimates as initial values for the fully exponential approximation. The algorithm then requires only relatively few iterations with the computationally expensive fully exponential corrections to $\widetilde{\boldsymbol{v}}$.

5.3 Application of the GP VAM assuming MAR

We apply the GP VAM with the package GPvam developed in Chapter 3 to the data set analyzed by Mariano et al. (2010), which is available in the supplementary material of McCaffrey and Lockwood (2011). According to McCaffrey and Lockwood (2011), the data come from vertically linked mathematics standardized test scores from grades 1–5 for a cohort of students from a large urban US school district.

The data have been pre-processed by McCaffrey and Lockwood (2011), and we further processed the data by removing observations with no student link, as well as observations missing both the test score and the teacher link. The resulting data set consists of 26019 observations on 9295 students over 5 years. For grades 1 through 5, there are 338, 318, 306, 321, and 259 teachers, respectively. This results in a total of 4781 teacher effects for the GP model to estimate.

We assume that unbalanced student profiles are due to observations that are missing at random (Little and Rubin, 2002). This amounts to assuming that the probability of a student being observed in a given year does not depend on his latent level of ability, his teacher history, or the score that he would have recorded if it had been observed. Applications of the MNAR model developed in Chapter 4 appear in Sections 5.4 and 5.5.

The estimated covariance matrices, with associated correlation matrices, are

$\boldsymbol{R}$:

$$\begin{pmatrix} 0.741 & 0.478 & 0.463 & 0.456 & 0.392 \\ 0.478 & 0.705 & 0.523 & 0.516 & 0.449 \\ 0.463 & 0.523 & 0.736 & 0.563 & 0.484 \\ 0.456 & 0.516 & 0.563 & 0.688 & 0.509 \\ 0.392 & 0.449 & 0.484 & 0.509 & 0.565 \end{pmatrix} \begin{pmatrix} 1.000 & 0.661 & 0.626 & 0.639 & 0.606 \\ 0.661 & 1.000 & 0.726 & 0.740 & 0.711 \\ 0.626 & 0.726 & 1.000 & 0.791 & 0.750 \\ 0.639 & 0.740 & 0.791 & 1.000 & 0.817 \\ 0.606 & 0.711 & 0.750 & 0.817 & 1.000 \end{pmatrix}$$

$\boldsymbol{\Gamma}_1$:

$$\begin{pmatrix} 0.443 & 0.121 & 0.120 & 0.107 & 0.095 \\ 0.121 & 0.100 & 0.088 & 0.084 & 0.077 \\ 0.120 & 0.088 & 0.087 & 0.083 & 0.076 \\ 0.107 & 0.084 & 0.083 & 0.080 & 0.074 \\ 0.095 & 0.077 & 0.076 & 0.074 & 0.069 \end{pmatrix} \begin{pmatrix} 1.000 & 0.575 & 0.610 & 0.568 & 0.541 \\ 0.575 & 1.000 & 0.941 & 0.941 & 0.920 \\ 0.610 & 0.941 & 1.000 & 0.994 & 0.986 \\ 0.568 & 0.941 & 0.994 & 1.000 & 0.995 \\ 0.541 & 0.920 & 0.986 & 0.995 & 1.000 \end{pmatrix}$$

$\boldsymbol{\Gamma}_2$:

$$\begin{pmatrix} 0.281 & 0.059 & 0.039 & 0.042 \\ 0.059 & 0.025 & 0.023 & 0.020 \\ 0.039 & 0.023 & 0.024 & 0.020 \\ 0.042 & 0.020 & 0.020 & 0.017 \end{pmatrix} \begin{pmatrix} 1.000 & 0.703 & 0.478 & 0.593 \\ 0.703 & 1.000 & 0.951 & 0.967 \\ 0.478 & 0.951 & 1.000 & 0.979 \\ 0.593 & 0.967 & 0.979 & 1.000 \end{pmatrix}$$

$\boldsymbol{\Gamma}_3$:

$$\begin{pmatrix} 0.248 & 0.032 & 0.024 \\ 0.032 & 0.015 & 0.015 \\ 0.024 & 0.015 & 0.015 \end{pmatrix} \begin{pmatrix} 1.000 & 0.516 & 0.394 \\ 0.516 & 1.000 & 0.979 \\ 0.394 & 0.979 & 1.000 \end{pmatrix}$$

$\boldsymbol{\Gamma}_4$:

$$\begin{pmatrix} 0.130 & 0.038 \\ 0.038 & 0.030 \end{pmatrix} \begin{pmatrix} 1.000 & 0.612 \\ 0.612 & 1.000 \end{pmatrix}$$

Table 5.5: Estimates for yearly means

	Estimate	Std. Error
Year 1	3.395	0.030
Year 2	3.996	0.029
Year 3	4.726	0.023
Year 4	5.309	0.022
Year 5	5.984	0.025

$\mathbf{\Gamma}_5 : 0.146$

Estimates for the yearly means appear in Table 5.5.

Using the EM algorithm, we obtain correlation patterns that are similar to those in Figures 2 and 3 of Mariano et al. (2010). However, we note that Mariano et al. (2010) obtained these results after careful choice of an informative prior that allowed for strong correlations between future year effects. In simulation studies, they found that a minimally informative Wishart prior for covariance parameters could result in posterior credible intervals for the correlations that did not include the true values. The EM algorithm gives maximum likelihood estimates that do not need any specifications of prior distributions.

5.4 Effects of Missing Data in an Urban School District

This section applies the joint model to data from grade-schools from an urban school district. The data set tracks a cohort of 2834 students from grades 4 through 6, recording their score on a standardized math test each year. The GP VAM is needed here because the tests are not vertically scaled; that is, the tests are not designed so that they can be compared across years on the same scale. Gain score VAMs may give misleading results in this setting. As expected, some students have missing observations. The absences could be the result of a medical appointment, the student skipping school, or the student's transfer to a new school district. The dataset is analyzed with the generalized persistence model assuming missing observations are ignorable, and with the joint model. Three different formulations of the missing data mechanism are used in the joint model in this example for a sensitivity analysis. Model MNAR-t fits only teacher random effects in the missing data mechanism, and MNAR-s fits only student effects. Model MNAR-b contains both random student and teacher effects in the missing data mechanism. When only teacher effects are included, the missing data mechanism models the probability that students who attended year j also attend year $j+1$, for $j = 1, \ldots, T-1$. When student effects are included, the student attendance in year 1 is also modeled.

The data set contains 102, 104, and 98 fourth, fifth, and sixth grade teachers, respectively. This leads to the modeling of 612 teacher effects. When combined with the student effects, this produces an $\widehat{\boldsymbol{\eta}}$ vector with 3446 elements. Only 0.08% of the components of the 6236×3446 $\boldsymbol{S}$ matrix are non-zero.

Because the future year teacher effects were all so highly correlated, we follow the suggestion of McCaffrey and Lockwood (2011) and average the future year effects for each teacher. For example, a fourth grade teacher's year-5 and year-6 effects are averaged to obtain a single estimated "future year" effect. Table 5.6 presents the results of applying the MAR GP VAM to the grade-school data. Estimates from the joint model containing teacher effects, student effects, and both teacher and student effects in the missing data mechanism appear in Tables 5.7, 5.8, and 5.9, respectively.

Table 5.6: MAR Model for Grade-School Data

Parameter	Estimate	Std. Error
Grade 4 mean score	501.2974	1.8201
Grade 5 mean score	520.9812	1.7517
Grade 6 mean score	541.6364	1.9685
σ_1^2	591.6821	31.7546
σ_2^2	416.6918	25.9969
σ_3^2	646.3503	32.3316
Student variance	1632.3315	54.4985
Grade 4: (1,1)	402.5404	70.9214
Grade 4: (2,1)	255.0205	54.7948
Grade 4: (3,1)	257.7022	57.2814
Grade 4: (2,2)	180.5257	51.5718
Grade 4: (3,2)	189.9864	49.0041
Grade 4: (3,3)	208.1485	59.7621
Grade 5: (1,1)	178.9935	34.9350
Grade 5: (2,1)	74.7577	22.4170
Grade 5: (2,2)	38.4672	20.1876
Grade 6: (1,1)	184.6173	37.9895

Table 5.7: Joint Model for Grade-School Data: Teacher Effects in Missing Data Mechanism

Parameter	Estimate	Std. Error
Grade 4 mean score	501.3168	1.8409
Grade 5 mean score	520.9576	1.7505
Grade 6 mean score	541.5879	1.9713
Grade 5 mean completion	0.6599	0.0319
Grade 6 mean completion	1.0343	0.0419
σ_1^2	591.4300	31.8124
σ_2^2	417.1465	26.0258
σ_3^2	646.7562	32.3652
Student variance	1631.3971	54.5238
Grade 4: (1,1)	401.0872	70.8279
Grade 4: (2,1)	243.6031	53.5346
Grade 4: (3,1)	248.4397	56.4094
Grade 4: (4,1)	2.1269	0.7616
Grade 4: (2,2)	165.3230	49.1631
Grade 4: (3,2)	176.5910	47.1283
Grade 4: (4,2)	1.3999	0.5969
Grade 4: (3,3)	195.7792	58.0564
Grade 4: (4,3)	1.5102	0.6301
Grade 4: (4,4)	0.0176	0.0154
Grade 5: (1,1)	183.2161	35.7010
Grade 5: (2,1)	77.1346	22.7912
Grade 5: (3,1)	1.2807	0.6721
Grade 5: (2,2)	38.8062	20.2178
Grade 5: (3,2)	0.7033	0.5072
Grade 5: (3,3)	0.0499	0.0258
Grade 6: (1,1)	186.9716	38.4489

Table 5.8: Joint Model for Grade-School Data: Student Effects in Missing Data Mechanism

