Bayesian Latent Variable Methods for Longitudinal Processes with Applications to Fetal Growth

James Christopher Slaughter

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, School of Public Health.

Chapel Hill
2007

Approved by:

Amy H. Herring
Katherine E. Hartmann
Lawrence L. Kupper
Chirayath Suchindran
Haibo Zhou

ABSTRACT

James Christopher Slaughter: Bayesian Latent Variable Methods for Longitudinal Processes with Applications to Fetal Growth
(Under the direction of Dr. Amy H Herring)

We consider methods for joint models of exposure and response in epidemiologic studies. In particular, we show how latent variable methods provide a structure for obtaining inference about multistate growth processes and multiple longitudinal and cross-sectional outcomes. Each model utilizes underlying, subject-specific latent variables to account for the correlation that arises from taking multiple observations on the same sampling unit. We also consider latent variable mixture models in order to more flexibly model the latent variable distributions and identify latent classes of subjects who are of particular scientific importance. We apply our methods to applications in reproductive health, obtaining interesting new insights while developing and applying statistical methodology.

We first consider the problem of estimating a multistate growth process with unknown initiation time to determine individual early fetal growth. Using cross-sectional data, we identify fetuses that have a latent tendency to grow relatively quickly and slowly and show that slow growth early in pregnancy is associated with an increased risk of future pregnancy loss. These results are important to researchers who use early ultrasounds to date pregnancies under the assumption that there is no measurable variability in early fetal growth.

Paper two is concerned with jointly modeling the unusual, asymmetric distributions of birth weight and gestational age. Using latent variable mixture models, we identify a latent class of subjects who are more likely to deliver early and have low weight. We also allow observed covariates to be associated with latent class membership. Our approach provides researchers a new method for examining low birth weight and pre-term birth.

In paper three, we aggregate multiple ultrasound measurements on fetal size and blood restriction using latent variables that follow mixture distributions to identify a latent class of subjects who are growth restricted during pregnancy. We then consider a joint model that examines the associations between covariates, early growth restriction, and outcomes measured at birth. Our methods are able to identify a latent class of subjects who have increased blood flow restriction and below average intrauterine size during the second trimester who are more likely to be growth restricted at birth.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. Amy Herring, for her constant support and guidance in writing this dissertation, and for being an excellent mentor in helping me grow as a researcher and biostatistician. I appreciated the insightful comments and suggestions of my committee members, Dr. Kathie Hartmann, Dr. Larry Kupper, and Dr. Haibo Zhou, and Dr. Chirayath Suchindran, and their willingness to offer both instruction and guidance throughout my education.

Finally, I would like to thank my family, friends, and fellow colleagues who have stood beside me as I have pursued this endeavor. Their support has made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	Abdominal circumference
ARS	Adaptive rejection sampling
BPD	Biparietal diameter
BMI	Body mass index
CDF	Cumulative distribution function
CI	Credible interval
CRL	Crown-rump length
EM	Expectation-Maximization
FL	Femur length
GLM	Generalized linear model
GLMM	Generalized linear mixed model
HC	Head circumference
LMP	Last menstrual period
MCMC	Markov chain Monte Carlo
MSD	Mean (gestational) sac diameter
PI	Pulsatility index
PIN	Pregnancy, Infection and Nutrition
PTB	Pre-term birth
RFTS	Right from the Start
RI	Resistance index

LIST OF ABBREVIATIONS

S/D	Systolic-diastolic (ratio)
SAB	Spontaneous abortion
SEM	Structural equation model
SGA	Small for gestational age

1 Introduction

Longitudinal data, in which repeated measurements are taken on the same subject over time, require special statistical methods because observations on the same subject tend to be correlated. A variety of approaches have been considered to account for the correlation (Diggle et al. 1994) with the random effects model being one of the most popular methods. In random effects models, observations on the same subject are assumed to be correlated due to the effect of some unobservable variables, called the random effects. The random effects are a type of latent variable, and random effects models are a type of latent variable method. We will consider several other types of models in which latent variables are used to account for correlation in longitudinal processes.

Latent variables can be broadly classified into two concepts, latent predictor variables and latent response variables. Latent predictors (exogenous latent variables) are determined outside of the model and can be related to mis-measured covariates. The daily average of particulate matter measurement from several monitoring stations is an example of error-prone realization of a latent true ambient particulate matter concentration predictor variable. Latent responses, also known as endogenous latent variables, are determined within the model. Latent responses include underlying variables that can only be measured indirectly through multiple items and are often useful for reducing the dimensionality of the data. For example, depression is a concept that cannot be measured directly, but can be indirectly estimated using a battery of questions. While an isolated response on the survey may not be very useful, latent variable methods provide a natural way of aggregating multiple responses to determine an individual's latent relative depression level. In our reproductive health examples, we quantify the latent tendency of an individual fetus to grow relatively quickly or slowly using information from a first and second trimester ultrasound and his or her age and weight at delivery.

Latent variables are commonly found in structural equation models (SEMs) used in the social sciences as a means of quantifying an unobservable concept based on several observed variables (Joreskog 1970). SEMs consist of a latent variable model and a measurement model. The latent variable model describes the association between latent predictor variables and latent response variables. The second part of the model, the measurement model, gives the relationship of measured outcomes and predictors with latent outcomes and latent predictors.

Structural equations models provide a general modeling framework that can be used for modeling several types of correlated data (Sanchez et al. 2005). Most work in this area considers SEMs as a model for multivariate normal data, but they can be thought of in a broader context. Ordinal data are directly incorporated into normal-theory models by using threshold models. Threshold models are based on the idea that, for example, the binary observed variable y takes on value one if some underlying continuous variable, y^*, is above a threshold value and zero otherwise. The latent y^* then replaces y in the SEM (Muthen 1983, 1984). Authors including Sammel et al. (1997) and Dunson (2000) propose additional extensions that allow for observed outcomes to be in the exponential family as in the generalized linear model.

Latent variables can also be used in the modeling of the incidence and progression of disease. Many diseases can be thought of as moving through several latent health states with the exact transition times between states unknown. Dunson and Baird (2002) describe the progression of a chronic condition, uterine fibroids, as progressing from (1) no disease, (2) preclinical disease, to (3) clinical disease. Information on the current state as well as several indicators of disease progress can be used to estimate a severity latent variable. This type of model can provide useful inference on the incidence and progression of disease using cross-sectional data.

1.1 Reproductive Health Applications

Reproductive and perinatal health spans a broad area of research in maternal and infant health including infertility, miscarriage, pregnancy complications, birth weight, pre-term delivery, birth defects, childhood cancer, and child development (Bracken 1984; Kiely 1991). It has been long recognized that events preceding and during pregnancy can influence the health and well being of both the mother and her child. Researchers have a dual responsibility to contribute to the practical knowledge that improves health as well as developing methodology so that we are better equipped to understand future health problems.

1.1.1 Fetal Growth

During pregnancy, fetal development is traditionally dated from the first day of a woman's last reported menstrual period (LMP). The literature intermittently refers to this method of dating as menstrual age and obstetricians often use the term gestational age interchangeably with menstrual age. However, structures related to growth should more accurately begin growing at some later time point, which we will refer to as conception. Conceptional age is also known as fetal age; for consistency, I will only use the terms gestational age and conceptional age.

In clinical practice, the conceptional age is rarely known so it is estimated based on the assumption of a midcycle ovulation (conceptional age = gestational age - 14 days). Estimates of conceptional age are inaccurate due to variability in the follicular phase distribution, which can vary both among women and within the same woman between different cycles (Zhou 2006). This variability can be particularly troubling when studying early fetal development because new structures appear every few days. Several studies, such as the Early Pregnancy Study (Wilcox et al. 1988), have been conducted to precisely date the time from LMP to clinical pregnancy. Using urinary biomarkers, Wilcox et al. (2001) estimated that the probability of conception on a given day of the menstrual cycle, conditional on reaching that day of the cycle was greatest on day 13 (Figure 1). The probability is less than 2% for each day before day eight and each day after day 21 of the cycle. Furthermore, the probability of clinical pregnancy on a given day can be significantly modified by covariates. Wilcox et al. (2001) found the distribution of the probability of clinical pregnancy by cycle day was more variable with a larger mode for women with irregular compared to regular cycles.

In our first paper, we are interested in modeling early fetal development as a growth process with an unknown initiation time. Conception, the unknown initiation time, is known to occur after the LMP so we estimate the time from LMP to conception. Individual growth rates are estimated using a latent growth variable that allows individuals to move between developmental states relatively quickly and slowly. In paper two, we focus on growth measured at birth using birth weight and gestational age. In that analysis, we attempt to identify factors that are related to an individual's underlying intrauterine growth rate and a tendency to be born earlier than average. In paper three, we identify growth restriction using multiple measurements of fetal size and blood flow restriction collected at two time points during the second trimester. We then examine the association between early growth restriction and growth restriction measured at birth. Papers one and two are motivated by the Right From the Start (RFTS) study of early pregnancy (Promislow et al. 2004), and paper three uses information from the Pregnancy, Infection, and Nutrition (PIN) study (Savitz et al. 1999).

1.1.2 Fetal Development during Pregnancy

Human pregnancies are divided into three trimesters, each normally lasting approximately 12-14 weeks. In paper one, we are concerned with development within the first trimester, specifically in the embryonic period which begins with fertilization and lasts for eight weeks. During the embryonic period, the embryo proceeds through several important developmental stages. These stages can be defined by the presence or absence of key features including the gestational sac, yolk sac, fetal pole, and cardiac activity. Each of these features may be observed by a first trimester ultrasound depending on the developmental progress of the pregnancy.

The gestational sac is the first structure visualized by sonography and can usually be seen by the fifth gestational week. The average internal diameter of the gestational sac is calculated as the mean of the anteroposterior diameter, the transverse diameter, and the longitudinal diameter. This measurement, known as the mean sac diameter

(MSD), provides a useful early estimator of age in a normal pregnancy. According to Filly and Hadlock (2000), the gestational sac can be observed when it reaches 2 to 3 mm MSD which occurs around 5 gestational weeks. MSDs up to 14 mm are very precise for predicating gestational age in normal pregnancies, but become less reliable as pregnancy progresses. According to Filly and Hadlock (2000), the predicted gestational age when the MSD is 14 mm is 6.5 weeks (95%CI: [6.0, 7.0]). The yolk sac is usually observed inside the gestational sac during the fifth gestational week. However, the dimensions of the yolk sac do not significantly improve the ability to predict gestational age, so we do not consider it in our analysis (Filly and Hadlock 2000).

After the gestational sac and yolk sac, the fetal pole is the next important structure able to be detected by ultrasound (Figure 2). At this time, the crown-rump length (CRL; also known as the fetal pole length) becomes the measurement of choice for estimating gestational age. The fetal pole, without normal cardiac activity, can be visualized when the CRL is as small as 2 mm, which occurs at 5.7 gestational weeks (95% CI: [5.2, 6.2]). Normal cardiac activity begins a few days later, by the sixth gestational week (Filly and Hadlock 2000). Hadlock et al. (1992) evaluated the association between CRL and gestational age in 416 women with good menstrual dating. They used a fourth order linear regression model in which CRL was able to predict 98.6% of the variation in the natural logarithm of gestational age. Their results are similar to predictions reported by other authors who often used only linear or quadratic CRL effects. An advantage of the Hadlock study is that it included a relatively large sample with a greater range of CRL measurements (from 2 mm to 120 mm) than other studies. The additional small crown-rump length measurements are particularly useful to our analysis.

Researchers have conducted studies to determine if gestational sac diameters or crown rump lengths measured by an early ultrasound are indicative of early pregnancy loss. Nyberg et al. (1987) collected data on 83 women who had two sonograms performed during the first trimester of pregnancy. The subjects were referred for a second ultrasound due to bleeding or pelvic pain (64 cases) or to confirm the pregnancy (19 cases). They found that gestational sac growth was significantly slower in women who eventually had abnormal pregnancies compared to women with normal pregnancies. However, in evaluating 254 viable singleton pregnancies, Brizot et al. (2001) were not able to confirm this result. Mantoni and Pedersen (1982) examined 67 patients with threatened abortion with regular, known menstrual cycles. They found that the crown rump length of these fetuses was smaller than expected based on their gestational age. There was also some evidence that the fetuses that eventually aborted had on average smaller crown rump lengths. Again, Brizot et al. (2001) could not replicate this result. Rather than consider gestational sac diameters and fetal pole lengths directly, we conceptualize that they are indicators of underlying fetal growth. In paper one, we estimate an early fetal growth latent variable and examine its association with the risk of pregnancy loss.

1.1.3 Birth Weight

Birth weight is an important predictor of perinatal, neonatal, and postnatal outcomes (Shan and Ohlsson 2002). Poor growth during the intrauterine period is associated with increased risk of perinatal and infant morbidity and mortality. Many epidemiological studies have been conducted to elucidate some of the determinants of birth weight. Factors including, but not limited to, race/ethnicity, maternal age, parity, previous history of low birth weight, and alcohol use are generally accepted as having a strong association. Other studies have found various environmental and occupational exposures, caffeine use, uterine fibroids, and stress may be associated with growth and require further research (Shan and Ohlsson 2002).

Zhang and Bowes (1995) attempted to create a standard for identifying small for gestational age (SGA) infants stratified on race, gender, and parity. Gestational age was calculated using LMP data for more than 95% of all subjects, but in some cases a clinician-based adjustment was needed to correct inaccurate LMPs. The authors do not report statistical test results indicating if they found significantly different growth curves by race, gender, or parity. However, they do provide graphs that indicate these covariates may be associated with growth as measured by birth weight between 25 and 42 weeks gestation. We will examine the association of race, gender, and parity with growth rate during the first trimester.