Parameter	Estimate	Std. Error
Grade 4 mean score	505.4803	1.9627
Grade 5 mean score	523.9156	1.8476
Grade 6 mean score	544.7444	2.0936
Grade 4 mean completion	0.8886	0.0332
Grade 5 mean completion	0.6457	0.0314
Grade 6 mean completion	0.7139	0.0317
σ_1^2	568.4458	30.6736
σ_2^2	424.6381	26.1150
σ_3^2	659.1867	32.8372
Student variance (1,1)	1652.9003	55.0909
Student variance (1,2)	0.4062	0.0416
Student variance (2,2)	11.4092	1.0200
Grade 4: (1,1)	401.9636	74.4837
Grade 4: (2,1)	256.2325	58.0930
Grade 4: (3,1)	268.0598	63.4438
Grade 4: (2,2)	184.9096	54.2676
Grade 4: (3,2)	201.8847	54.3640
Grade 4: (3,3)	229.4734	68.1920
Grade 5: (1,1)	179.4790	35.1303
Grade 5: (2,1)	74.5944	22.7574
Grade 5: (2,2)	37.8510	20.6074
Grade 6: (1,1)	186.4737	38.3054

Table 5.9: Joint Model for Grade-School Data: Both Student and Teacher Effects in Missing Data Mechanism

Parameter	Estimate	Std. Error
Grade 4 mean score	505.0338	2.0427
Grade 5 mean score	522.1291	2.4057
Grade 6 mean score	545.0489	2.3237
Grade 4 mean completion	0.8601	0.0336
Grade 5 mean completion	0.6285	0.0343
Grade 6 mean completion	0.5449	0.0681
σ_1^2	569.3392	30.7202
σ_2^2	424.3269	26.0705
σ_3^2	660.5157	32.8708
Student variance (1,1)	1643.5646	54.7924
Student variance (1,2)	0.3194	0.0482
Student variance (2,2)	10.6332	1.0073
Grade 4: (1,1)	414.7861	75.6862
Grade 4: (2,1)	263.9356	59.2370
Grade 4: (3,1)	277.4264	64.5767
Grade 4: (4,1)	0.9300	0.8443
Grade 4: (2,2)	188.4893	55.7737
Grade 4: (3,2)	206.9267	55.0179
Grade 4: (4,2)	0.5480	0.6798
Grade 4: (3,3)	232.6256	69.3750
Grade 4: (4,3)	0.5683	0.7430
Grade 4: (4,4)	0.0048	0.0190
Grade 5: (1,1)	181.8541	35.5557
Grade 5: (2,1)	74.4134	22.4873
Grade 5: (3,1)	1.1783	1.2934
Grade 5: (2,2)	38.5864	20.2441
Grade 5: (3,2)	-0.4298	0.9402
Grade 5: (3,3)	0.2119	0.0846
Grade 6: (1,1)	190.1575	39.1850

Figures 5.1, 5.2, 5.3, 5.4, and 5.5 plot the estimated current and future year teacher effects for each grade from the MAR model and from MNAR-t. The smallest correlation among these plots is .9891. Figures 5.6, 5.7, 5.8, 5.9, and 5.10 plot the estimated current and future year teacher effects for each grade from the MAR model and from MNAR-s. The smallest correlation between the estimated teacher effects from the MAR and this joint model is .9917. Figures 5.11, 5.12, 5.13, 5.14, and 5.15 plot the estimated current and future year teacher effects for each grade from the MAR model and from MNAR-b. The smallest correlation between the estimated teacher effects from the MAR and this joint model is .9791.

Despite the change in parameter estimates between the models, the teacher effects do not change noticeably. The correlations between the estimated effects from the MAR and joint models are all .9791 or greater. This finding is consistent with what McCaffrey and Lockwood (2011) find using selection and pattern mixture models, modeling missingness as a function of students only, when studying a different data set. Their models yield correlations between 0.97 and 1 for all teacher effects between the joint models and the MAR model, using a variable persistence structure for the teacher effects. Aaronson et al. (2007) rank teachers by the quartile of the relevant effect that their individual estimate falls in. None of the teacher rankings move more than a single quartile between the MAR and joint models. The stability of the teacher rankings between the models MAR, MNAR-t, MNAR-s, and MNAR-b indicates that the rankings from the GP VAM for this data set are not sensitive to the presence of potentially nonignorable missing data. Of course, this does not provide sufficient evidence for concluding that the dropout process is ignorable, but it does provide a stronger argument than fitting the MAR model alone.

Figure 5.1: Joint v. MAR for Grade-School 4th Grade Current Year Effects with Teachers included in Missing Data Model

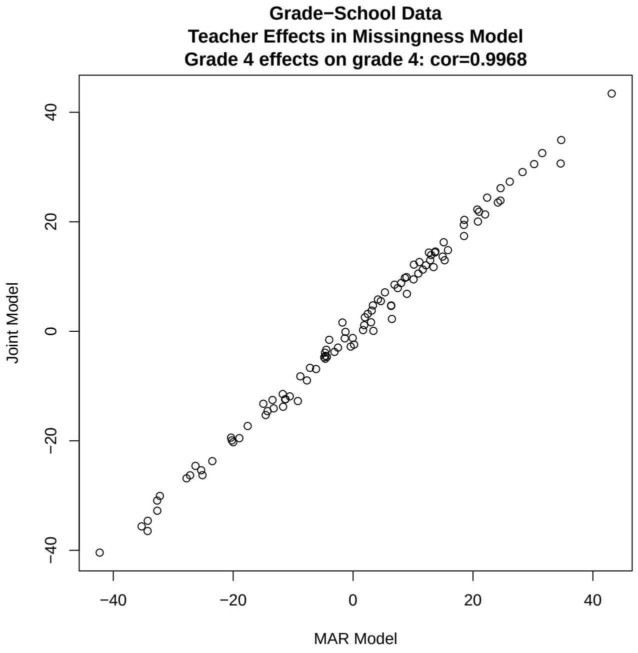


Figure 5.2: Joint v. MAR for Grade-School 4th Grade Future Year Effects with Teachers included in Missing Data Model

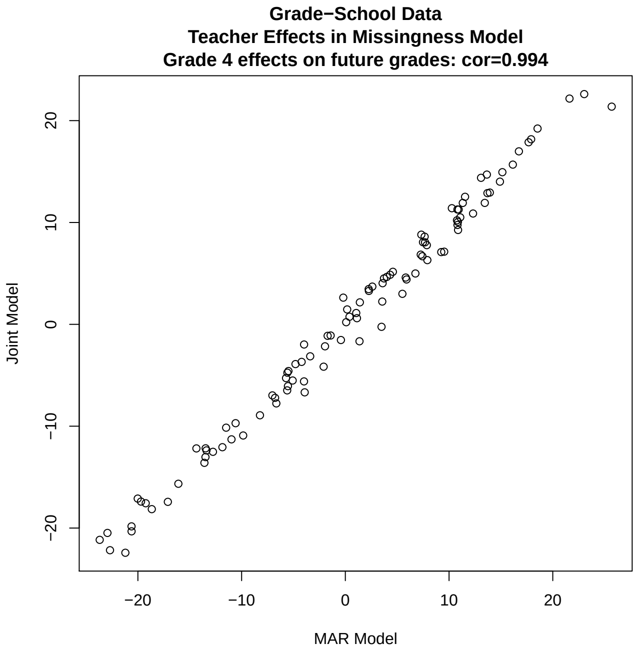


Figure 5.3: Joint v. MAR for Grade-School 5th Grade Current Year Effects with Teachers included in Missing Data Model

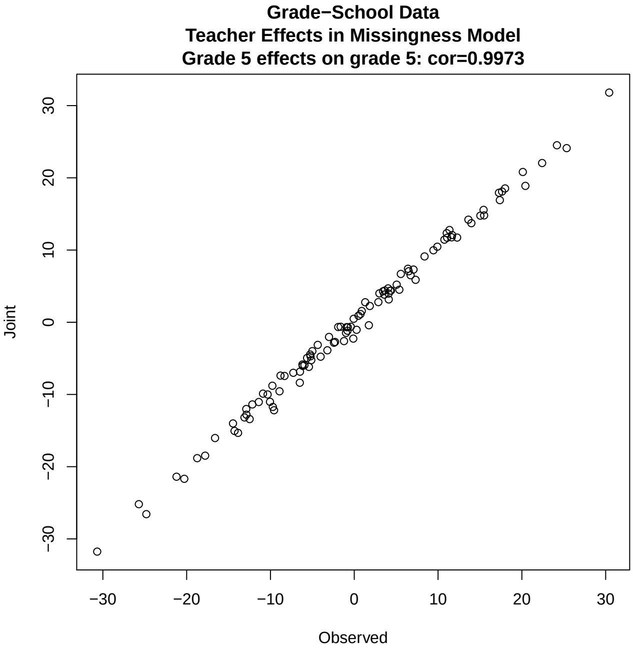


Figure 5.4: Joint v. MAR for Grade-School 5th Grade Future Year Effects with Teachers included in Missing Data Model

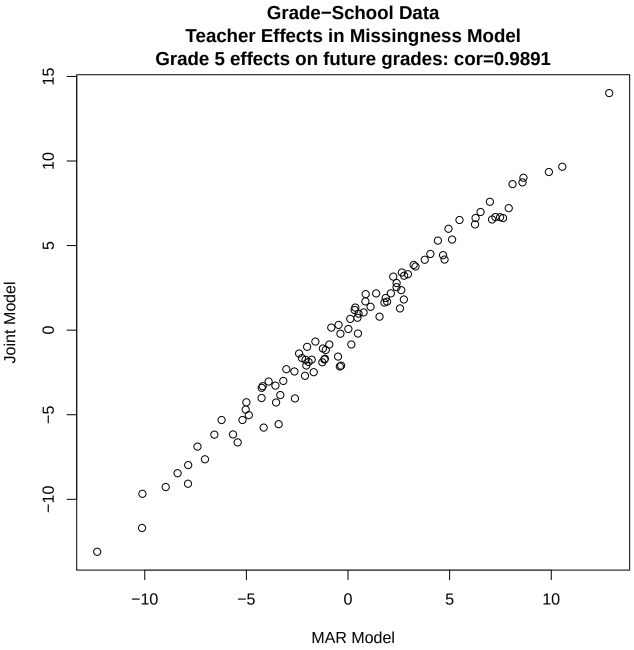


Figure 5.5: Joint v. MAR for Grade-School 6th Grade Current Year Effects with Teachers included in Missing Data Model