Weight at birth is a function of both the amount of time from conception to birth and the intrauterine growth rate of a fetus. Low birth weight can result from slow intrauterine growth, early conceptual age, or a combination of the two. Figure 3 plots birth weight versus gestational age to help describe the different type of outcomes that could be observed in a population of newborns. Most newborns are normal age and weight, but some will be born below the 10th percentile of weight for a given gestational age (small for gestational age, SGA) and others may be born before the 37th week of gestation (pre-term birth). Some newborns will be both pre-term and SGA. In papers two and three, we capture the correlation between birth weight and gestational age using an immaturity latent variable that follows a finite mixture distribution. Using the mixture distribution approach also allows us to identify a latent class of subjects with early gestational age and low weight. In reproductive health, epidemiologists refer to these types of subjects as belonging to the "residual" distribution (Wilcox et al. 2001), and they are of special importance because they are at increased risk for mortality and other forms of morbidity (Buekens et al. 2000). We attempt to find covariates that are associated with belonging to the residual distribution.

1.2 Other Applications

Our research is motivated by an application in reproductive health, but the methods we propose can be applied to other areas as well. We explain models using examples from other disciplines throughout this dissertation, and provide a brief overview of the most common applications here. SEMs are primarily found in the social sciences (Bollen 1989), and Sanchez et al. (2005) advocates their use in environmental epidemiology. In the statistics literature, reproductive toxicology has been used to motivate both multistate growth models (Sternberg and Satten 1999) as well as mixed continuous and discrete outcomes (Catalano and Ryan 1992). Other examples of multistate models come from AIDS research (DeGruttola and Lagakos 1989), the development of uterine fibroids (Dunson and Baird 2002), and breast cancer screening (Duffy et al. 1995).

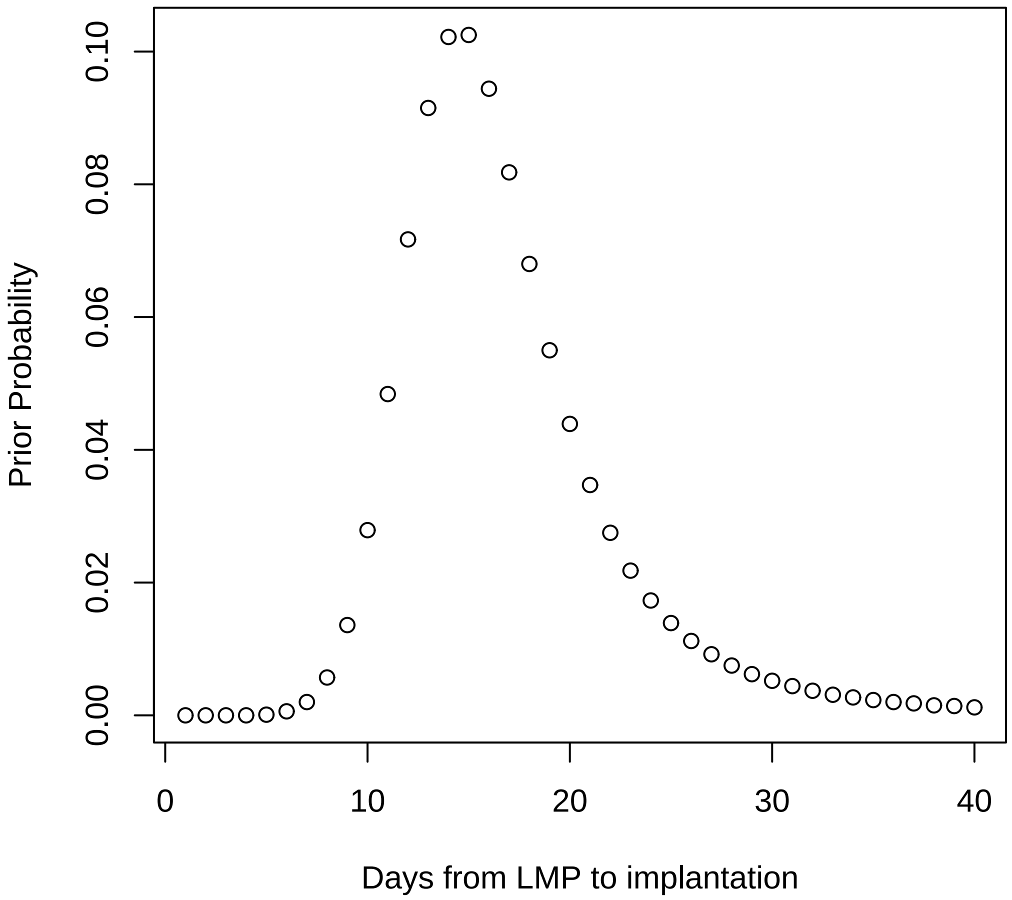

Figure 1: Prior probability of conception on a given day of the menstrual cycle, conditional on reaching that day of the cycle.

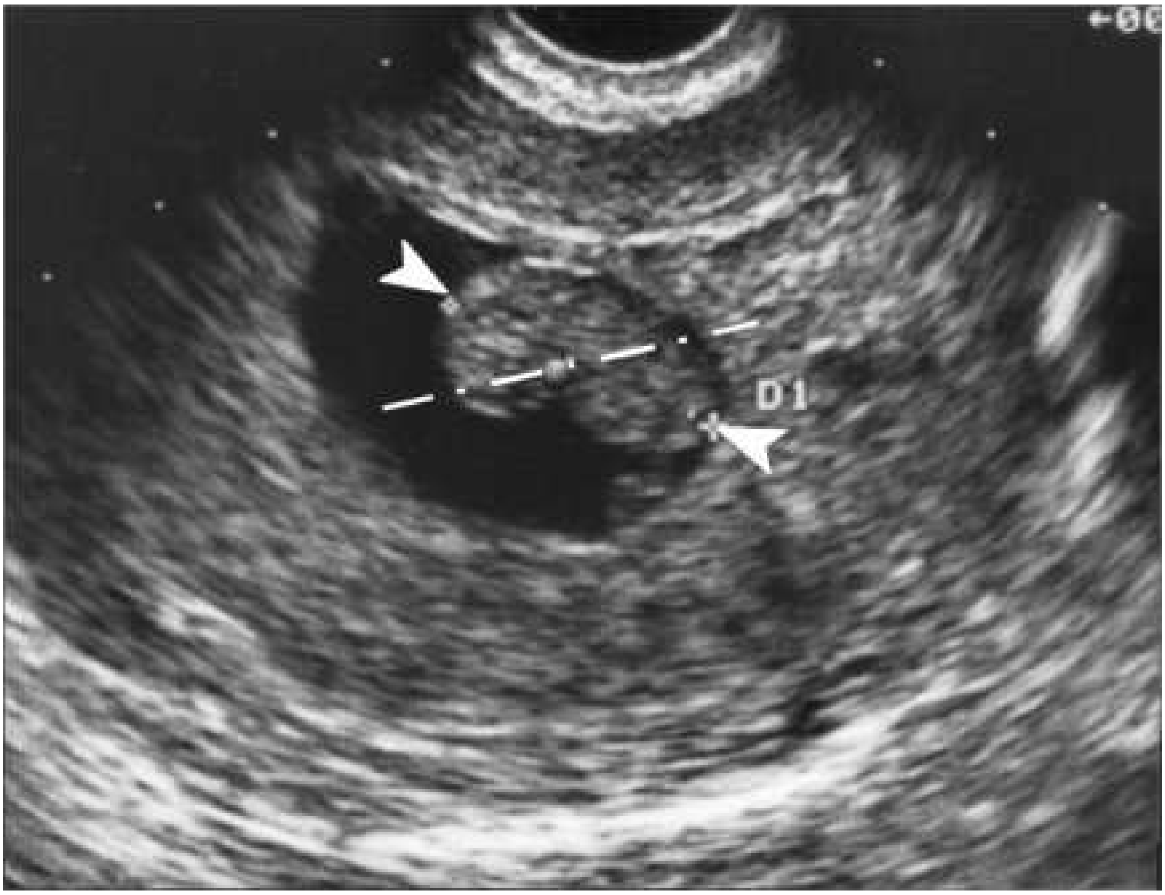

Figure 2: Early fetal pole visualization by ultrasound. The pole is indicated by the two arrowheads.

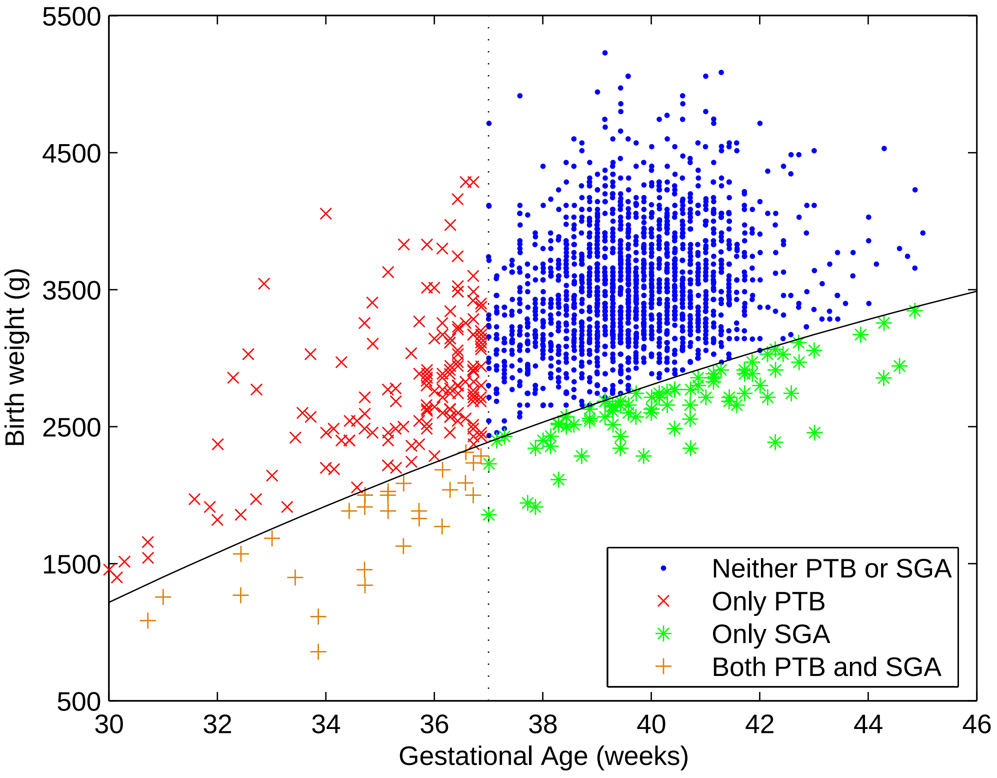

Figure 3: Definitions of birth weight, pre-term birth, and small for gestational age.

2 Latent Variable Methods for Longitudinal Data

Latent variable methods can capture many statistical concepts including random effects, missing data, sources of variation in hierarchical data, finite mixtures, latent classes, and clusters (Muthen 2002; Sanchez et al. 2005). In this section, we review several types of latent variable methods for longitudinal data.

2.1 Random Effects Models

Random effects models are commonly used for longitudinal data in which repeated or otherwise correlated measurements are taken on the same subject. The linear random effects model as proposed by Laird and Ware (1982) is specified by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, ..., n \tag{2.1}$$

where $\boldsymbol{y}_i$ is $n_i \times 1$, $\boldsymbol{X}_i$ is an $n_i \times p$ matrix of fixed predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{Z}_i$ a $n_i \times q$ vectors of covariates for the $q \times 1$ vector of random effects $\boldsymbol{b}_i$, and $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ vector of random errors. It is standard to assume that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent and both normally distributed

$$\boldsymbol{b}_i \sim N_q\left(\boldsymbol{0}, \boldsymbol{D}\right) \tag{2.2}$$

and

$$\boldsymbol{\epsilon}_i \sim N_{n_i}\left(\boldsymbol{0}, \boldsymbol{R}_i\right) \tag{2.3}$$

so that, marginally,

$$\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}_i, \boldsymbol{D} \sim N_{n_i}\left(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{R}_i + \boldsymbol{Z}_i'\boldsymbol{D}\boldsymbol{Z}_i\right) \tag{2.4}$$

Random effects models can also be specified for non-continuous outcomes. The generalized linear model (GLM) extends the methods of regression analysis to outcome variables that follow other distributions in the exponential family (McCullagh and Nelder 1989). For example, the binomial distribution is often used for binary outcomes and the Poisson distribution for counts. The generalized linear mixed model (GLMM) is the GLM generalization of the linear random effects model and is specified by Diggle et al. (1994) as

- Given $\boldsymbol{b}_i$, the responses $\boldsymbol{y}_i$ are mutually independent for each i and follow a GLM with density $f(y_{ij}|\boldsymbol{b}_i) = \exp[\{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))\}/\phi + c(y_{ij}, \phi)]$. The conditional moments $\mu_{ij} = \mathrm{E}(y_{ij}|\boldsymbol{b}_i) = \psi'(\theta_{ij})$ and $v_{ij} = \mathrm{Var}(y_{ij}|\boldsymbol{b}_i) = \psi''(\theta_{ij})\phi$, satisfy $h(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i$ and $v_{ij} = v(\mu_{ij})\phi$ where h and v are known link and variance functions, respectively, and $\boldsymbol{z}_{ij}$ is a subset of $\boldsymbol{x}_{ij}$

- The random effects $\boldsymbol{b}_i$ are mutually independent with a common underlying distribution, F

Frequentist likelihood-based estimation of the GLMM can be computationally difficult because high dimensional integrals are needed to evaluate the marginal likelihood (Breslow and Clayton 1993). In contrast, likelihood-based Bayesian methods are relatively straightforward to carry out using the Gibbs sampler because we can sample parameters conditionally on the random effects (Gelfand and Smith 1990).