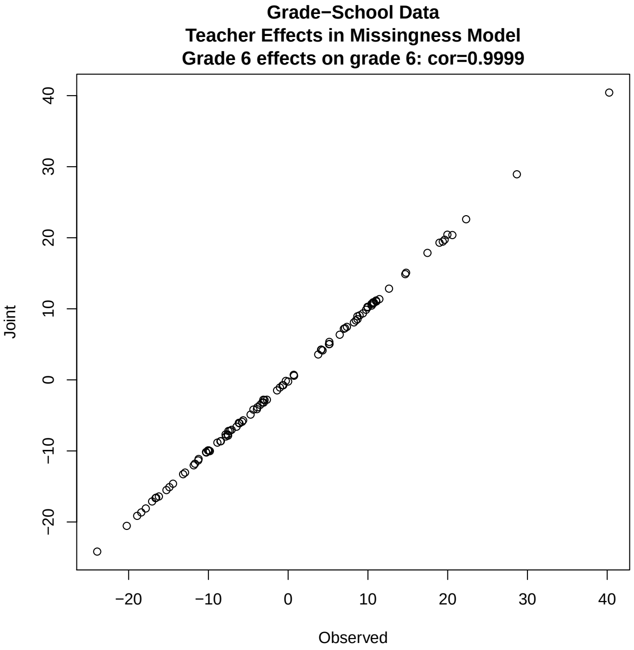


Figure 5.6: Joint v. MAR for Grade-School 4th Grade Current Year Effects with Students included in Missing Data Model

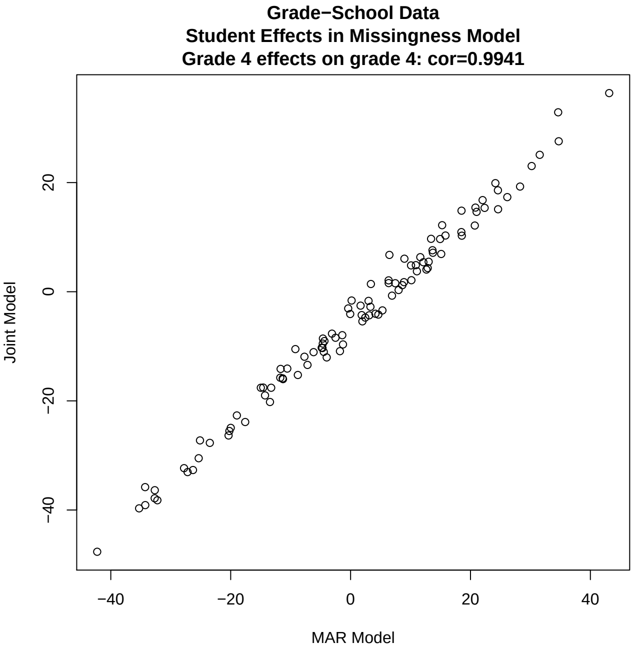


Figure 5.7: Joint v. MAR for Grade-School 4th Grade Future Year Effects with Students included in Missing Data Model

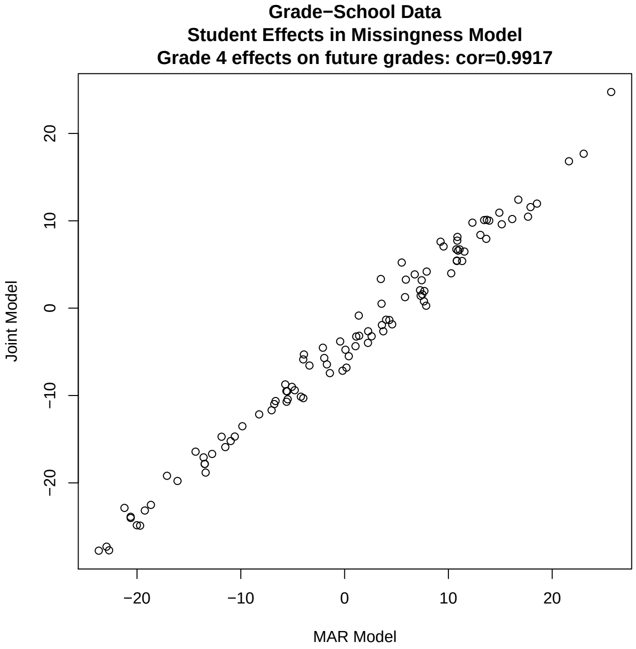


Figure 5.8: Joint v. MAR for Grade-School 5th Grade Current Year Effects with Students included in Missing Data Model

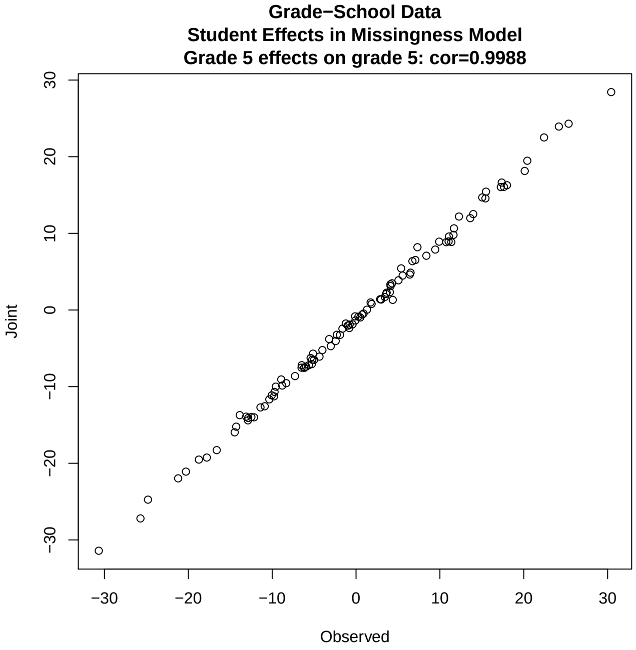


Figure 5.9: Joint v. MAR for Grade-School 5th Grade Future Year Effects with Students included in Missing Data Model

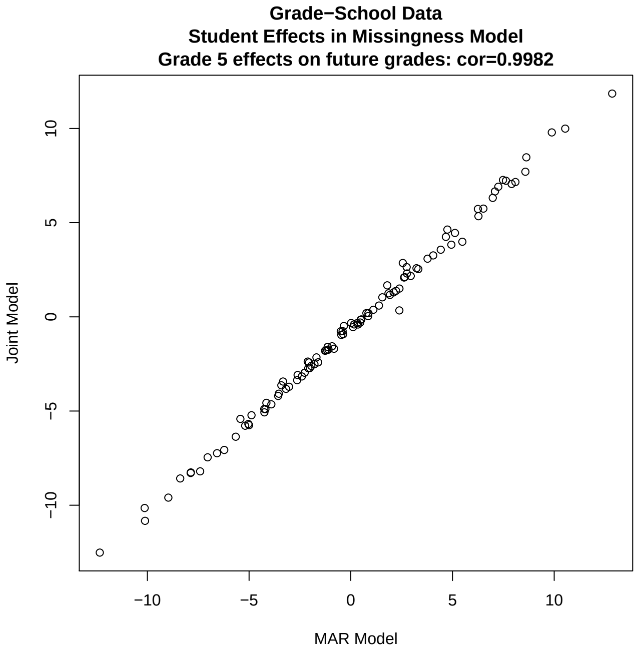


Figure 5.10: Joint v. MAR for Grade-School 6th Grade Current Year Effects with Students included in Missing Data Model

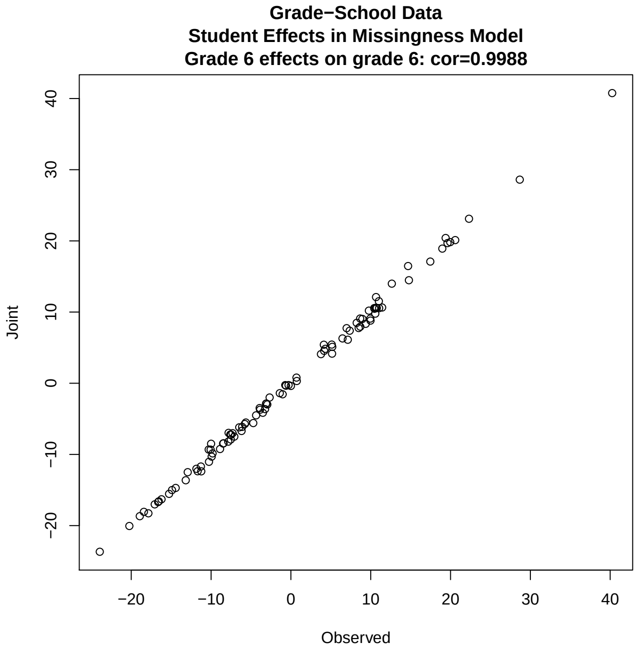


Figure 5.11: Joint v. MAR for Grade-School 4th Grade Current Year Effects with Teachers and Students included in Missing Data Model

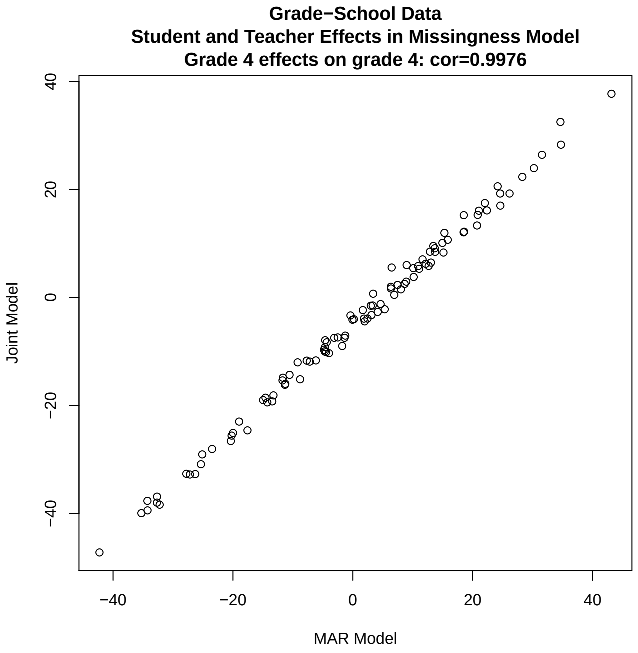


Figure 5.12: Joint v. MAR for Grade-School 4th Grade Future Year Effects with Teachers and Students included in Missing Data Model

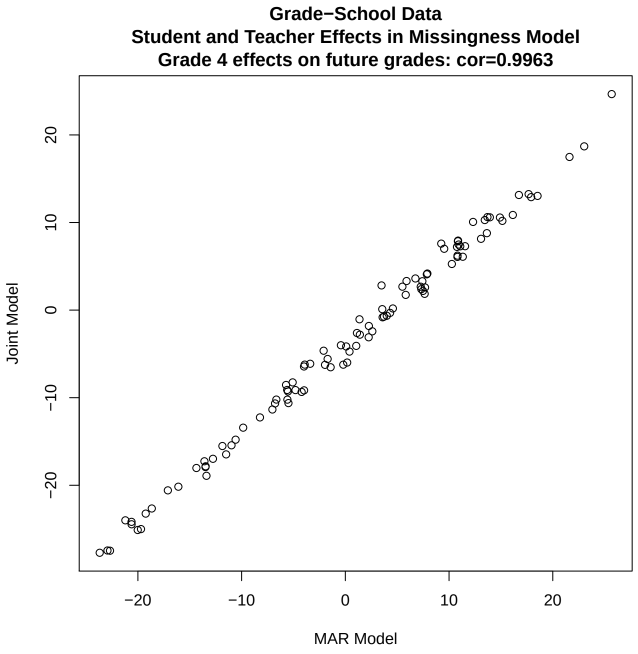


Figure 5.13: Joint v. MAR for Grade-School 5th Grade Current Year Effects with Teachers and Students included in Missing Data Model

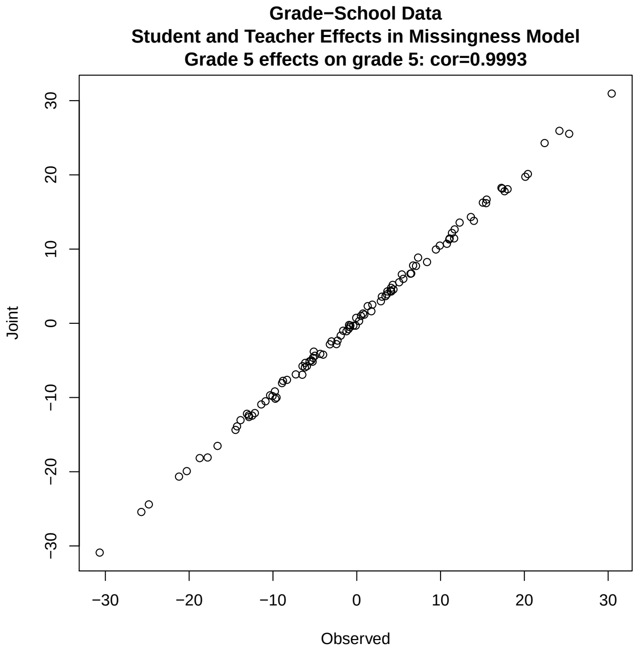


Figure 5.14: Joint v. MAR for Grade-School 5th Grade Future Year Effects with Teachers and Students included in Missing Data Model

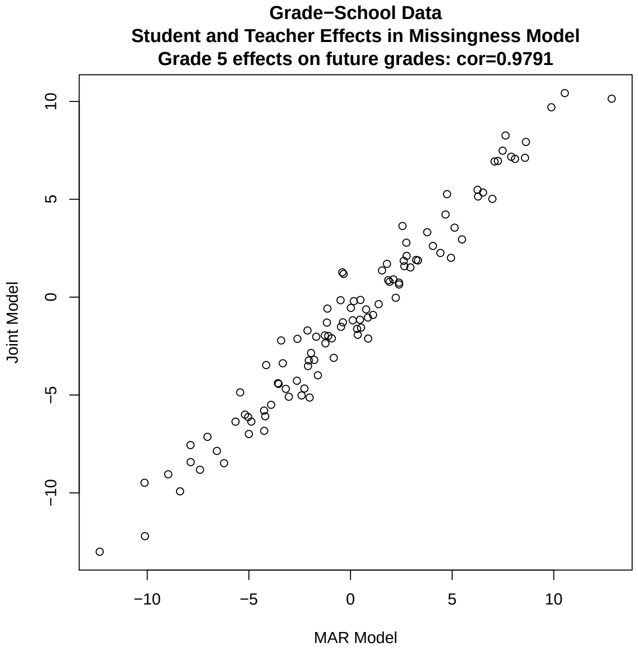


Figure 5.15: Joint v. MAR for Grade-School 6th Grade Current Year Effects with Teachers and Students included in Missing Data Model

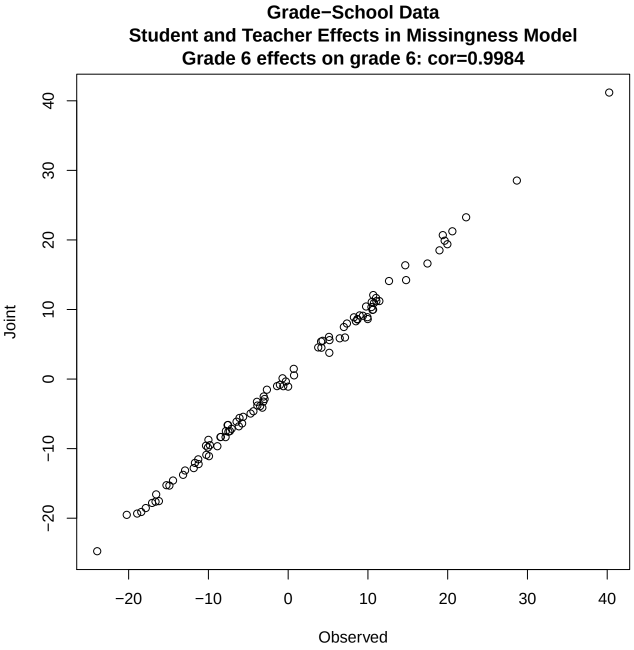
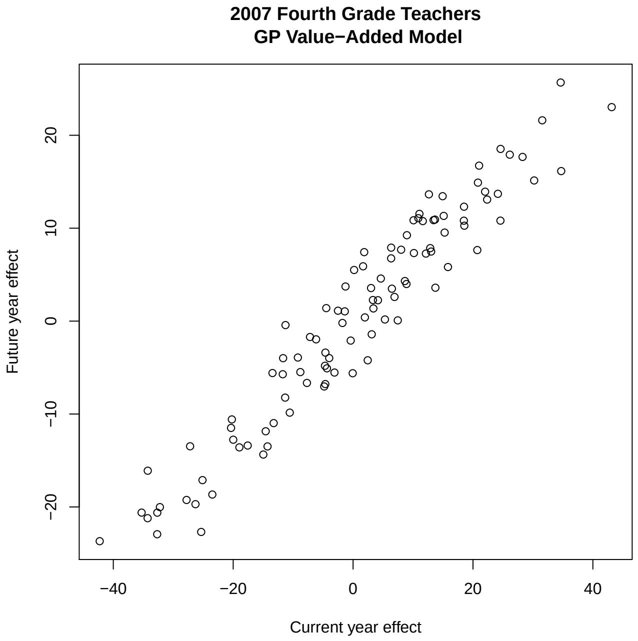


Figure 5.16: Fourth Grade Effects from GP model



It is also useful to examine the bivariate plot of current and future year teacher effects for each grade, even though the estimated model parameters indicate that they are positively correlated. Figures 5.16 and 5.17 plot the current and future year effects for fourth and fifth grade teachers, respectively, from the GP model. Figures 5.18 and 5.19 plot the corresponding effects from the joint model, where in this case the joint model included both student and teacher effects in the missing data mechanism. The positive correlations between these effects indicate that teachers whose classes perform well tend to graduate students who go on to perform well in future years. We reiterate that the interpretation of these effects as “teacher effects,” as they are being interpreted by state education departments, relies on the randomization of students to classrooms: the VAM attempts to account for the nonrandom assignment through the inclusion of random student intercepts. The warnings from Draper (1995) about the importance of randomization in multilevel models are still relevant.