In summary, the basic idea underlying a random effects model is that there is a natural heterogeneity across individuals that can be represented in their regression coefficients. Observations on the same individual are correlated due to sharing some unobservable variables, $\boldsymbol{b}_i$. A model of this type is sometimes referred to as a kind of latent variable model (Diggle et al. 1994) with the unobserved random effects being considered latent variables.

2.2 Latent Class Trajectory Models

Growth models, which model the changes in individuals over time, are one traditional application for random effects models. A growth model that allows for a quadratic effect of time for subject i measurement j at age t_{ij} could be specified using the notation of Laird and Ware by

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, t_{ij} + (\beta_2 + b_{2i})\, t_{ij}^2 + \epsilon_{ij} \tag{2.5}$$

where $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim$ i.i.d.$N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i})$. An equivalent representation of this model in terms of subject-specific latent variables η_{0i} for the intercept, η_{1i} for the linear effect and η_{2i} for the quadratic effect is

$$y_{ij} = \eta_{0i} + t_{ij}\eta_{1i} + t_{ij}^2\eta_{2i} + \epsilon_{ij} \tag{2.6}$$

where $\boldsymbol{\eta}_i \sim N\left([\beta_0, \beta_1, \beta_2]', \boldsymbol{D}\right)$ and $\boldsymbol{\epsilon}_i \sim$ i.i.d.$N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i})$.

It may be of interest how the shape of an individual growth trajectory, as measured by the latent variables, is related to some outcome measure. For example, Muthen and Shedden (1999) are interested in the association between the shape of an alcohol use trajectory from ages 18-25 and the risk of alcohol dependence at age 30. Directly relating the η coefficients to alcohol dependence using a logistic regression model is problematic because the meaning of a given $\boldsymbol{\eta}_i$ coefficient depends on the values of the other $\boldsymbol{\eta}_i$ coefficients. For example, a subject with a positive linear slope ($\eta_{1i} > 0$) could have decreasing dependence over time (if $\eta_{i2} < 0$), a purely linear increase over time (if $\eta_{i2} = 0$), or increase very rapidly over time (if $\eta_{i2} > 0$). A better option is a latent class trajectory model.

A latent class model summarizes shared features of the η coefficients using an underlying categorical variable. Specifically, the m_i-dimensional $\boldsymbol{\eta}_i$ are related to the K-dimensional latent categories $\boldsymbol{c}_i$ and p-dimensional covariates $\boldsymbol{x}_i$ using

$$\boldsymbol{\eta}_i = \boldsymbol{A}\boldsymbol{c}_i + \boldsymbol{\Gamma}_\eta \boldsymbol{x}_i + \boldsymbol{\zeta}, \qquad \boldsymbol{\zeta}_i \sim N_{m_i}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{2.7}$$

with conforming parameter matrices $\boldsymbol{A}(m_i \times K)$ and $\boldsymbol{\Gamma}_\eta(m_i \times p)$. The latent categorical variable, $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})$ follows a multinomial distribution with $c_{ik} = 1$ if individual i falls in class k and zero otherwise. A r-dimensional vector of observed dichotomous outcome variables $\boldsymbol{u}_i$ can then be related to the latent categories using a logistic regression model

$$\text{logit}\left(\Pr(u_{ij} = 1 | \boldsymbol{c}_i)\right) = \boldsymbol{\Lambda}_{u,j}\boldsymbol{c}_i \tag{2.8}$$

where each $\boldsymbol{\Lambda}_{u,j}$ is an $1 \times K$ parameter matrix, $j = 1, \ldots, K$. To complete the specification of the model, Muthen and Shedden (1999) allow the categorical latent variables $\boldsymbol{c}$ to be related to covariates $\boldsymbol{x}$ using a multinomial logit model for unordered polytomous responses. Defining $\pi_{ik} = \Pr(c_{ik} = 1 | \boldsymbol{x}_i)$ and the $\text{logit}(\boldsymbol{\pi}_i) = (\log\left[\pi_{i1}/\pi_{iK}\right], \ldots, \log\left[\pi_{i,K-1}/\pi_{iK}\right])$,

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\alpha}_c + \boldsymbol{\Gamma}_c \boldsymbol{x}_i \tag{2.9}$$

with a $(K-1)$ vector of intercepts $\boldsymbol{\alpha}_c$ and $(K-1) \times p$ parameter matrix $\boldsymbol{\Gamma}_c$. With this specification, the conditional distributions of $\boldsymbol{y}$ and $\boldsymbol{u}$ given $\boldsymbol{x}$ are influenced by parameters that vary across the categories of $\boldsymbol{c}$.

Several recent articles have built on these results. Guo et al. (2006) extend the latent class regression model so that it can include regression on latent predictors. Miglioretti (2003) discusses additional challenges that arise when the longitudinal measurements are mixtures of continuous, binary, and count data. Lin et al. (2000) use a latent class model to uncover subpopulation structure for both biomarker trajectories and the probability of disease outcome in highly unbalanced longitudinal data.

2.3 Structural Equation Models

SEMs are a flexible class of models originally proposed by Joreskog (1970) that allow modeling of both multivariate data and multiple, closely related predictors. The SEMs described in the book by Bollen (1989) incorporate a measurement model and a latent variable model for multivariate normal data. Using the notation of Bollen, the latent variable model is

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{2.10}$$

where $\boldsymbol{\eta}$ is and $m \times 1$ vector of latent responses, $\boldsymbol{B}$ an $m \times m$ coefficient matrix, and $\boldsymbol{\xi}$ are latent predictor variables ($n \times 1$) with coefficient matrix $\boldsymbol{\Gamma}$ ($m \times n$). The latent class model given in (2.7) is a special case of (2.10) where $\boldsymbol{B} = \boldsymbol{0}$, $\boldsymbol{\Gamma} = [\boldsymbol{A}; \boldsymbol{\Gamma}_\eta]$, and $\boldsymbol{\xi} = [\boldsymbol{c}_i; \boldsymbol{x}_i]$. The measurement model is then given by

$$\boldsymbol{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_y) \tag{2.11}$$

$$\boldsymbol{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \quad \boldsymbol{\delta} \sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma}_x) \tag{2.12}$$

The vectors $\boldsymbol{y}(p \times 1)$ and $\boldsymbol{x}(q \times 1)$ are observed variables with coefficient matrices, referred to as "factor loadings" in SEMs, $\boldsymbol{\Lambda}_y(p \times m)$ and $\boldsymbol{\Lambda}_x(q \times n)$ and with errors $\boldsymbol{\epsilon}(p \times 1)$ and $\boldsymbol{\delta}(q \times 1)$, respectively. Identifiability is an important consideration in SEMs, and is often accomplished by constraining covariance matrices to be diagonal, factor loading matrices to be equal to one, and/or the variance of the latent variables to be unity (Bollen 1989).

Some researchers have given particular attention to models that contain observed variables that are multiple indicators and multiple causes of a single latent response variable (MIMIC) as we have in the early pregnancy example (Joreskog and Goldberger 1975). In the MIMIC model, $\boldsymbol{\Lambda}_y = \boldsymbol{I}_q$ and the latent predictor variables are observed without error ($\boldsymbol{x} = \boldsymbol{\xi}$) so that η is a stochastic function of $\boldsymbol{x}$ with coefficients $\boldsymbol{\Gamma}$ and error component ζ_1:

$$\eta = \boldsymbol{\Gamma}\boldsymbol{x} + \zeta_1, \quad \zeta_1 \sim N(0, \sigma^2). \tag{2.13}$$

We say that η is affected by one or more x variables and it is associated with one or more y variables such that

$$\boldsymbol{y} = \boldsymbol{\Lambda}_y \eta + \boldsymbol{\epsilon} \tag{2.14}$$

with errors $\boldsymbol{\epsilon}$ and factor loadings $\boldsymbol{\Lambda}_y$. This model will be identifiable if there are at least two y variables, at least one $\boldsymbol{x}$, and η is provided a scale by either fixing the variance of ζ_1 or a factor loading term (e.g. $\zeta_1 \sim N(0,1)$ or $\lambda_1 = 1$). Identifiability of these models as well as more general models with multiple latent variables is often a concern and is discussed by Stapelton (1977), Robinson (1974), and Bollen (1989) among others.

Mixed effects models are closely related to SEMs where the random effects are latent response variables that capture the correlation among observations taken on the same sampling unit. While conceptually similar, there are some important differences in how they specify the mean and covariance models. To clarify the differences, consider a p-dimensional vector of responses $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ip}]'$ that has a linear relationship with dose. A linear random effects model as in (2.1) could be fit

$$y_{ij} = \beta_{0j} + \beta_{1j} * \text{dose}_i + b_{i1} + b_{i2} * \text{dose}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{2.15}$$

The same data could be analyzed with the SEM

$$\begin{aligned} y_{ij} &= \mu_j + \lambda_j * \eta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2) && (2.16) \\ \eta_i &= \gamma_1 * \text{dose}_i + \delta_i, \quad \delta_i \sim N(0, 1) && (2.17) \end{aligned}$$

The SEM estimates additional factor loading terms $\lambda_j (\lambda_j \geq 0)$ that multiply the latent variable to allow for a non-exchangeable correlation structure as well as differences in scale between the measured outcomes (Dunson 2006b). The correlation structure for the random effects model is allowed to vary by dose, but for the control group (dose = 0) simplifies to a compound-symmetric structure. In the SEM, the correlation between y_{ij} and $y_{ij'}$ is given by

$$\text{corr}(y_{ij}, y_{ij'}) = \frac{\lambda_j \lambda_{j'}}{\sqrt{\lambda_j^2 + \sigma_j^2}\sqrt{\lambda_{j'}^2 + \sigma_{j'}^2}} \tag{2.18}$$

which can differ for different sets of outcomes j and j'. A potentially important advantage to the random effects model is that it allows heterogeneity in the regression coefficients for dose across different responses. Models that combine the benefits of random effects models and SEMs are an area of current research (Dunson 2006b). Lin et al. (2000) and Roy et al. (2003) consider other approaches for allowing differences in scale when modeling multiple continuous outcomes.

While Bollen considers SEMs as models for multivariate normal data, they can be considered in a wider context. Longitudinal data can be measured on binary scale, categorical (ordinal or nominal) scale, metric scale (discrete or continuous), or in various combinations. Other researchers have considered latent variable models which allow measured and latent variables to have a broader class of distributions. Dunson (2000), Sammel et al. (1997), and Moustaki and Knott (2000) consider distributions in the exponential family for modeling variables that are collected on a variety of scales. In papers two and three, we relax the normality assumption and consider options including a latent class model in which the latent variable follows a mixture distribution.

2.4 Bayesian Structural Equation Models

The majority of literature on SEMs using latent variables is frequentist in nature, but Bayesian approaches have been proposed by several authors. There is a long history for Bayesian analysis of factor models, which are a special case of SEMs, in the psychometrics literature (Ansari and Jedidi 2000; Lee 1981; Martin and McDonald 1975). Early work on more general SEMs was done by Lee (1992) and Bauwens (1984).

There are several important differences between a Bayesian and frequentist approach to latent variable models. While frequentists usually assume normal distributions for latent variables, Bayesians must specify prior distributions for all unknown parameters in both the latent variable and measurement models. The priors allow for information from previous studies or theory to inform about the nature of the structural relationships, or, in the absence of such information, vague priors can be used. Computationally, Bayesian methods typically rely on Markov chain Monte Carlo (MCMC methods), which can be computationally expensive but provide exact posterior distributions for functions of any unknown parameters.

A Bayesian perspective allows us to estimate the posterior distribution of latent variables as well as parameters without making asymptotic assumptions. As pointed out by Palomo et al. (2007), estimating the joint posterior of latent variables has several advantages over a frequentist approach. For one, we can obtain point and interval estimate of factor scores for each individual in the study. We can then compare different factor scores among individuals (through posterior probabilities that, for example, one score is higher than another) and if an individual's factor score changes over time. We can also identify outlying individuals who are at the tail of the distribution.

A downside of a Bayesian approach to SEMs is that high autocorrelations can lead to slow convergence of the Gibbs sampler. It may take several hours of sampling until the Monte Carlo error in sampling is negligible. Recent articles have focused on MCMC methods to implement Bayesian analysis in complex cases including non-linear systems (Arminger 1998; Lee and Song 2004) and multi-level data (Dunson 2000; Song and Lee 2004). We discuss computational issues in more detail in section 3.

2.5 Multistate Models

A multistate process is a stochastic process that can take on a finite number of states K where each state describes the current condition (Jewell 2005; Kalbfleisch and Prentice 2002). In a general form where we allow time to be continuous, we define $Y(t)$ to be the state of the process at time t, $Y(t) \in \{1, 2, \ldots, K\}$ and $t \geq 0$. A Markov process is the simplest possible model that can be expressed for $Y(t)$. For a homogeneous population with no covariates, the transition rate from state i to j for an individual who is in state i at time t^- is given by

$$\begin{aligned} d\Lambda_{ij}(t) &= P\left[Y(t^- + dt) = j | Y(u), 0 \leq u < t, Y(t^-) = i\right] \\ &= P\left[Y(t^- + dt) = j | Y(t^-) = i\right] \end{aligned}$$

for all $Y(u), 0 \leq u < t, j \neq i$. The Markov assumption is that the process is memoryless in that only the current occupied state is need to specify the transition rates. Transition rates are allowed to depend on t, the amount of time since the beginning of the study.