Figure 5.17: Fifth Grade Effects from GP model

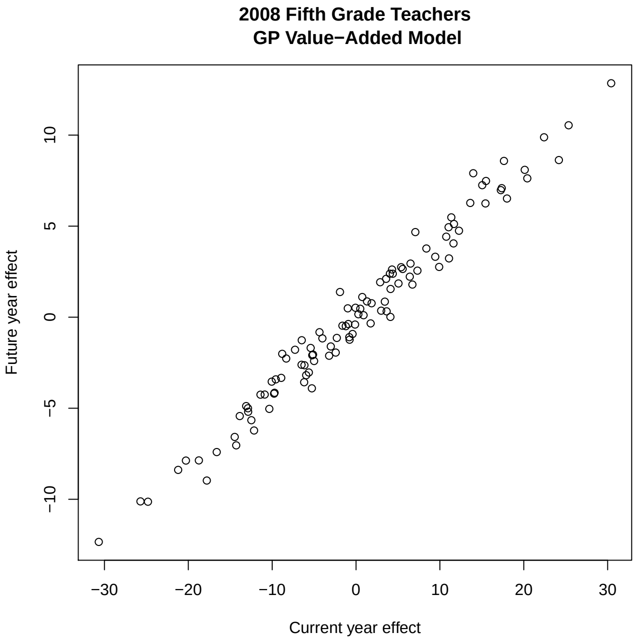


Figure 5.18: Fourth Grade Effects from Joint model

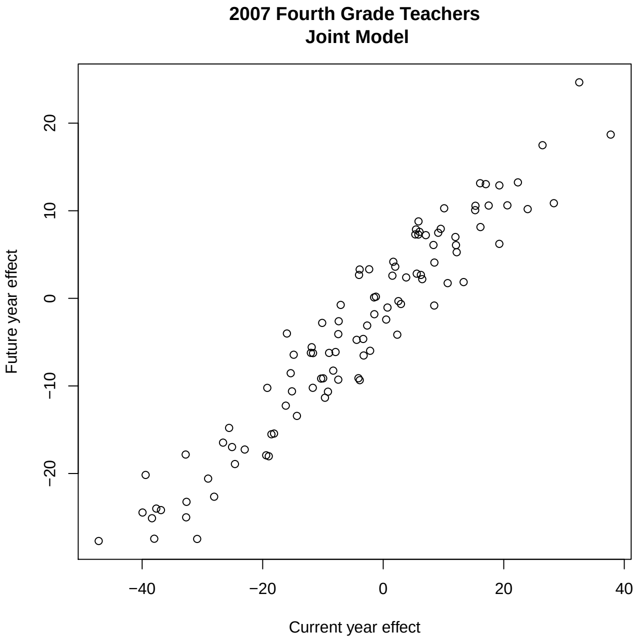
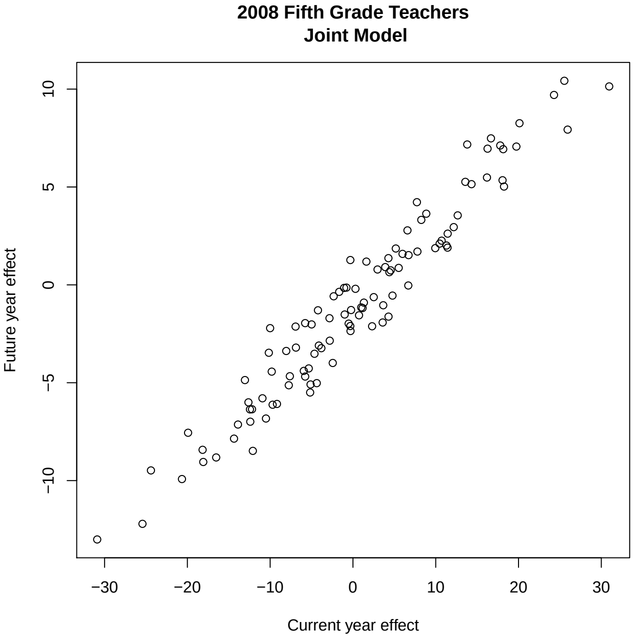


Figure 5.19: Fifth Grade Effects from Joint model



5.5 Effects of Missing Data in Calculus Classes

This section applies the joint model to data on calculus grades from a large public university. Broatch and Lohr (2011) used a subset of these data in their analyses. The data set tracks 3561 students who took calculus 2 and possibly calculus 3 at the university. A total of 184 calculus 2 classes are included from Fall 2000 through Spring 2005. In addition, 144 calculus 3 classes from Spring of 2001 through Spring of 2006 are included. Students who took only calculus 3 during the study are omitted. Analysis focuses on the grades assigned to students, which are converted to the corresponding value on a four-point scale. The scores in the data set are collectively centered and standardized.

This calculus example provides us with an ideal setting for testing for the presence of informative missing data. While not every student who takes calculus 2 does so with the intention of taking calculus 3, we may expect to see, on average, a certain proportion of calculus 2 students going on to complete calculus 3. In this example, we construct the missing data mechanism to measure the proportion of students from calculus 2 classes who complete calculus 3. To perform a sensitivity analysis, we fit an MAR model and compare its parameter estimates and estimated teacher effects to three different nonignorable models.

We fit the GP model both singly (assuming missing data are ignorable) and jointly with a missing data mechanism that includes random teacher effects, though in the second case under an assumption of MAR which is enforced by setting the correlation between the random effects in the two sub-models to be zero. In the model we will call MNAR-t, we include a random teacher effect in the missing data mechanism that is correlated with the corresponding teacher effects from the observed data mechanism and measures the proportion of each teacher's students who go on to complete calculus 3. The model MNAR-s models missingness as a function of student random effects. Even though only one binary observation is made on each student, we are able to fit this model because the predicted student effects in the missing data mechanism borrow strength from their correlation with the student effects from the observed data mechanism. Finally, MNAR-b contains both random student and teacher effects in the missing data mechanism. The appropriate missing data process cannot be chosen by empirical investigation of the observed data (including examination of the log-likelihood) since the observed data do not provide information to support one particular MNAR model over another (Fitzmaurice et al., 2004; Xu and Blozis, 2011). Instead, we compare the parameter estimates for the observed data mechanism across the different models, looking for sensitivity to the assumptions about the nature of the missing data.

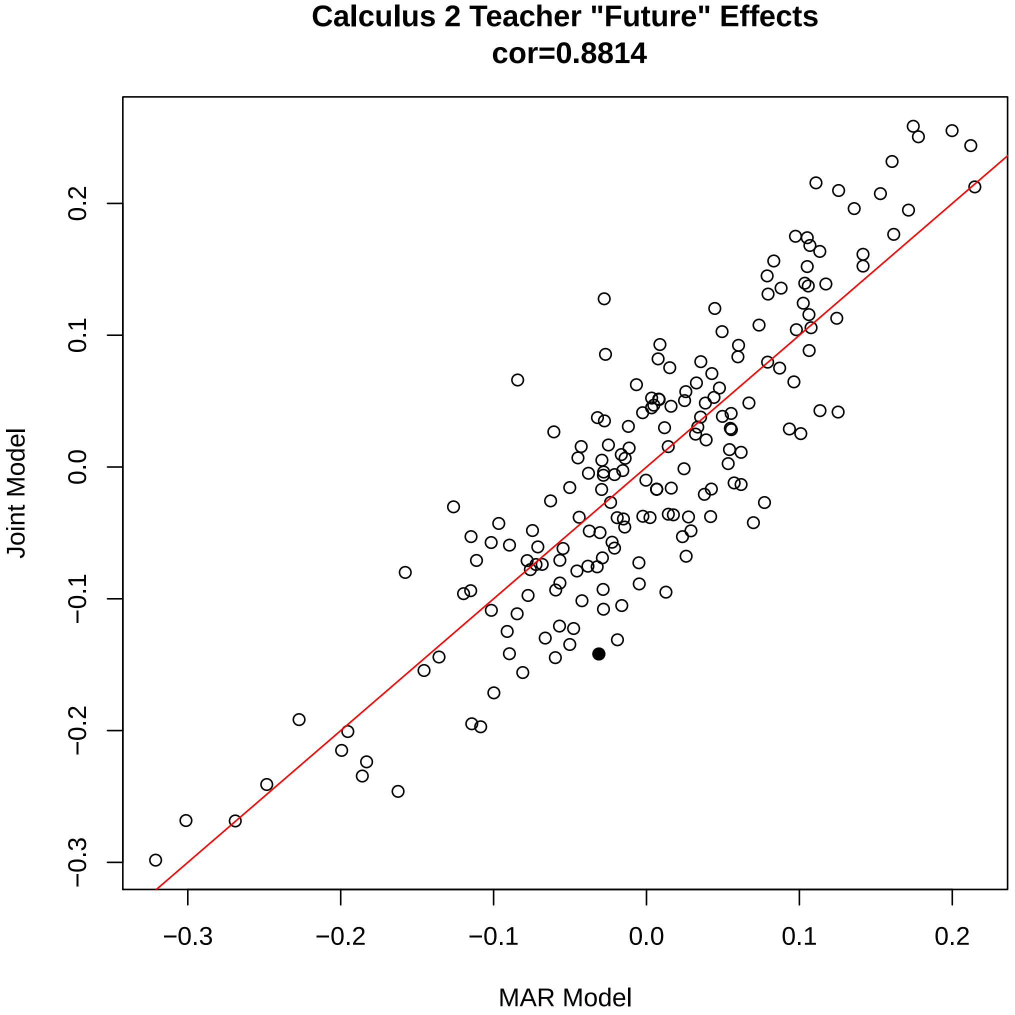
The parameter estimates from the models appear in Table 5.10. The yearly means in the observed data model are represented by μ_i^y, for $i = 1, 2$. The value μ_2^r gives the estimated proportion, e.g. $\Phi(0.2459) = 0.5971$, of Calculus 2 students who complete Calculus 3. The other parameters follow the same notation as used in Chapter 4. Also listed for each model are -2 times the Laplace approximated log-likelihood ($-2l$) and the correlation (ρ) of the predicted calculus 2 future year effects with those from the MAR model. Because the student scores come from non-standardized class grades, the current year teacher effects reflect the tendency of individual teachers to assign above- or below-average grades, and not necessarily the effectiveness of their teaching. The future year effects of calculus 2 teachers, however, reflect how well each teacher's former students performed in comparison to their new calculus 3 classmates. The correlation (ρ) of these effects in the MNAR models to the MAR model provides a summary of the sensitivity of the teacher rankings to nonignorable dropout under different models for the missing data mechanism. Using selection and pattern mixture models to attribute MNAR data to students, McCaffrey and Lockwood (2011) found values of ρ that were all greater than 0.97. MNAR-s provides the analog of their models using correlated random effects, and yields $\rho = .994$. Likewise, MNAR-b does not produce teacher effects that are substantially different from the MAR model. However, MNAR-t reorders the teacher effects, producing $\rho = .881$. Aaronson et al. (2007) rank teachers by the quartile of the relevant effect that their individual estimate falls in. Analyzing the calculus data with MNAR-t leads to different classifications than those produced by MAR model. Thus, a teacher may receive a different evaluation based on the model assumed (either tacitly or explicitly) for the missing data mechanism. Using the method of Aaronson et al. (2007), some teachers move two (or even three) quartiles when evaluated with MNAR-t. Figure 5.20 plots the calculus 2 future year teacher effects from MNAR-t against the future effects from the MAR model.

While computing estimates for MNAR-t, we are able to calculate an entire EM iteration with a fully exponential correction for $\widetilde{\boldsymbol{\eta}}$ in just under 2.5 minutes. Once this algorithm converges, we include the corrections to $\widetilde{\boldsymbol{v}}$, thereby increasing the iteration time to around 30 hours. The full algorithm then requires only a few further iterations to converge. The differences are, for practical purposes, negligible between the estimates obtained from including all of the fully exponential corrections and the estimates obtained by including the corrections to $\widetilde{\boldsymbol{\eta}}$ only. As such, when estimating the parameters of the models MNAR-s and MNAR-b in Section 5.3, we include the corrections from Equation (4.16) only, since the dimension of $\widetilde{\boldsymbol{v}}$ in these examples is greater than 8000×8000.

Following the suggestion of Molenberghs et al. (2008), we examine the fit of MNAR-t to MAR to see which teachers make the missing data mechanism appear to be MNAR. The large amount of missing data in certain calculus classrooms means that the effects of those classrooms are shrunk toward zero due to the shrinkage properties of EBLUPs. This shrinkage property is normally desirable in VAMs, but in the case of nonignorable dropout, we lose information. For illustration, we examine the records of one of the teachers most greatly down-weighted by MNAR-t, indicated by a solid black circle in Figure 5.20. Only 20% of the students from this classroom completed calculus 3 (most of them failed the calculus 2 course), and those that did all received below-average grades in their respective calculus 3 classrooms. The calculus 2 teacher's effect on calculus 3 in the MAR model is less than 0, but is severely shrunk because only a few observations are present. This is representative of the types of teachers that seem to be most affected in the joint model MNAR-t: their effects are decreased. Also affected are the teachers who have the greatest proportion of students completing calculus 3. Their scores are increased. By contrast, the aforementioned calculus 2 teacher has the future year effect increased when student effects are included in the missing data mechanism. The change in teacher rankings produced by MNAR-t does not appear in MNAR-b or MNAR-s because these models seem to attribute the missingness of students with low grades to the students.

For convenience, the correlation matrix for the effects of calculus 2 teachers from MNAR-t appears in Equation (5.1). The last column of these matrices, "3 comp.", yields information about the correlation of the completion effect of the calculus 2 teachers. A larger completion effect means that relatively more of a teacher's students go on to complete calculus 3. This effect is positively correlated with both the "2 on 2" effect, indicating relatively how high a calculus 2 teacher's average assigned grade in a class, and with the "2 on 3" effect, indicating relatively how well the calculus 2 teacher's former students did in calculus 3. However, the current

Figure 5.20: Calculus 2 Future Year Effects: MAR vs. MNAR-t



and future year effects for calculus 2 teachers are not correlated. Observing that a teacher gives above- or below-average grades yields no information about how well the students of that teacher performed in calculus 3. Applications of VAMs to standardized test score data usually show a positive correlation between the current and future teacher effects (Mariano et al., 2010).

Table 5.10: Sensitivity analysis for university data. Standard errors are in parentheses.

	Ignorable	MAR	MNAR-t	MNAR-s	MNAR-b
μ_1^y	-0.095 (0.027)	-0.095 (0.028)	-0.097 (0.028)	-0.092 (0.027)	-0.094 (0.028)
μ_2^y	-0.154 (0.034)	-0.154 (0.035)	-0.161 (0.035)	-0.282 (0.035)	-0.284 (0.035)
μ_2^r		0.256 (0.026)	0.246 (0.026)	0.307 (0.065)	0.304 (0.041)
σ_1^2	0.388 (0.023)	0.388 (0.023)	0.385 (0.023)	0.328 (0.020)	0.330 (0.020)
σ_2^2	0.292 (0.019)	0.292 (0.019)	0.293 (0.019)	0.330 (0.019)	0.329 (0.019)
$\mathbf{\Gamma}_{stu}[1,1]$	0.618 (0.026)	0.618 (0.026)	0.620 (0.026)	0.680 (0.026)	0.674 (0.025)
$\mathbf{\Gamma}_{stu}[2,1]$				0.637 (0.128)	0.640 (0.065)
$\mathbf{\Gamma}_{stu}[2,2]$				0.600 (0.633)	0.610 (0.261)
$\mathbf{\Gamma}_1[1,1]$	0.082 (0.015)	0.082 (0.015)	0.085 (0.015)	0.077 (0.013)	0.082 (0.015)
$\mathbf{\Gamma}_1[2,1]$	-0.004 (0.009)	-0.004 (0.010)	-0.001 (0.010)	-0.006 (0.009)	-0.002 (0.010)
$\mathbf{\Gamma}_1[3,1]$			0.043 (0.011)		0.017 (0.013)
$\mathbf{\Gamma}_1[2,2]$	0.028 (0.011)	0.028 (0.011)	0.031 (0.011)	0.028 (0.010)	0.030 (0.011)
$\mathbf{\Gamma}_1[3,2]$			0.021 (0.009)		0.010 (0.012)
$\mathbf{\Gamma}_1[3,3]$		0.034 (0.013)	0.040 (0.014)		0.052 (0.022)
Γ_2	0.080 (0.015)	0.080 (0.015)	0.082 (0.015)	0.082 (0.015)	0.082 (0.015)
$-2l$		20047.9	20022.7	19447.6	19436.7
ρ	1	1	0.881	.994	.984

$$cor(\mathbf{\Gamma}_1) = \begin{array}{c} \\ \text{2 on 2} \\ \text{2 on 3} \\ \text{3 comp.} \end{array} \begin{array}{c} \begin{array}{ccc} \text{2 on 2} & \text{2 on 3} & \text{3 comp.} \end{array} \\ \begin{pmatrix} 1 & -0.03 & 0.74 \\ -0.03 & 1 & 0.60 \\ 0.74 & 0.60 & 1 \end{pmatrix} \end{array} \tag{5.1}$$

The sensitivity analysis illustrates the influence that assumptions about the nature of missing data may have on the resulting teacher rankings. A challenge with MNAR models is that their fit may not be tested empirically, since data are missing. Thus the choice between MNAR-b, MNAR-t, or MAR depends in part on a subject-matter decision. As shown by our sensitivity analysis, that decision has direct implications for the estimated teacher rankings.

5.6 Discussion

The analyses of the urban school district and university data sets both result in correlated longitudinal and missingness mechanisms. However, only the university data shows a substantial change to the teacher rankings between the models. We believe this behavior may be due in part to the amount of missing data in each example. In the university data set, there are some calculus 2 classrooms from which only 20% of the students complete calculus 3. No classroom in the grade-school data has less than 70% observed data. The large amount of missing data in certain university classrooms means that those classroom effects are shrunk to zero due to the shrinkage properties of EBLUPs. This shrinkage property is normally desirable in VAMs, but in the case of informative dropout, this property may produce biased estimates for teacher effects.

The insensitivity of the teacher rankings from the elementary school data set to assumptions about the missing data mechanism is consistent with the findings of McCaffrey and Lockwood (2011). Besides the amount of missing data from each class, two other important differences between the elementary and the college data sets are that 1) the scores from the elementary setting come from standardized exams and 2) college students have much greater latitude in selecting their future courses. Ideally, the type sensitivity analysis we ran for each of our data sets should be performed whenever value-added models are used to evaluate teachers. In situations like our college example where the sensitivity analysis leads to substantially shuffled rankings for the teachers, the VAM should not be used for high-stakes decisions. Otherwise, some teachers will be unjustly punished or rewarded due to a choice of unverifiable modeling assumptions. Recalling the quote from Molenberghs and Kenward (2007), "ignoring MNAR models is no different an option than shifting to one particular MNAR model, it is just much more convenient."

Chapter 6

CONCLUSION

In this dissertation, we have extended the capabilities of value-added models for use in educational and other applications. We first devised a method for efficient computation of maximum like estimates of the most flexible VAM in use, the GP model of Mariano et al. (2010). We then constructed a new model that allows examination of possible effects of missing data on VAM scores.

We have developed an efficient and stable EM algorithm to obtain maximum likelihood estimates (MLEs) of the generalized persistence (GP) (Mariano et al., 2010) VAM. Although the model may be specified in software such as SAS, the multi-membership structure produces several large matrices which must be manipulated and inverted, and the random effects representing the future year effects of teachers are often highly correlated. For even medium-sized data sets, SAS runs out of memory, runs so slow as to be impractical, or fails when its Newton algorithm steps out of the parameter space. By contrast, our package GPvam obtains the MLEs relatively quickly and reliably. Although the computational methods and software for GPvam are developed in the educational setting, they can be used in many other applications as well, substituting the level-1 units for “students” and the level-2 units for “teachers”. Similar models have been proposed for studying contributions of physicians toward patients’ health outcomes. The multi-membership structure also arises in social network data (Airoldi et al., 2008). In another example, Browne et al. (2001) and Goldstein et al. (2000) describe a multi-membership model used to study Belgian household migration with complete persistence, measuring the propensity of individuals to change household membership. The GP model may be a good candidate for the Belgian household data since the similarity of former roommates may decrease over time.

We have extended the EM algorithm and computational methods implemented in GPvam to accommodate the estimation of a new model which we have presented for jointly modeling nonignorable missing data with the student scores of a value-added model. In addition to the challenges faced during the estimation of the GP model, the joint model faces a high-dimensional integration problem due to the non-linear functions introduced to the log-likelihood by the binary attendance indicators. The joint model provides flexibility in the specification of the missing data mechanism, allowing the attendance indicators to be modeled as a function of a combination of random student and teacher effects. These effects are allowed to be correlated with their counterparts in the VAM, producing a means of modeling nonignorable dropout. The flexibility in the random effects structure for the missing data mechanism furnishes the capacity for performing a sensitivity analysis.