In a general multistate model, subjects are allowed to transition from any state k to another state k'. Multistate growth models impose some restrictions on the multistate model. For one, growth models are unidirectional. That is, state transitions occur in a distinct order, only moving from state k to state $k+1$, $k = 1, \ldots, K-1$. Also, all subjects begin at the same initial state, $k = 1$. Traditional survival analysis, where subjects only move from "at risk" to "failed" states is a simple example of a unidirectional multistate process with $K = 2$.

In a regular Markov model, we are able to observe the states directly so that the state transition probabilities are the only parameters of interest. In a hidden Markov model the complete state history of a multistate process $Y(t)$ is not available at every time point t, but variables that are influenced by the state are observed. The influencing variables can be linked to underlying latent progress variables that are indicative of the waiting time spent in a state (Dunson and Baird 2002). In a longitudinal study, the state may be observed at several time points for each subject but the exact transition times are only known to occur within an interval. In a cross-sectional design, the state information is only available at one time point. We are concerned with the cross-sectional case where state transition times are interval censored and developmental progress covariates are measured once for each subject.

We consider multistate growth models where the moment the process begins, called the initiation time, is not known. The earliest approaches for multistate models with unknown initiation times relied on the restrictive Markov assumption (Kalbfleisch and Lawless 1985), which is not appropriate in our application. Semi-Markov models for fitting interval censored data with unknown initiation times (Satten and Sternberg 1999) have been developed that allow transition rates to depend on the amount of time spent in the state. In the fetal growth example, we consider the LMP date to be a known, pre-initiation time and conception the unknown initiation time. With this fixed time scale, it would also be possible to use the semi-Markov methods of DeGruttola and Lagakos (1989) or Sternberg and Satten (1999) by considering the pre-initiation time to be the first stage in the network. If the waiting times in different states are independent, then the semi-Markov assumption will hold. We avoid the semi-Markov assumption and allow individual transition rates between states to be influenced by a subject-specific latent variable. Our approach is most similar to Dunson and Baird (2002).

3 Computation

When proposing structural equation models, it is particularly important to consider which parameters in the model are identifiable and which must be constrained. In many cases, the marginal likelihood can be examined to determine how the data informs about each parameter. In a Bayesian analysis, fitting MCMC models can be somewhat of an art form so that alternative strategies may be needed to achieve dependable results (Gelfand and Sahu 1999). For example, including parameters that are non-identifiable, with appropriately vague priors, can help improve convergence. In this section, we first provide a brief introduction to Bayesian methods. We then define identifiability and Bayesian identifiability and explain one way in which including non-identifiable parameters in a model can improve computational performance while introducing a more general class of conditionally conjugate priors. This section is concluded by describing a data augmentation algorithm useful for fitting probit and multivariate probit models.

3.1 Bayesian Methods for Data Analysis

Bayesian methods are a useful tool for the applied statistician with applications in a wide range of areas due to the development of MCMC methods and the corresponding expansion of computing power. The details of a Bayesian approach to data analysis (Carlin and Louis 2000; Gelman et al. 2004) are well beyond the scope of this dissertation. Instead we will summarize a few important concepts that serve as the foundation for following sections.

The flexibility of the Bayesian framework allows it to deal with very complex analytical problems while using relatively simple conceptual methods. Bayesian methods are based on estimating the posterior distribution of p parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\eta}$, $f\left(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\eta}\right)$. Using Bayes' Theorem,

$$f\left(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\eta}\right) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta}) \tag{3.1}$$

where $L(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{\eta})$ represents the likelihood and $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ the prior distribution on $\boldsymbol{\theta}$. Inference is then based on the posterior distribution of parameters θ_i.

It is often difficult to directly calculate the complete joint posterior distribution, $f\left(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\eta}\right)$, when p is not small. Gibbs sampling (Gelfand and Smith 1990; Geman and Geman 1984) is a commonly used MCMC method to draw elements of $\boldsymbol{\theta}$ individually or in small groups. The Gibbs sampler generates values from the joint posterior distribution by iteratively drawing samples from the complete conditional distributions $f(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j, \boldsymbol{x}, \boldsymbol{\eta}, i \neq j)$ for each $i = 1, \ldots, p' \leq p$. After sufficient iterations, draws from the individual complete conditionals will eventually converge to being draws from the desired posterior distribution $f\left(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\eta}\right)$. When conjugate priors are chosen, it is often easy to sample from the complete conditionals because they are of known forms. For complete conditionals of unknown form, sampling can be done using the Metropolis-Hastings Algorithm (Chib and Greenberg 1995; Hastings 1970) with proposal densities generated from adaptive rejection samplings (ARS; Gilks and Wild 1992) or ARMS (ARS with a metropolis step; Gilks et al. 1995. Iterations continue until the parameters are judged to converge by a variety of diagnostic techniques.

3.2 Identifiability

A parameter θ is (frequentist) identifiable by observations of a random vector $\boldsymbol{y}$ if distinct values for θ yield distinct distributions for $\boldsymbol{y}$ (Basu 1983). Bayesian identifiability (Dawid 1979) focuses on the posterior distribution and is concerned with whether the data and prior provide information about the parameters. Situations can arise where parameters are not frequentist identifiable, but can identified in a Bayesian analysis that uses informative priors. In these cases, parameters that are only Bayesian identifiable need to be recognized and interpreted appropriately.

Specifying a proper prior ensures a proper posterior. However, improper priors are often used to reflect ignorance of parameters or for mathematical convenience. Gelfand and Sahu (1999) explore the use of improper priors for the generalized linear model and developed the following theorem:

> Consider the Gaussian linear model $y = X\beta + \epsilon$ with $X_{n\times p}, \beta_{p\times 1}$, with $\text{rank}(X) = r < p$ and $\epsilon \sim N(0, \sigma^2 I)$. Suppose that the prior $f(\beta, \sigma^2)$ takes the form $f(\beta)f(\sigma^2)$, where $f(\beta) = f(\gamma)$, a proper prior on $\gamma = \Omega\beta$ with $\Omega\beta$ estimable. Then $f(\beta, \sigma^2|y)$ is improper.

A proper prior will need to be specified on some non-estimable parameters for the posterior to be proper. In the frequentist setting, this is usually accomplished by constraining certain parameters to be equal to some value (usually 0). Gelfand and Sahu (1999) give a logistic growth curve example to illustrate that this is not an ideal solution when performing Gibbs sampling. They show that placing non-informative (but not improper) proper priors on all parameters in a less than full rank model can significantly decrease the autocorrelation of estimable parameters. The worst convergence of the Markov chain is attained by imposing constraints. Gelman et al. (2003) also notes for the generalized linear model more diffuse working prior densities improves mixing. Recent papers on parameter expansion elaborate on these ideas.

3.3 Parameter Expansion

Parameter expansion adds new parameters that are not identifiable to a model. An example is replacing a parameter θ by a product $\phi\psi$ so that inference can be obtained only about the product and not the individual parameters. This technique has been used to improve the convergence rate of the EM algorithm by Liu et al. (1998) in a non-Bayesian setting. Gelman (Gelman 2004; Gelman et al. 2003) uses parameter expansion as a computational aid within a Gibbs sampling framework, which we will find more useful.

Gelman considers the following hierarchical model expressed as a regression with coefficients in M batches.

$$y = \sum_{m=1}^{M} X^{(m)}\beta^{(m)} + \text{error}$$

where $X^{(m)}$ is the mth submatrix of predictors and $\beta^{(m)}$ is the mth subvector of regression coefficients. The J_m coefficients in each subvector $\beta^{(m)}$ have an exchangeable $N(0, \sigma_m^2)$ prior distribution for $j = 1, \ldots, J_m$. He uses this general notation to allow for several different types of models. We are interested in a mixed effects model where, for the fixed effects, $\sigma_m = \infty$ and, for the random effects, σ_m is estimated from the data.

This hierarchical model can become stuck when σ_m happens to be near 0. In the updating stage of the Gibbs sampler, $\beta^{(m)}$ will be shrunk to 0, at the next update σ_m

will be near 0, and so on. Simulations show that it can take considerable time for the Gibbs sampler to escape this situation (Gelman et al. 2003).

Gelman (2004) advocates using the following parameter-expanded model where each component of the regression model is multiplied by a new parameter, α_m^*. Let the old model be

$$y = \sum_{m=1}^{M} X^{(m)} \beta^{(m)} + \text{error}$$

and the new, expanded model be

$$y = \sum_{m=1}^{M} \alpha_m^* X^{(m)} \beta^{*(m)} + \text{error}$$

The estimable function $|\alpha_m^*| \sigma_m^*$ in the new model maps to σ_m in the old model and $\alpha_m^* \beta_j^{*(m)}$ from the new model maps to $\beta_j^{(m)}$ from the old model. The individual parameters α_m^*, $\beta_j^{*(m)}$, and σ_m^* are not frequentist identifiable, but if proper priors are specified, they will be Bayesian identifiable.

Gelman (2006) examines different choices for prior distributions in the hierarchical model. He expands the family of conditionally-conjugate prior distributions by applying a redundant multiplicative reparameterization to the general hierarchical model:

$$\begin{aligned} y_{ij} &\sim N(\mu + \xi \eta_j, \sigma_y^2) \\ \eta_j &\sim N(0, \sigma_\eta^2) \end{aligned}$$

There is an implicitly conditionally conjugate prior distribution on the random effects standard deviation, $\sigma_\alpha = |\xi| \sigma_\eta$. Using the usual conjugate priors for ξ (normal) and σ_η^2 (inverse-gamma), σ_α has the distribution of the absolute value of a noncentral-t variate. This is referred to as a folded noncentral t distribution. Gelman also discusses a special case of the folded non-central t, the half-t distribution. This half-t is appealing because the distribution of ξ will be symmetric about 0, however the half-t family is not conditionally conjugate.

In summary, Gelman's findings and advice include:

- Start with a noninformative uniform prior density on the standard deviation σ_α
- For a non-informative but proper density, try a half-normal centered at 0, $\xi \sim N(0, 100^2)$
- Avoid the usual inv-gamma(ϵ, ϵ) priors on σ_α^2 especially when small values of σ_α are possible. Inference is very sensitive to ϵ and the prior distribution is not non-informative as desired.
- To restrict σ_α from large values, use a half-t family

We found these suggestions particularly helpful in the early fetal growth analysis. Convergence rates were improved while using a less informative prior than the usual gamma prior on precision parameters.

3.4 Data Augmentation

Data augmentation is a general scheme in which the observed data are enhanced so as to make it easier to analyze. Dempster et al. (1977) popularized the method with their paper on the EM algorithm for maximizing a likelihood function, but their approach was not called "data augmentation" until Tanner and Wong (1987). From a Bayesian perspective, latent data augments the observed data so we can use both the latent and observed data to calculate the posterior distribution of the parameters of interest (van Dyk and Meng 2001). For observed binary data, data augmentation postulates the existence of an underlying continuous variable such that the event is observed to occur if the continuous variable is above some threshold value. Using this idea, we can use latent variables to facilitate model computation for categorical outcomes.

The data augmentation algorithm outlined by Albert and Chib (1993) connects a probit regression model on an observed dichotomous outcome with a normal linear regression model on a continuous latent outcome. Their algorithm involves sampling latent outcome variables from the truncated normal distribution, with the truncation conditional on the observed dichotomous outcome variable. Let $y_1, \ldots, y_n$ be the observed dichotomous outcome on subject i and y_i^* the underlying latent response, where y_i^* are i.i.d. $N(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$. Define $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise so that the y_i are independent Bernoulli random variables with $Pr(y_i = 1) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The complete conditional distribution of the latent y_i^* is then given by

$$y_i^* | y_i, \boldsymbol{x}_i, \boldsymbol{\beta} \sim \begin{cases} N_1\left(\boldsymbol{x}_i'\boldsymbol{\beta}, 1\right) I(y_i^* > 0) & \text{if} \quad y_i = 1 \\ N_1\left(\boldsymbol{x}_i'\boldsymbol{\beta}, 1\right) I(y_i^* < 0) & \text{if} \quad y_i = 0 \end{cases} \tag{3.2}$$

With a conjugate prior specification for $\boldsymbol{\beta}$, Gibbs sampling can then be used to sample from the posterior distribution of $\boldsymbol{\beta}$ using well known linear regression results. To simplify subsequent notation, the general truncated normal distribution as given in (3.2) with $\sigma^2 = 1$ will be specified as $TN\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2, y_i\right)$. The underlying latent normal approach can naturally be extended to the multinomial probit model (McCulloch and Rossi 1994) for univariate ordinal categorical data and the multivariate probit model (Chib and Greenberg 1998) for multivariate binary correlated outcomes.