When applied to standardized math scores from an urban grade-school district, none of the three formulations of the joint model (including random student, teacher, and both student and teacher effects in the missing data mechanism) produce substantially different teacher rankings from the GP model (which assumes missing data are ignorable). However, when applied to calculus 2 and 3 grades from a large university, the joint model with random teacher effects in the missing data mechanism produces rankings for the future-year effect of calculus 2 teachers that have a correlation of only .88 with those from the GP model. Some of individual teacher rankings moved two (or even three) quartiles between the two models. The difference in these rankings are a consequence of modeling the data under two unverifiable assumptions about the missing data (that they are ignorable and that they are nonignorable according to a certain parametric structure). The nonignorable missing observations violate the missing at random assumption of the GP VAM. This sensitivity is an important finding, because of the potentially high-stakes applications of the teacher rankings.

Similar to the results of our application to the grade-school data, McCaffrey and Lockwood (2011) did not find an appreciable difference in the results of their ignorable and nonignorable models. However, McCaffrey and Lockwood (2011) analyzed data from elementary school standardized scores, attributing the missingness to student, but not teacher, characteristics. Two important differences between the calculus and the elementary school data are the lack of standardization in the calculus grades and the greater potential for the calculus attendance trajectories of students to vary by teacher, due to the greater choice college students have in selecting future courses. These factors may help explain the more profound changes to calculus teacher rankings resulting from the joint model MNAR-t. The results of our application suggest that at the university level, and other settings in which there is more discretion for course progression, considering missing data explicitly as a function of teacher effects can result in different teacher rankings. If the university in the application were to use the data for personnel decisions, some teachers would receive different evaluations based on the modeling assumptions made about dropout.

Our correlated parameter model provides a different perspective in the joint, missing-data analysis of teacher effects over the work of McCaffrey and Lockwood (2011) by allowing the missingness of test scores to depend on teacher history. The missing data mechanism could be further refined by distinguishing between different types of missing data. For example, suppose that in the college setting students are being tracked across calculus 2 and 3. Some students will have missing calculus 3 scores because they did not take the class. Others, however, will be missing scores because they enrolled in the class and later dropped out. Information on this partial enrollment is available from college data in the form of a ‘W’ on the student’s transcript. By contrast, part-year enrollment information may not be available at the grade-school level, or may not indicate if the student was enrolled for a week or nearly the entire year. Thus, future work will consider more specific dropout mechanisms for different educational settings.

A major concern in the application of VAMs is the sensitivity of the teacher rankings to the choice of fixed effects that are included in the model, as well as sensitivity to potential measurement error in the covariates. For example, the presence of covariate measurement error tends to bias the teacher estimates of the standardized gain model (Reback, 2008) and the student growth percentile model (Betebenner, 2009), with teachers of minority and impoverished students being more likely to be rated as ineffective (Wright, 2010). Simpler VAMs based on “mean gain” are subject to increased estimation error with measurement error on covariates (McCaffrey et al., 2009). Besides measurement error in the covariates, Briggs and Domingue (2011) found significant differences in teacher rankings in the data set analyzed by the Los Angeles Times (Felch et al., 2010) depending on the covariates that were included in the model. We recommend further investigation of the impact of omitted or incorrectly measured fixed effect covariates on VAM teacher rankings.

Another avenue for future work involves developing robust estimation methods for value-added models. Outlier-robust regression methods may be useful for studying the impact of outliers on the teacher rankings. Another option would be to use a nonparametric mixed model, or semiparametric methods which improve the fit of incorrectly specified parametric models with a certain amount of nonparametric fit (Waterman et al., 2006). Employing these different methods would allow for an analysis of the sensitivity of teacher rankings to the assumptions of the potentially misspecified linear model. This work could be combined with the development of influence diagnostics for the GP model to evaluate the fit of the linear mixed model based GP VAM.

The methods developed in this dissertation can be generalized or modified for applications to other non-linear mixed models with multi-membership or non-nested designs. For example, non-linear mixed models are popular in the pharmaceutics industry in the study of pharmacokinetics. It should be possible to adapt the methods we have developed to compute the estimates of these types of models, possibly yielding greater flexibility in the level of complexity of the models that may be fit. It is also interesting to explore the comparative performance of the Laplace approximation, the fully-exponential Laplace approximation, and, when possible, penalized quasi-likelihood methods in SAS software.

In another application, it should be possible to use multi-membership generalized linear mixed models in the ranking of sports teams when only win/loss information may be used, such as with the BCS college football rankings. There would be several advantages, both computational and theoretical, of treating teams as random effects instead of fixed effects. Treating teams as fixed effects leads to difficulties with complete separation (Allison, 2008) when the data contain teams with perfect records: this is not an issue when the teams are treated as random effects. However, treating teams as random effects leads to a multi-membership design matrix for the random effects, producing an intractable integral in the likelihood of dimension equal to the number of teams in the data. Work in this direction has been limited by computational requirements which are met by the methods we have developed.

Value-added models are widely acknowledged to be imperfect instruments for measuring teacher effectiveness (Koretz, 2008; Braun et al., 2010). They depend on the quality of the test as a measuring instrument and are sensitive to model assumptions and unmeasured covariates. When students are not randomly assigned to teachers, effects ascribed to teachers may actually be more properly attributed to the students who take those teachers. However, these difficulties are also faced by other measures of teacher performance. Despite their shortcomings, VAMs have great potential to provide information that can be used to improve the educational system (Harris and McCaffrey, 2010).

Value-added models have been implemented for teacher evaluation by more than 20 states as a result of Race to the Top and other educational initiatives. Many of the properties of different VAMs, however, have not been well studied to date, in particular the effects of past teachers and the potential effects of missing data. The work in this dissertation provides methods for computing persistence of teacher effects and proportions of variability coming from different parts of the educational system. This information can be used to target resources, suggest designed experiments, and improve quality of education. The ability to model mechanisms for missing data and explore their effects on parameter estimates and VAM scores gives researchers a valuable new tool for understanding influences in education.

REFERENCES

Aaronson, D., Barrow, L., and Sander, W. (2007), “Teachers and Student Achievement in the Chicago Public High Schools,” *Journal of Labor Economics*, 25, 95–135.

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken: John Wiley & Sons, 2nd ed.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), “Mixed Membership Stochastic Blockmodels,” *Journal of Machine Learning Research*, 9, 1981–2014.

Allison, P. D. (2008), “Convergence Failures in Logistic Regression,” SAS Global Forum 2008, no. 360-2008, www2.sas.com/proceedings/forum2008/360-2008.pdf.

Ballou, D., Sanders, W., and Wright, P. (2004), “Controlling for Student Background in Value-Added Assesment of Teachers,” *Journal of Educational and Behavioral Statistics*, 29, 37–65.

Bates, D. and Maechler, M. (2011), *Matrix: Sparse and Dense Matrix Classes and Methods*, R package version 1.0-1.

Betebenner, D. W. (2009), “Norm- and Criterion-Referenced Student Growth,” *Educational Measurement: Issues and Practices*, 28, 42–51.

Braun, H. (2005), “Value-Added Modeling: What Does Due Dilligence Require?” in *Value Added Models in Education: Theory and Applications*, ed. Lissitz, R., Maple Grove, MN: JAM Press, pp. 19–38.

Braun, H. I., Chudowsky, N., and Koenig, J. (2010), *Getting Value Out of Value-Added*, Washington, DC: National Academies Press.

Breslow, N. E. and Lin, X. (1995), “Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion,” *Biometrika*, 82, 81–91.

Briggs, D. and Domingue, B. (2011), “Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the Los Angeles Times.” *National Education Policy Center*, www.nepc.colorado.edu/publication/due–diligence.

Broatch, J. and Lohr, S. (2011), “Multidimensional Assesment of Value Added by Teachers to Real-World Outcomes,” *Journal of Educational and Behavioral Statistics*, in press.

Browne, W. J., Goldstein, H., and Rasbash, J. (2001), "Multiple Membership Multiple Classification (MMMC) Models," *Statistical Modelling*, 1, 103–124.

Casella, G. and Berger, R. L. (2001), *Statistical Inference*, Pacific Grove: Duxbury Press, 2nd ed.

de Leeuw, J. and Meijer, E. (2008), *The Handbook of Multilevel Analysis*, New York: Springer.

Demidenko, E. (2004), *Mixed Models Theory and Applications*, Hoboken: Wiley-Interscience.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Doran, H. C. and Lockwood, J. R. (2006), "Fitting Value-Added Models in R," *Journal of Educational and Behavioral Statistics*, 31, 205–230.

Draper, D. (1995), "Inference and Hierarchical Modeling in the Social Sciences," *Journal of Educational and Behavioral Statistics*, 20, 115–117.

Evans, M. and Swartz, T. (1995), "Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems," *Statistical Science*, 10, 254–272.

— (2000), *Approximating Integrals via Monte Carlo and Deterministic Methods*, New York: Oxford University Press.

Felch, J., Song, J., and Smith, D. (2010), "Who's teaching L.A.'s kids?" http://articles.latimes.com/2010/aug/14/local/la-me-teachers-value-20100815.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004), *Applied Longitudinal Analysis*, Hoboken: John Wiley & Sons.

Follmann, D. and Wu, M. (1995), "An Approximate Generalized Linear Model With Random Effects For Informative Missing Data," *Biometrics*, 51, 151–168.

Goldstein, H., Rasbash, J., Browne, W., Woodhouse, G., and Poulain, M. (2000), "Multilevel Models in the Study of Dynamic Household Structures," *European Journal of Population*, 16, 373–387.

Harris, D. N. and McCaffrey, D. F. (2010), "Value-Added: Assessing Teachers' Contributions to Student Achievement," in *Teacher Assessment and the Quest for Teacher Quality*, ed. Kennedy, M. M., San Francisco: Jossey-Bass, pp. 251–282.