3.5 Multivariate Probit Models

Chib and Greenberg (1998) detail an approach for modeling multivariate correlated binary data that is a natural extension of the univariate probit model framework. Let y_{ij} be a binary response for the ith observation unit and jth variable, $j = 1, \ldots, p$. Chib and Greenberg (1998) postulates the existence of an underlying multivariate normal random vector

$$\boldsymbol{y}_i^* \sim N_p(\boldsymbol{X\beta}, \boldsymbol{\Sigma}) \tag{3.3}$$

such that $y_{ij} = I(y_{ij}^* > 0)$ and $\boldsymbol{\Sigma}$ is a correlation matrix for identifiability. We can then extend this representation for joint modeling of continuous and binary correlated outcomes. Let $\boldsymbol{z}_{i1}$ be a p_1-dimensional vector of observed normal continuous response and $\boldsymbol{z}_{i2}$ a p_2-dimensional vector of binary responses with underlying responses $\boldsymbol{z}_{i2}^*$. The joint distribution of $\boldsymbol{z}_{i1}$ and $\boldsymbol{z}_{i2}^*$ is given by

$$\begin{pmatrix} \boldsymbol{z}_{i1} \\ \boldsymbol{z}_{i2}^* \end{pmatrix} \sim N_{p1+p2} \left(\boldsymbol{X\beta}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) \tag{3.4}$$

where the diagonal elements of $\boldsymbol{\Sigma}_{22}$ are fixed at one for identifiability. The problem is now reduced to modeling normally distributed correlated data using observed and latent continuous random variables.

While the relatively direct probit model framework for analyzing mixed continuous and discrete outcomes is conceptually appealing, it can be computationally demanding. In a Bayesian analysis, high autocorrelations within parameters and cross correlations between parameters can make convergence of the Gibbs sampler excruciatingly slow. A number of approaches have been suggested to improve mixing and the convergence rates of longitudinal models and the multivariate probit model that we may be able to adapt for modeling mixed continuous and discrete outcomes.

Gelfand et al. (1995) presents hierarchical centering reparamaterizations that often have improved convergence for a broad class of normal linear mixed models. Hierarchical centering can be explained by considering the simple random effects model $y_{ij} = \mu + b_{i1} + \epsilon_{ij}$, where $b_{i1} \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_e^2)$ i.i.d for each subject i and repeated measure j. In order to create a better-behaved posterior surface, a second level hierarchy is introduced by defining $\eta_i = \mu + b_{i1}$ such that $\eta_i|\mu \sim N(\mu, \sigma_b^2)$. The hierarchical centering transformation can break the high posterior correlations among parameters that commonly occur in random effects models and greatly improve convergence of the Gibbs sampler.

The benefits of blocking, which involves updating parameters in groups, in hierarchical longitudinal models are advocated by Chib and Carlin (1999). They also specifically consider longitudinal binary probit models of the form $\Pr(y_{ij} = 1|\boldsymbol{b}_i) = \Phi(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i)$ where Φ is the standard normal CDF and $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, \boldsymbol{D})$. Then the conditional distribution of the latent outcome, $y_{ij} = I(y^*_{ij} > 0)$, is given by $y^*_{ij}|\boldsymbol{b}_i \sim N(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i, 1)$. A simple algorithm for sampling from the joint posterior that updates parameters one at a time will include sampling (1) $\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{b}, \boldsymbol{D}$ and (2) $y^*_{ij}|y_{ij}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{D}$ as well as (3) $\boldsymbol{b}|\boldsymbol{y}^*, \boldsymbol{\beta}, D$ and (4) $\boldsymbol{D}|\boldsymbol{b}$. Chib and Carlin (1999) refine this algorithm by marginalizing the distribution of $\boldsymbol{y}^*_i$ over the random effects so that $\boldsymbol{y}^*_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, I_{ni} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i)$. Steps (1) and (2) of the simple algorithm are blocked so that we sample both $\boldsymbol{\beta}$ and $\boldsymbol{y}^*_i$ from $[\boldsymbol{\beta}, \boldsymbol{y}^*_i|\boldsymbol{y}, \boldsymbol{D}]$ by sampling (a) $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{D}$ and (b) $\boldsymbol{y}^*_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{D}$. In example datasets, this approach and some similar variations significantly reduce autocorrelations for $\boldsymbol{\beta}$, but are less effective on improving convergence for the random effect variance, $\boldsymbol{D}$.

Imai and Van Dyk (2005) and Nobile (1998) propose methods for improving the convergence of the multinomial probit model. They realize that one of the reasons for the poor convergence of probit models is the variance of the latent response variable is not identifiable and is fixed at one by convention. Imai and Van Dyk (2005) introduce a new latent variable $\tilde{y}_i^* = \alpha y_i^*$ so that $\tilde{y}_i^* | \beta, \alpha, y_i \sim TN(\alpha \boldsymbol{X}_i \boldsymbol{\beta}, \alpha^2, \boldsymbol{y})$. An inverse gamma prior is specified for α^2 and it is included as part of the Gibbs sampling algorithm. Nobile (1998) develops a hybrid Markov chain in which a Metropolis step is performed after each cycle of the Gibbs sampler to change the scale of the current state. Both methods improve the computational efficiency of simulated data, with the more recent paper of Imai and Van Dyk (2005) having the best performance in most situations.

4 Bayesian Modeling of Embryonic Growth using Latent Variables

4.1 Abstract

In a growth model, individuals move progressively through a series of states where each state is indicative of their developmental status. Interest lies in estimating the rate of progression through each state while incorporating covariates that might affect the transition rates. We develop a Bayesian discrete time multistate growth model for inference from cross-sectional data with unknown initiation times. For each subject, data are collected at only one time point at which we observe the state as well as covariates that measure developmental progress. We link the developmental progress variables to an underlying latent growth variable that can affect the transition rates. We also examine the association between latent growth and the probability of future events. We use a Markov chain Monte Carlo algorithm for posterior computation and apply our methods to a novel study of embryonic growth and pregnancy loss in which we were able to find evidence in favor of a previously hypothesized but unproven association between slow growth early in pregnancy and increased risk of future loss.

4.2 Introduction

Fetal growth is important both clinically and in epidemiologic studies relating growth to later pregnancy and developmental outcomes. To date, large cohort studies have relied on the last menstrual period (LMP) and onset of bleeding to determine the time of miscarriage. This approach is centrally flawed because it ignores the developmental state of the fetus prior to the loss. Development may stop days to weeks prior to the onset of bleeding so that, for example, a loss classified as a miscarriage in the 10th week of gestation may have been the loss of an appropriately-grown 10 week fetus, or a fetus with development arrested at 7 weeks. Coupling cross-sectional ultrasound information with the ability to estimate growth provides an opportunity to estimate probable developmental state prior to miscarriage. This can help researchers more accurately identify which insults must have occurred prior to the pregnancy loss and rule out exposures that occur after development arrest but before the onset of bleeding.

Another complication in determining the miscarriage date is uncertainty in pregnancy dating. After the occurrence of a positive pregnancy test, time of conception is traditionally dated as two weeks after the LMP. Such dating is notoriously imprecise due to variation in follicular phase length, the time between menses and ovulation, and, in rarer cases, missed menses (Kramer et al. 1988; Savitz et al. 2002). However, the distribution of the follicular phase length has been well characterized (Wilcox et al. 2001) and can be used to obtain more realistic estimates of the conception time. With these difficulties in mind, our goal in the early pregnancy example analysis is two-fold: (1) to determine whether there exists variation in fetal growth before 20 weeks gestation, and (2) if variation in growth exists, to determine if growth is associated with the probability of future pregnancy loss. In particular, if (2) holds the current practice of using ultrasound dating in early pregnancy loss studies may need to be re-evaluated.

This vexing problem can be posed statistically using a latent multistate growth process with unknown initiation time. In a multistate model the current condition of a subject is summarized by assignment to one of a finite number of K states (Jewell 2005). A growth model assumes that all subjects begin at the same initial state and advance unidirectionally through subsequent states. That is, all subjects progress in order from state k to $k+1$, $k = 1, \ldots, K-1$ without skipping any states or regressing to previous states. This article focuses on the problem of estimating such models using cross-sectional data in which transition times are not observed. For each subject, we have information at one time point on the current state as well as measurements of developmental progress that are surrogates of the amount of time spent in a state. Also, the moment the growth process begins, referred to as the initiation time, is not known, though some time point before initiation is known.

Multistate growth models with interval censored transition times appear in several applications. Often the initiation time is also unknown but not considered in the analysis. One example involves studying tumor growth in mice. Typically, a mouse is exposed to a carcinogenic compound and a tumor may develop some time following the exposure. At a later date, the mouse is sacrificed to determine how far the tumor has progressed (Albert and Shih 2003; Dewanji and Kalbfleisch 1986; Ryan and Orav 1988). In these studies, the investigators analyze time from exposure rather than time from tumor initiation. We focus on measuring growth as a function of time since initiation and demonstrate that, in our application, failing to account for the unknown initiation time can have a large impact on the analysis. Other examples of multistate models include AIDS progression (DeGruttola and Lagakos 1989), breast cancer (Duffy et al. 1995), and the development of uterine fibroids (Dunson and Baird 2002). Our research is motivated by the Right from the Start (RFTS) study of embryonic development.

RFTS is a prospective cohort study that identified women who were planning to conceive or in early pregnancy (Promislow et al. 2004). Investigators enrolled women when they had a positive pregnancy test, and then promptly provided an early first trimester ultrasound. At the time of the ultrasound, each fetus was assigned to one of three states based on the presence or absence of important developmental features: (1) only gestational sac present, (2) fetal pole present without regular cardiac activity, and (3) fetal pole with normal cardiac activity. Lack of cardiac activity is a natural state of development that every fetus experiences for hours to days while the heart forms, becomes detectable by ultrasound, and before a heart rate is established. It is not necessarily indicative of a problem pregnancy. The ultrasound also provides two measurements of developmental progress, the mean gestational sac diameter and fetal pole length. Both of these covariates have been studied extensively and found to be strongly associated with time since LMP (Filly and Hadlock 2000; Hadlock et al. 1992).

We develop a latent variable method for incorporating development progress variables using structural equations (Joreskog 1970) in a multistate growth model. Our approach differs from previous analyses that used developmental progress covariates directly. For example, Albert and Shih (2003) jointly model tumor onset and growth after onset using a single tumor volume measurement. When multiple measures of developmental progress are available, previous direct approaches have attempted to categorize disease severity into a set of levels (Craig et al. 1999). The number, volume and location of a tumor could be used to assign subjects into severity levels ranging from high to low, although defining the cut points for the levels may be unclear. Instead, we follow a latent variable approach similar to Dunson and Baird (2002) that naturally combines multiple measurements of developmental progress into the underlying latent growth concept. We allow the latent variable to be continuous, which is more flexible than the direct approach in that it does not pre-specify arbitrary severity boundaries in the classification procedure. We are also able to consider associations between the latent variable and future outcomes. For example, in our early pregnancy analysis we identify an important relationship between latent growth and the risk of pregnancy loss by 20 weeks.

Bayesian methods are well-suited to the embryonic growth application because abundant prior knowledge is available. We propose statistical models in a specific form that allows prior information about the time from LMP to conception and time to state transitions to be readily incorporated. Available methodology using Markov models (Kalbfleish and Lawless 1985) would not be appropriate in our application because it assumes that the transition rates are independent of the amount of time spent in state. Semi-Markov models have also been developed for fitting interval censored data with unknown initiation times. While Satten and Sternberg (1999) allow the transition rate to depend on the time spent in the current state, they do not incorporate surrogates of developmental progress that are functions of time since initiation in addition to state membership. Bayesian methods that consider surrogates of developmental progress for interval censored data have been proposed (Dunson and Baird 2002), but do not allow the developmental progress covariates to be functions of an unknown initiation time. In addition, Dunson and Baird (2002) treat the multistate growth model as the primary outcome of interest, while we additionally consider joint modeling of growth with a future pregnancy outcome.

Section 4.3 describes the underlying data structure, the general stochastic and latent variable models, and the specific models we used in the RFTS example analysis. In section 4.4 we provide a Bayesian approach to fitting such a model with the necessary conditional distributions in Appendix A. Section 4.5 contains the application of our methods to a study of early pregnancy and section 4.6 discusses the results.

4.3 Model Description

4.3.1 Data Structure

Suppose that a K state growth process has an unknown initiation time and some known pre-initiation time as depicted in Figure 4 for a $K = 3$ state model. In our RFTS example analysis, the date of the last menstrual period (pre-initiation) is known while the subsequent time of conception (initiation) is unknown. Let t_i^I be the time interval between pre-initiation and initiation for subject $i, i = 1, \ldots, n$. We then standardize the time axis so that the initiation time is zero for each subject and measure other time points relative to this zero point. On this scale, the pre-initiation time is $-t_i^I$ as indicated in Figure 4. Let $T_i^{(k)}$ be the unobserved transition time from state k to $k+1$ with each subject's current state ascertained at time W_i. W_i is not known though the sum $W_i + t_i^I$ is observed.

To obtain information on the amount of time since initiation (W_i) and state progression, let there be one or more measurements of development progress available at W_i for each subject. For the measurements to be useful, we should expect the continuous measurements will increase (or decrease) stochastically as a function of time since initiation and possibly other covariates. Similarly, for dichotomous measurements, the probability that they are present or absent should be a function of time since initiation. In the RFTS embryonic growth example, two continuous variables are available, the fetal pole length and gestational sac diameter. Both of these variables have been shown to increase with time since LMP (Filly and Hadlock 2000; Hadlock et al. 1992), and it is reasonable to assume that they are more accurately a function of time since conception.

4.3.2 Latent Variable Model

Regardless of state, let there be C measures of developmental progress, $P_{1i}, P_{2i}, \ldots P_{Ci}$, available at time W_i, $i = 1, \ldots, n$. Similar to Dunson and Baird (2002), we link the $n \times 1$ vectors $\boldsymbol{P}_c$ to an underlying latent growth variable $\boldsymbol{Z}^*(n \times 1)$ and time since initiation ($\boldsymbol{W}$)

$$\boldsymbol{P}_c = g_c(\boldsymbol{Z}^*, \boldsymbol{W}; \boldsymbol{\Lambda}_c, \sigma_c), \qquad c = 1, \ldots, C \tag{4.1}$$

where $g_c(\cdot)$ is a function involving parameters $\boldsymbol{\Lambda}_c$ with error component σ_c. For each subject, we assume that the latent growth variable $Z_i^* \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma_{z^*}^2)$ as in a Bollen (1989) structural equation model. Z_i^* incorporates the concept of individual growth and models the correlation between the developmental progress variables. We allow Z_i^* to have an expected value that can be a function of covariates $\boldsymbol{X}_i$ and parameters $\boldsymbol{\beta}$. With no covariates, Z_i^* can be thought of as a random intercept from a mixed effects model (cf. Laird and Ware 1982).

4.3.3 Stochastic Model

We next describe the model for the state transition rates, $\alpha^{(k)}$, from state k to $k+1$. Using a discrete time model, we partition the time axis into J disjoint intervals $I_j = (t_{j-1}, t_j]$, with $0 < t_1 < \cdots < t_J, t_J > (W_i + t_i^I)\ \forall\ i$. In the RFTS example analysis, we use time intervals of length one day because ultrasound and LMP information are collected as dates. We then characterize the general transition rate for subject i in interval j from state k to $k+1$ to be

$$\alpha_{ij}^{(k)} = \Pr\left(T_i^{(k)} \in I_j | T_i^{(k)} > t_{j-1}, T_i^{(k-1)} \leq t_{j-1}\right) \tag{4.2}$$

for $k = 1, \ldots, K-1$. We assume that the time intervals are sufficiently small so that there is zero probability of two transitions within the same interval. For individuals in state 1, the transition times $T_i^{(1)}, \ldots, T_i^{(K-1)}$ are right censored. For individuals in state $s, s \in \{2, \ldots, K-1\}$, entry times $T_i^{(1)}, \ldots, T_i^{(s-1)}$ are left censored while $T_i^{(s)}, \ldots, T_i^{(K-1)}$ are right censored. Finally, for individuals in state K, all entry times $T_i^{(1)}, \ldots, T_i^{(K-1)}$ are left censored. Since we do not directly observe $T_i^{(k)}$, we sample the unknown intervals of entry as a step in the MCMC algorithm.

We also connect Z_i^* to the state transition model as defined in (4.2)

$$\alpha_{ij}^{(k)} = h_k(Z_i^*, \boldsymbol{X}_{ki}; \omega_{jk}, \boldsymbol{\Gamma}_k), \qquad k = 1, \ldots, K-1 \tag{4.3}$$

where $h_k(\cdot)$ is some smooth, monotone function mapping from the real numbers to the probability space. It involves baseline transition probability parameter ω_{jk}, covariate parameters $\boldsymbol{\Gamma}_k$ and possibly additional covariates $\boldsymbol{X}_{ki}$. We anticipate that subjects who develop quickly will have large values of Z_i^* while slow-growing subjects will have relatively small values.

We consider the transition likelihood for a K state model where the initiation time is fixed at zero for each subject and the transition times are complete. Let $l_i^{(k)} = \left\{l : T_i^{(k)} \in I_l\right\}$ and $m_i = \{m : W_i \in I_m\}$ be the discrete time intervals in which $T_i^{(k)}$ and W_i, respectively, fall. For an individual in state 1 at W_i, their contribution to the transition likelihood is $\prod_{j=1}^{m_i}\left(1-\alpha_{ij}^{(1)}\right)$. Letting $l_i^{(0)} = 0$, an individual in state $s, s \in \{2,\ldots,K-1\}$, contributes $\left\{\prod_{k=1}^{s-1}\left[\alpha_{i,l_i^{(k)}}\prod_{j=l_i^{(k-1)}+1}^{l_i^{(k)}-1}\left(1-\alpha_{ij}^{(k)}\right)\right]\right\}\prod_{j=l_i^{(s-1)}+1}^{m_i}\left(1-\alpha_{ij}^{(s)}\right)$ to the complete transition likelihood. Finally, a subject in state K with all $T_i^{(k)}$ complete contributes
$\prod_{k=1}^{K-1}\left[\alpha_{i,l_i^{(k)}}\prod_{j=l_i^{(k-1)}+1}^{l_i^{(k)}-1}\left(1-\alpha_{ij}^{(k)}\right)\right]$.

4.3.4 Fetal Growth Model

In the RFTS analysis, we use the inverse probit function for $h_k(\cdot)$ in a $K = 3$ state model. The discrete time probit regression models describing the 1 → 2 and 2 → 3 state transition rates for embryo i in interval I_j are

$$\alpha_{ij}^{(1)} = \Phi\left(\boldsymbol{M}_j\boldsymbol{\omega} + \gamma_1 Z_i^*\right) \tag{4.4}$$

$$\alpha_{ij}^{(2)} = \Phi\left(\nu + \gamma_2 Z_i^*\right) \tag{4.5}$$

where $\boldsymbol{M}_j$ is the j-th row of a design matrix that provides regression splines for the baseline transition probabilities $\boldsymbol{\omega}$. The most flexible model would fit one parameter per time interval, but we use a spline function with four degree of freedom to decrease the dimensionality. We also found that an intercept term, ν, is sufficient to model this baseline transition probability from state two to three.

For the latent growth model, we allow $\boldsymbol{Z}^*$ to be a function of time since conception ($\boldsymbol{W}$) and some dichotomous variable $\boldsymbol{D}$. We use a log transformation of gestational sac diameter ($\boldsymbol{Z}_1$) and crown rump length ($\boldsymbol{Z}_2$) and a quadratic time effect consistent with previous approaches (Filly and Hadlock 2000)

$$\begin{aligned} \boldsymbol{Z}^* &= \beta_1 \boldsymbol{D} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n \tau_{z^*}^{-1}) \qquad (4.6) \\ \boldsymbol{Z}_1 &= \lambda_{01} + \lambda_{11}\boldsymbol{W} + \lambda_{21}\boldsymbol{W}^2 + \boldsymbol{Z}^* + \boldsymbol{\delta}_1 \\ &= \boldsymbol{X}_1\boldsymbol{\Lambda}_1 + \boldsymbol{Z}^* + \boldsymbol{\delta}_1, \quad \boldsymbol{\delta}_1 \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n \tau_{z_1}^{-1}) \qquad (4.7) \\ \boldsymbol{Z}_2 &= \lambda_{02} + \lambda_{12}\boldsymbol{W} + \lambda_{22}\boldsymbol{W}^2 + \boldsymbol{Z}^* + \boldsymbol{\delta}_2 \\ &= \boldsymbol{X}_2\boldsymbol{\Lambda}_2 + \boldsymbol{Z}^* + \boldsymbol{\delta}_2, \quad \boldsymbol{\delta}_2 \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n \tau_{z_2}^{-1}) \qquad (4.8) \end{aligned}$$

In the RFTS analysis, D_i is either an indicator variable for infant's gender, mother's black ethnicity, or if the mother had any previous live births (parity). There is evidence to indicate that these covariates are associated with growth between week 25 and 42 (Zhang and Bowes 1995), but no evidence that they are influential earlier in pregnancy.

The structural equations specified by (4.6) - (4.8) include constraints that are necessary for frequentist identifiability. The factor loadings for $\boldsymbol{Z}^*$ are fixed at one so we are able to identify the precision parameters τ_{z_1} and τ_{z_2}. With two measurements of developmental progress, we are also limited to one latent variable that models the covariance between fetal pole length and gestational sac diameter. General identifiability conditions covered by Bollen (1989) among others can be difficult to derive, so we only discuss identifiability in our example.

Finally, we examine the association between Z_i^* and future events, Y_i, that occur after W_i. In RFTS, we looked for an association with the probability of a spontaneous abortion (SAB) by week 20 using a probit regression model, $\Pr(Y_i = 1 | Z_i^*) = \Phi(\mu_0 + \mu_1 Z_i^*)$. In this analysis, we include all subjects who did not have a SAB as well as subjects who had a SAB at some time point following their ultrasound. We exclude subjects who had evidence of a SAB prior to their ultrasound because the date of the loss could not be determined as accurately as the date of the ultrasound. In subjects with an early loss, the developmental progress variables would be function of time from conception to loss rather than time to ultrasound.

The path diagram in Figure 5 illustrates the dependencies in the early fetal growth model. We use boxes to represent observed variables, both with (solid lines) and without (dashed lines) distributional assumptions. Circles represent unobserved variables and arrows indicate association so that lack of arrows signifies conditional independence. Latent growth (Z^*) models the correlation between the gestational sac diameter and fetal pole length and is allowed to influence both the state transition rates and the risk of pregnancy loss by 20 weeks. We estimate the unknown time from conception to the ultrasound (W) by subtracting the estimated time from LMP to conception (t^I) from the observed total time from LMP to ultrasound ($W + t^I$). The two measurements of developmental progress obtained by the ultrasound are allowed to increase as a function of time since conception so that latent growth is not a function of W. Finally time from LMP to conception (t^I) is associated with the state transitions, but not through the transition rate like latent growth. Instead, time from LMP to conception indicates the starting point of the growth process, from which we measure the number of days to the first and subsequent state transitions.

4.4 Bayesian Analysis

We use Bayesian methods, which estimate the joint posterior distribution of the parameters given the data, to conduct our analysis. The joint distribution of all parameters can be written in general as

$$\begin{aligned}
&f(\boldsymbol{\beta}, \tau_{z^*}, Z_i^*, \boldsymbol{\mu}, \boldsymbol{\omega}, \gamma_1, \nu, \gamma_2, \mathbf{\Lambda}_1, \tau_1, \mathbf{\Lambda}_2, \tau_2, t_i^I) \\
&\quad = \quad f(\boldsymbol{\beta}, \tau_{z^*}) f(Z_i^* | \boldsymbol{\beta}, \tau_{z^*}) f(\boldsymbol{\mu} | Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) f(\boldsymbol{\omega}, \gamma_1 | \boldsymbol{\mu}, Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) f(\nu, \gamma_2 | \boldsymbol{\omega}, \gamma_1, \boldsymbol{\mu}, Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) \\
&\qquad \times f(\mathbf{\Lambda}_1, \tau_1 | \nu, \gamma_2, \boldsymbol{\omega}, \gamma_1, \boldsymbol{\mu}, Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) f(\mathbf{\Lambda}_2, \tau_2 | \mathbf{\Lambda}_1, \tau_1, \nu, \gamma_2, \boldsymbol{\omega}, \gamma_1, \boldsymbol{\mu}, Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) \\
&\qquad \times f(t_i^I | \mathbf{\Lambda}_2, \tau_2, \mathbf{\Lambda}_1, \tau_1, \nu, \gamma_2, \boldsymbol{\omega}, \gamma_1, \boldsymbol{\mu}, Z_i^*, \boldsymbol{\beta}, \tau_{z^*}) \qquad (4.9)
\end{aligned}$$

where $f(\cdot|\cdot)$ denotes a conditional distribution. It is natural to incorporate additional conditional independencies so the joint distribution in (4.9) simplifies to the hierarchical model

$$\begin{aligned}
f(\boldsymbol{\beta}, \tau_{z^*}) f(Z_i^* | \boldsymbol{\beta}, \tau_{z^*}) f(\boldsymbol{\mu} | Z_i^*) f(\boldsymbol{\omega}, \gamma_1 | Z_i^*) f(\nu, \gamma_2 | Z_i^*) \times \ldots \\
\times f(\mathbf{\Lambda}_1, \tau_1 | Z_i^*) f(\mathbf{\Lambda}_2, \tau_2 | Z_i^*) f(t_i^I | \mathbf{\Lambda}_1, \tau_1, \mathbf{\Lambda}_2, \tau_2, Z_i^*)
\end{aligned}$$

To facilitate computation of probit models, we use data augmentation as outlined by Albert and Chib (1993). Briefly, their algorithm involves sampling latent outcome variables from the truncated Normal distribution, with the truncation conditional on the observed dichotomous outcome variable. Details are provided in Appendix A. This approach connects a probit regression model on an observed dichotomous outcome with a normal linear regression model on the continuous latent outcome.

We also utilize parameter expansion, a technique in which non-identifiable parameters are added to a model, to improve computational performance while providing truly non-informative priors (Gelman 2004). Specifically, we redundantly multiply the random effects Z_i^* in models (4.7) - (4.8) by the parameter ξ. Simulations indicate (Gelfand and Sahu 1999; Gelman et al. 2003) that placing a non-informative, proper prior on ξ significantly decreases the autocorrelation of estimable parameters. Gelman (2006) also shows that the using conjugate priors for ξ (Normal) and τ_{z^*} (Gamma) is superior to the usual choice of gamma(ϵ, ϵ) on τ_{z^*} while fixing $\xi = 1$. In the later case, inference can be very sensitive to the choice of ϵ and the prior distribution is often not vague as desired. This expansion scheme is adapted from a similar approach used for the EM algorithm by Liu et al. (1998).

We use a MCMC algorithm programmed in Matlab for posterior calculations. Gibbs sampling (Gelfand and Smith 1990) proceeds by iterating the complete conditionals given in Appendix A until convergence is established.

4.5 Application

4.5.1 Dataset

We apply our methods to 2029 women with singleton pregnancies enrolled in the RFTS study mentioned in section 4.2. Ultrasonographers measured fetal pole lengths on 2005 subjects and gestational sac diameter diameters on 1007 subjects. All subjects had at least one of these two readings so we could determine individual developmental progress and report the date of their LMP. Twenty-one subjects had only a gestational sac present (state 1), 15 had a fetal pole (state 2), and the remaining 1993 subjects had a fetal pole with normal cardiac activity (state 3) at the time of their ultrasound. Sixty-three women who had viable pregnancies and were in state 3 went on to have a loss by week twenty.