Harville, D. A. (2008), *Matrix Algebra from a Statistician's Perspective*, New York: Springer.

Henderson, C. R. (1950), “The Estimation of Genetic Parameters,” *The Annals of Mathematical Statistics*, 21, 309–310.

— (1975), “Best Linear Unbiased Estimation and Prediction under a Selection Model,” *Biometrics*, 31, 423.

Jamshidian, M. and Jennrich, R. I. (2000), “Standard Errors for EM Estimation,” *Journal of the Royal Statistical Society, Series B*, 62, 257–270.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, New York: Springer.

Kass, R. E. and Steffey, D. (1989), “Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Emperical Bayes Models),” *Journal of the American Statistical Association*, 84, 717–726.

Koretz, D. (2008), “A Measured Approach,” *American Educator*, Fall, 18–39.

Lehmann, E. L. and Romano, J. (2010), *Testing Statistical Hypotheses*, New York: Springer, 3rd ed.

Leonhardt, D. (2010), “When Does Holding Teachers Accountable Go Too Far?” www.nytimes.com/2010/09/05/magazine/05FOB-wwln-t.html.

Lin, H., Liu, D., and Zhou, X.-H. (2009), “A Correlated Random-Effects Model for Normal Longitudinal Data with Nonignorable Missingness,” *Statistics in Medicine*, 29, 236–247.

Lin, X. and Breslow, N. E. (1996), “Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion,” *Journal of the American Statistical Association*, 91, 1007–1016.

Lindfield, G. and Penny, J. E. T. (1988), *Microcomputers in Numerical Analysis*, New York: Halsted Press.

Lindstrom, M. J. and Bates, D. M. (1988), “Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data,” *Journal of the American Statistical Association*, 83, 1014–1022.

— (1990), “Nonlinear Mixed Effects Models for Repeated Measures,” *Biometrics*, 46, 673–687.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006), *SAS for Mixed Models*, Cary, NC: SAS Institute, Inc., 2nd ed.

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, New York: John Wiley, 2nd ed.

Lockwood, J., McCaffrey, D., Mariano, L., and Setodji, C. (2007), “Bayesian Methods for Scalable Multivariate Value-Added Assesment,” *Journal of Educational and Behavioral Statistics*, 32, 125–150.

Lockwood, J. R., Doran, H., and McCaffrey, D. F. (2003), “Using R for Estimating Longitudinal Student Achievement Models,” *The Newsletter of the R Project*, 3, 17–23.

Louis, T. A. (1982), “Finding the Observed Information Matrix when Using the EM Algorithm,” *Journal of the Royal Statistical Society Series B*, 44, 226–233.

Louisiana Department of Education (2010), “Louisiana Adopts Value-Added Teacher Evaluation Model,” Press Release, http://www.doe.state.la.us/offices/publicaffairs/press_release.aspx?PR=1428.

Magnus, J. R. and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Hoboken: John Wiley & Sons Ltd, revised ed.

Mariano, L. T., McCaffrey, D. F., and Lockwood, J. (2010), “A Model for Teacher Effects from Longitudinal Data Without Assuming Vertical Scaling,” *Journal of Educational and Behavioral Statistics*, 35, 253–279.

McCaffrey, D., Lockwood, J., Mariano, L. T., and Setodji, C. (2005), “Challenges for Value-Added Assesment of Teacher Effects,” in *Value Added Models in Education: Theory and Applications*, ed. Lissitz, R., Maple Grove, MN: JAM Press, pp. 111–144.

McCaffrey, D., Lockwood, J. R., Koretz, D., Louis, T., and Hamilton, L. (2004), “Models for Value-Added Modeling of Teacher Effects,” *Journal of Educational and Behavioral Statistics*, 29, 67–101.

McCaffrey, D. F., Han, B., and Lockwood, J. R. (2009), “Turning Student Test Scores into Teacher Compensations Systems,” in *Performance Incentives: Their Growing Impact on American K-12 Education*, ed. Springer, M. G., Washington, DC: Brookings Institution Press.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., and Hamilton, L. S. (2003), *Evaluating Value-Added Models for Teacher Accountability*, Pittsburgh: The RAND Corporation.

McCaffrey, D. F. and Lockwood, J. R. (2011), “Missing Data in Value-Added Modeling of Teacher Effects,” *Annals of Applied Statistics*, 5, 773–797.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall, 2nd ed.

McCulloch, C. E. (1994), “Maximum Likelihood Variance Components Estimation for Binary Data,” *Journal of the American Statistical Association*, 89, 330–335.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, Hoboken: John Wiley & Sons, 2nd ed.

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, Hoboken: John Wiley & Sons, 2nd ed.

McLean, R. A., Sanders, W. L., and Stroup, W. W. (1991), “A Unified Approach to Mixed Linear Models,” *The American Statistician*, 45, 54–64.

Molenberghs, G. and Kenward, M. G. (2007), *Missing Data in Clinical Studies*, West Sussex: John Wiley.

Molenberghs, G., Kenward, M. G., Verbeke, G., Beunckens, C., and Sotto, C. (2008), “Every Missingness Not at Random Model Has A Missingness at Random Counterpart with Equal Fit,” *Journal of the Royal Statistical Society, Series B*, 70, 371–388.

Nocedal, J. and Wright, S. J. (1999), *Numerical Optimization*, New York: Springer.

Petersen, K. B. and Pedersen, M. S. (2008), “The Matrix Cookbook,” Version 20081110.

Pinheiro, J. C. and Bates, D. M. (1995), “Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model,” *Journal of Computational and Graphical Statistics*, 4, 12–35.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raudenbush, S. and Bryk, A. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Thosand Oaks, CA: Sage, 2nd ed.

Reback, R. (2008), “Teaching to the Rating: School Accountability and the Distribution of Student Achievement,” *Journal of Public Economics*, 92, 1394–1415.

Rizopoulos, D. (2010), “JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data,” *Journal of Statistical Software*, 35, 1–33.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009), “Fully Exponential Laplace Approximations for the Joint Modelling of Survival and Longitudinal Data,” *Journal of the Royal Statistical Society, Series B*, 71, 637–654.

Rowan, B., Correnti, R., and Miller, R. J. (2002), "What Large-Scale, Survey Research Tells Us About Teacher Effects on Student Achievement: Insights From the Prospects Study of Elementary schools," *Teachers College Record*, 104, 1525–1567.

Ryan, R. M. and Weinstein, N. (2009), "Undermining Quality Teaching and Learning: A Self-Determination Theory Perspective on High-Stakes Testing," *Theory and Research in Education*, 7, 224–233.

Sanders, W., Saxton, A., and Horn, B. (1997), "Grading teachers, grading schools. Is student achievement a valid evaluation measure?" Thousand Oaks, CA: Corwin Press, Inc, chap. The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assesment., pp. 137–162.

SAS Institute Inc. (2011), *SAS 9.2 Help and Documentation*, Cary, NC: SAS Institute, Inc.

Shun, Z. (1997), "Another Look at the Salamander Mating Data: A Modified Laplace Approximation Approach," *Journal of the American Statistical Association*, 92, 341–349.

Shun, Z. and McCullagh, P. (1995), "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society, Series B*, 57, 749–760.

Steele, B. M. (1996), "A Modified EM Algorithm for Estimation in Generalized Mixed Models," *Biometrics*, 52, 1295–1310.

The National Academies (2009), "Letter Report to the U.S. Department of Education on the Race to the Top Fund," www.nap.edu/catalog/12780.html.

Tierney, L. and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716.

U.S. Department of Education (2009), "Race to the Top Program Executive Summary," www2.ed.gov/programs/racetothetop/executive-summary.pdf.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Vonesh, E. F. (1996), "A Note on the Use of Laplace's Approximation for Nonlinear Mixed-Effects Models," *Biometrika*, 83, 447–452.

Vonesh, E. F., Greene, T., and Schluchter, M. D. (2006), “Shared Parameter Models for the Joint Analysis of Longitudinal Data and Event Times,” *Statistics in Medicine*, 25, 143–163.

Waterman, M. J., Birch, J. B., and Schabenberger, O. (2006), “Linear Mixed Model Robust Regression,” Tech. rep., http://www.web-e.stat.vt.edu/dept/web-e/tech_reports/TechReport07-3.pdf.

Wolfinger, R. and O’Connell, M. (1993), “Generalized Linear Mixed Models: A Pseduo-likelihood Approach,” *Journal of Statistical Computation and Simulation*, 48, 233–243.

Wolfinger, R., Tobias, R., and Sall, J. (1994), “Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models,” *SIAM Journal of Scientific Computing*, 15:6, 1294–1310.

Wolfinger, R. D. and Lin, X. (1997), “Two Taylor-series Approximation Methods for Nonlinear Mixed Models,” *Computational Statistics and Data Analysis*, 25, 465–490.

Wright, S. P. (2004), “Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assesment Without Imputation,” Presented at the 2004 National Evaluation Institute, July 8-10, 2004, Colorado Springs, CO, www.sas.com/resources/asset/educational–value–added–assessment.pdf.

— (2010), “An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education,” Tech. rep., http://www.sas.com/resources/whitepaper/wp_16975.pdf.

Wright, S. P., White, J. T., and Sanders, W. L. (2010), *SAS EVAAS Statistical Models*, Cary, NC: SAS Institute, www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf.

Wu, C. (1983), “On the Convergence Properties of the EM Algorithm,” *Annals of Statistics*, 11, 95–103.

Wu, M. C. and Carroll, R. J. (1988), “Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process,” *Biometrics*, 44, 175–188.

Xu, S. and Blozis, S. A. (2011), “Sensitivity Analysis of Mixed Models for Incomplete Longitudinal Data,” *Journal of Educational and Behavioral Statistics*, 36, 237–256.

Yuan, Y. and Little, R. J. A. (2009), “Mixed-Effect Hybrid Models for Longitudinal Data with Nonignorable Dropout,” *Biometrics*, 65, 478–486.