Human pregnancies are divided into three trimesters, each normally lasting approximately 12-14 weeks. During the embryonic period, which begins with fertilization and lasts for eight weeks, the embryo may be particularly susceptible to chemical and environmental insults (Kiely 1991). Almost every first trimester ultrasound in RFTS was performed during this important period. From the early ultrasound, we specify states based on the presence or absence of key developmental features including the gestational sac, fetal pole, and cardiac activity. We also use measurements obtained from the ultrasound that have been shown to be excellent predictors of menstrual age. For example, Hadlock and colleagues (1992) used a fourth order linear regression model in which crown-rump length (CRL) was able to predict 98.6% of the variation in the natural logarithm of gestational age. Their results are similar to predictions reported by other authors who often used only linear or quadratic CRL effects. The gestational sac is also believed to be very precise for predicting menstrual age in early pregnancies (Filly and Hadlock 2000).

4.5.2 Bayesian Prior Specification

Bayesian methods are particularly suited to modeling embryonic development because important prior information can be incorporated in the analysis. We use results from the Early Pregnancy Study conducted by Wilcox et al. (1995), which analyzed urinary biomarkers from a group of 221 women to precisely date the length of time from LMP to clinical pregnancy (Wilcox et al. 2001). Due to its high cost and participant burden, it would very difficult to replicate these procedures in large exposure studies like RFTS. The Wilcox study provides day-specific estimates of the probability of clinical pregnancy on a given day of the menstrual cycle, conditional on reaching that day of the cycle (t^I), used in specifying the multinomial prior distribution depicted in Figure 6.

We do not have prior information that we can directly apply to the state transition rates ($\alpha_{ij}^{(k)}$), but we do have information that allows us to specify informative priors on functions of these parameters. Hadlock and others developed models that predict menstrual age in early pregnancy using crown rump length and gestational sac diameter measurements (Filly and Hadlock 2000; Hadlock et al. 1992). They show that the fetal pole, without normal cardiac activity, can be visualized when the crown rump length is 2mm, which occurs at 5.7 menstrual weeks (95% CI: +/- 3 days). Normal cardiac activity begins a few days later, by the sixth menstrual week. Hadlock provides us with prior information about the state progression probabilities for a fetus growing at a normal rate, $\boldsymbol{\omega}$ and ν. For each ω_j, we used independent Normal priors with mean specified to correspond with regression splines consistent with Hadlock's results. The next transition to a normal heart rate occurs quickly, so we used a $N(1.5, 0.1)$ prior for ν. We also explored models with weaker prior assumptions.

For all other parameters where we do not have prior information, we used proper but appropriately vague priors. For each of the regression parameters $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$, $\boldsymbol{\beta}$, γ_1, γ_2, ξ, and $\boldsymbol{\mu}$ we used independent $N\left(\mu_p, \sigma_p^2\right)$ prior distributions, with $\mu_p = 0, \sigma_p^2 = 100$. For precision parameters τ_{z_1}, τ_{z_2}, and τ_{z^*} we use Gamma(a, b) priors, with $a = b = 0.5$.

4.5.3 Analysis

We performed our analysis by iterating the MCMC algorithm given in Appendix A. We monitored parameter autocorrelations and used a variety of diagnostics from the CODA package in the R statistical program (R Development Core Team 2004). All parameters except γ_1 converged quickly, so we used one million iterations to ensure that γ_1 also satisfied convergence diagnostics. Simulation work indicates that γ_1 converges much more quickly when there are more subjects observed in the early states than we had in RFTS.

Figure 7 summarizes the posterior means by subject of the latent growth variable $\boldsymbol{Z}^*$. We did not find any association between the mean of latent growth and male gender, black ethnicity of the mother, or being multiparous. Using (4.6) and the parameter expanded version of (4.7) and (4.8), these associations are measured by examining the estimable function $\theta = \xi\beta_1$. For the association with gender, we found $\Pr(\theta > 0) = 0.21$, for ethnicity $\Pr(\theta > 0) = 0.37$ and in the parity model, $\Pr(\theta > 0) = 0.25$.

We did find an association between latent growth and both state transition rates. Table 1 summarizes the posterior means and 95% credible intervals for γ_1 and γ_2 and indicates that subjects with large values of Z_i^* make transitions more quickly than subjects with relatively small values. To more easily interpret these results, we identify a slow-, normal-, and fast-growing individual based on the 5th, 50th, and 95th percentiles of $\boldsymbol{Z}^*$ and compared their estimated transition probabilities graphically. Figure 8a displays the day-specific conditional probability of moving from state $1 \rightarrow 2$ for these subjects. Our results are consistent with studies from Hadlock et al. (1992) which indicate that a normal fetus should develop a fetal pole around day 26 after conception. When the transition rate is approximately 0.5, the slow and fast growing subject are separated by three days. Figure 8b shows the cumulative probability of having made a transition from state 2 → 3. Most subjects make this transition quickly, but the transition may take longer in slow-developing subjects. Using an early ultrasound to date the pregnancy may be inaccurate because this method assumes that subjects have the same early growth rate.

Fetal development is traditionally dated starting from then LMP and then assumes that conception occurs exactly two weeks later. However, there can be considerable variability among women in time from LMP to conception (t_i^I). Figure 6 gives the prior and posterior distributions of t_i^I for two noteworthy subjects and all subjects combined. Overall, the prior and posterior distributions for all subjects are similar. Among individual subjects, there is significant variability in the shape of the posterior distributions of t_i^I. For example, subject A has a clear peak around day 10 while subject B is relatively flat from day 15 to 50. Ninety-nine percent of subjects have a within-subject variance of t_i^I that is more similar to subject A than subject B, with 7% percent of subjects having a posterior variance greater than 25 days^2. For most subjects, like subject A, we are confident in our estimate of t_i^I and, even for subject B, it is likely that t^I is longer than two weeks.

To evaluate the impact of not estimating t_i^I, we conducted a comparison analysis in which we fixed the time from LMP to conception at two weeks for each subject. In this simpler model, individual variability in the time from LMP to conception is now included in the time to the first state transition. Failing to account for the unknown initiation time alters both the distribution of the latent growth variable, Z^*, and posterior distribution of γ_1. The posterior mean for γ_1 is greatly increased while the posterior distribution for γ_2 changes little (Table 1). Second, some subjects change from having moderate to extreme values of the latent growth variable, Z^*. A post-hoc comparison indicates that, for subjects who would have unusually large t_i^I, fixing t_i^I at two weeks results in large negative value of Z^*. Conversely, subjects who would have small values of t_i^I have large positive values of Z^* in the comparison analysis ($\text{corr}\left(t^I, Z^*\right) = -0.75$). This strong negative correlation is removed when we model the unknown conception time ($\text{corr}\left(t^I, Z^*\right) = -0.04$).

We also found an association between latent growth rate and the probability of having a spontaneous abortion by the 20th week of pregnancy. Embryos that were relatively slow growing were more likely to subsequently spontaneously abort. Specifically, we found that a one standard deviation increase in latent growth is associated with a -2.11 unit change in the probit of the probability of having a SAB (95% credible interval [-2.42, -1.90]). According to the model, approximately 1.7% (95% CI [0.8%, 2.9%]) of embryos with an average value of Z_i^* will be lost before the 20th week while 3.8% (95% CI [2.4%, 5.4%]) of embryos with a value of Z_i^* one standard deviation below the mean will be lost. We note that this analysis included only embryos that were viable at the time of the ultrasound so that the pregnancy loss occurred after the ultrasound.

Because our primary inference is about a latent variable and we make a number of distributional assumptions, it is important to evaluate the fit of our model. To do so, we calculated the posterior predictive distributions for observed data and then compared these distributions to our observed data following Lynch and Western (2004). Figure 9 shows the posterior distributions and observed data for fetal pole length by the observed time from LMP to ultrasound. This plot represents 65 subjects, with one subject randomly chosen for each unique value of time from LMP to ultrasound. We also include the posterior predictive distribution that would be obtained using the simplified analysis that assumes that time from LMP to conception is fixed at two weeks. Figure 9 clearly indicates that our model fits observed fetal pole lengths well, and is superior to the simpler analysis option. Model evaluation using gestational sac diameters yields the same conclusions.

4.6 Discussion

In this article we describe a Bayesian approach for analyzing a multistate growth process where the initiation time is not known. We use an example from human embryonic growth and pregnancy loss both to motivate the methodology and serve as an example analysis. Our analysis is complicated by only having cross-sectional data so that initiation and transition times are not observed. An ideal study would have repeated measures on each subject with a known time of initiation. However, factors including cost and participant burden make such studies difficult to conduct. A similar study that determines state and measures developmental progress variables at one time point could use our analysis approach.

In our early embryonic growth example, most subjects are observed to be in state 3 while an optimal study would observe more subjects in early states. We expected few subjects in states 1 and 2 due to difficulties in achieving very early ultrasound and the short waiting time in state 2. Subjects in state 3 do provide censored transition times for both the state transitions as well as the developmental progress covariates which are indicative of the waiting time spent in the state. Our Bayesian approach allows us readily incorporate respected prior information on the baseline transition rates. For example, in Figure 8a, the general S-shape of the curve is stabilized by prior information from Hadlock et al. (1992) while differences among subjects are based on the data and a diffuse prior.

Our primary goal was to identify embryos that are developing relatively quickly and relatively slowly using a latent growth model. The latent variable approach has advantages over previous methods that use developmental progress variables directly. In the direct approach, it is often not clear how different surrogates should be combined which makes any final classification system somewhat arbitrary. Past research indicates that that short gestational sac diameters (Nyberg et al. 1987) or fetal pole lengths (Mantoni and Pedersen 1982) may be associated with increased risk of pregnancy loss. However, a more recent study (Brizot et al. 2001) could not replicate either result. Rather than consider these developmental progress variables directly, we conceptualize that they are indicative of underlying embryonic growth. Our approach automatically differentiates individuals using one overall, continuous measurement regardless of partial missingness, measurements being taken at different times, or unknown initiation times. We can then incorporate the latent variable in joint models for multistate growth processes and future events.

Applying our methods to the RFTS study led to new insights about early pregnancy. We were able to identify individual embryos with different growth rates while accounting for the variable time to conception and time to ultrasound. Our comparison analysis indicates that failing to model the unknown initiation time has a large impact on the model parameters. In many applications, the latent variable is a nuisance parameter, but it is an important part of our analysis. Gender, ethnicity and parity have been shown to be possible predictors of growth starting in the 25th week of pregnancy (Zhang and Bowes 1995), but we were not able to confirm that latent growth is modified by these covariates during the first trimester. We were able to find evidence in favor of a previously hypothesized but unproven association between slow growth early in pregnancy and increased risk of future pregnancy loss.

Model	Parameter	Mean	Median	Lower 2.5%	Upper 97.5%
t^I estimated	γ_1	0.63	0.60	0.32	1.13
	γ_2	1.00	1.03	0.28	1.37
t^I fixed	γ_1	4.09	4.09	3.84	4.37
	γ_2	1.03	1.07	0.17	1.42

Table 1: Posterior summaries of the parameters characterizing the association between $\boldsymbol{Z}^*$ and state transition probabilities. The proposed model, t^I estimated, is compared to a simpler analysis that fixes t^I at two weeks.

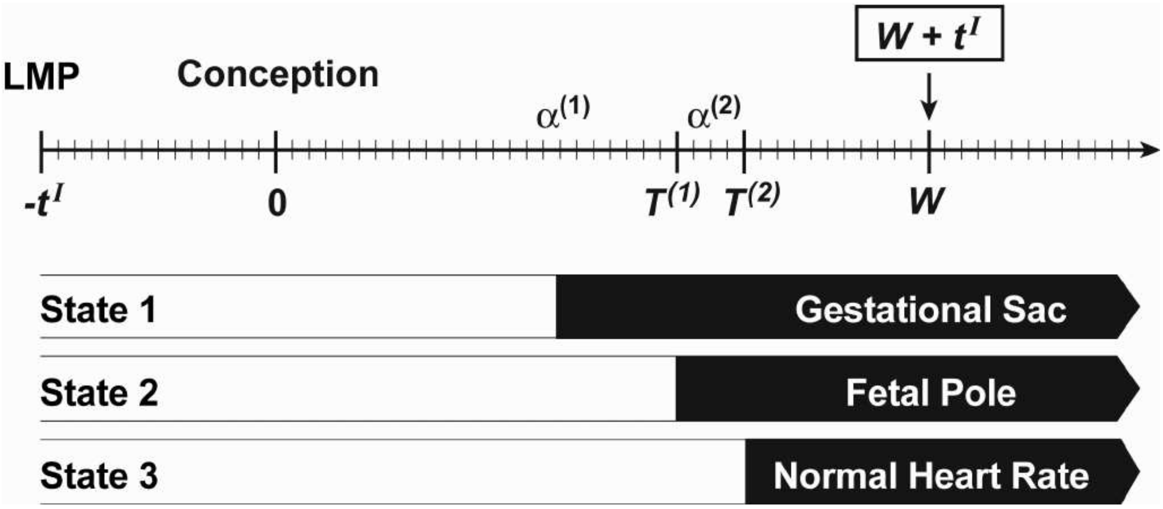

Figure 4: Timeline for a subject who has a fetal pole with normal heart rate at time W. W is not observed but $W + t^I$ is known.

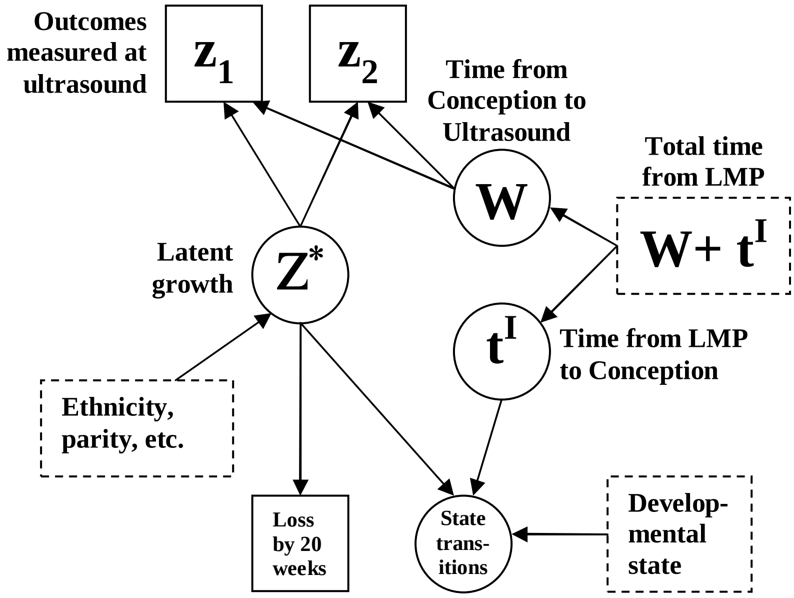

Figure 5: Path diagram illustrating the dependencies in the fetal growth model. Circles represent latent variables, squares indicate observed variables, and arrows show association.

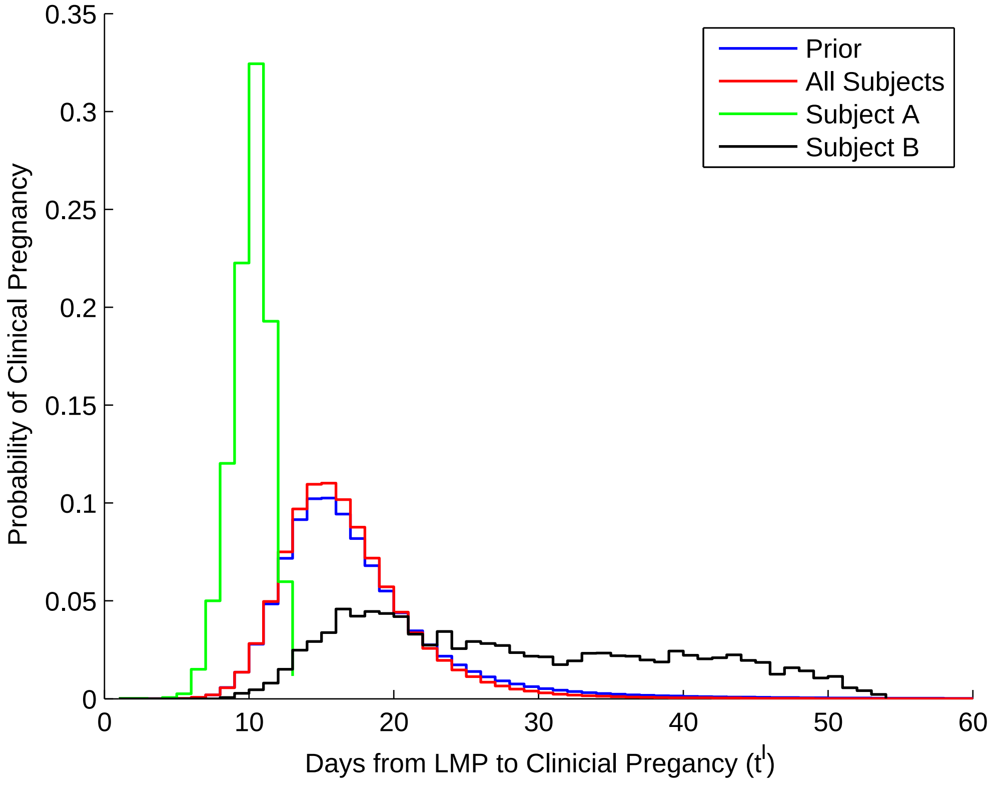

Figure 6: Prior and posterior distributions of the probability of conception on a given day of the menstrual cycle, conditional on reaching that day of the cycle. The green and black lines represent posterior distributions for individual subjects, and the red line all subjects combined. The vast majority (99%) of individual subjects had posterior distributions with a variance more similar to subject A than subject B.

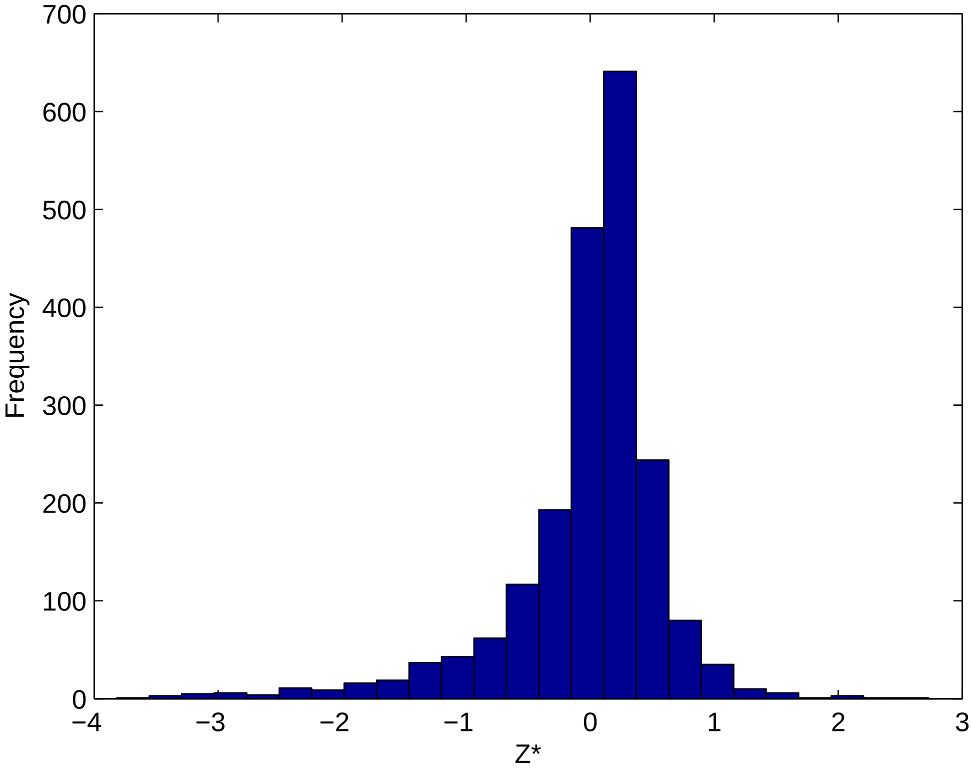

Figure 7: Distribution of the posterior means of $\boldsymbol{Z}^*$.

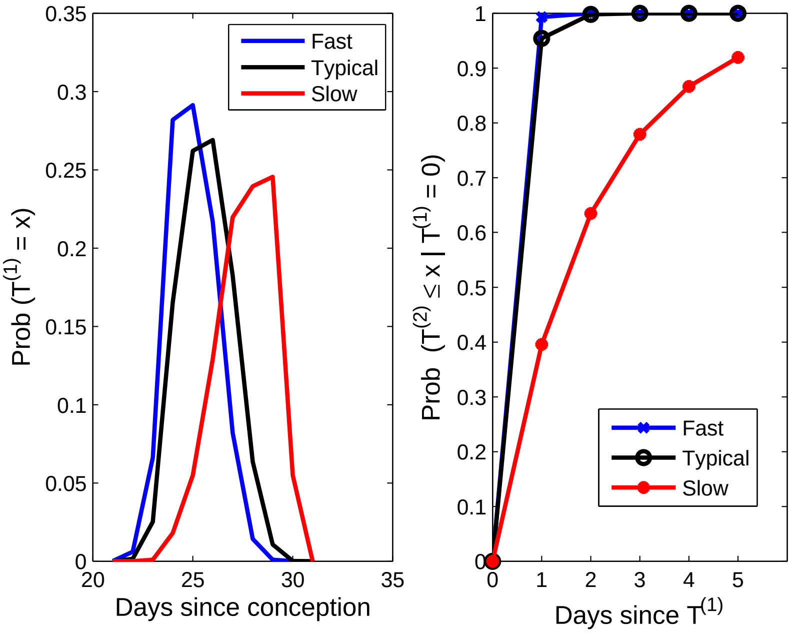

Figure 8: Discrete transition rates from a gestational sac to fetal pole ($1 \rightarrow 2$, a) and cumulative probability of observing a normal heart rate ($2 \rightarrow 3$, b) for three subjects.

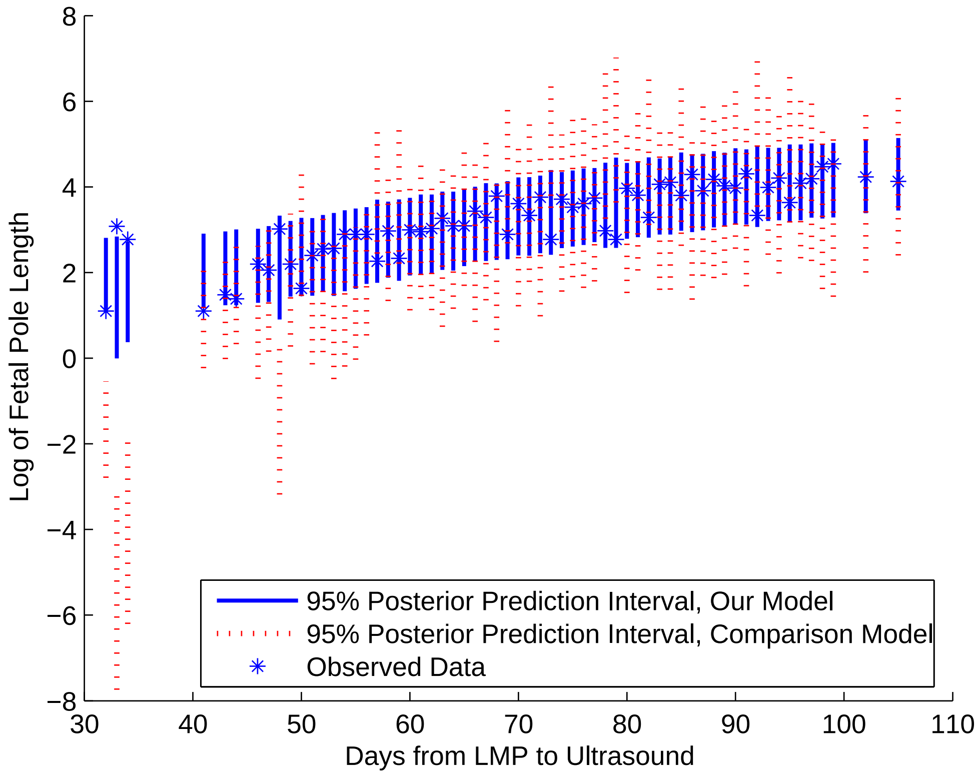

Figure 9: Posterior prediction intervals and observed fetal pole lengths for 65 subjects.

5 A Mixture Model for Birth Weight and Gestational Age

5.1 Abstract

The distributions of birth weight and gestational age at delivery are heavily skewed and can be described as arising from a predominant component and a residual component that is indicative of immature birth (Wilcox et al. 2001). A baby from the residual component will, on average, have lower birth weight, earlier gestational age, and greater risk of mortality than a baby from the predominant component. We propose a Bayesian latent variable mixture model to identify subjects in the residual component by jointly modeling birth weight and gestational age at delivery. Our methods do not rely on predefined cut points, like pre-term birth or low birth weight, and account for uncertainty when classifying babies into the high risk group. Additionally, we use latent variable models to summarize first trimester ultrasound growth measurements in order to estimate individual early fetal growth restriction. In pregnancies that proceed to term, our results indicate that growth restriction found in the first trimester is associated with both lower birth weight and growth restriction at delivery. Furthermore, we found significant associations between belonging to an immaturity latent group and smoking status, race, and education level, but no associations with maternal age or parity. An analysis using our approach may be more powerful than an analysis that uses pre-term birth and very

pre-term birth, respectively, as the outcomes of interest.

5.2 Introduction

In reproductive epidemiology, birth weight and gestational age at delivery have been extensively studied due to their strong association with infant mortality and weaker associations with other morbidity events. Birth weight and gestational age are inherently related in that early gestational age is strongly associated with lower birth weight. Similarly, growth measurements obtained by an ultrasound on the same fetus, such as the fetal pole length and gestational sac diameter, are positively correlated. In this paper, we develop latent variable methods that use these correlated outcomes to jointly model: (1) birth weight and gestational age, which allows us to classify babies born to a immaturity latent class, and (2) first trimester ultrasound measurements, in order to address early growth restriction.

Both gestational age and birth weight follow primarily normal distributions with extended lower tails (the “residual” distribution) in the direction of early births and low birth weights, respectively (e.g. Figure 10). The vast majority of births lie within the predominant portion of the distributions while, for birth weight in particular, about 2% to 5% of births have been found to lie in the residual component of the distribution (Wilcox et al. 2001). A population with a higher proportion of births in the residual component will be at an increased risk for infant mortality and later morbidity outcomes such as diabetes (Harder et al. 2007; Shan and Ohlsson 2002). Conversely, two populations may differ in their predominant distributions without a corresponding change in outcome. For example, the predominant birth weight distribution for Mexican-American babies is shifted to the left compared to US non-Hispanic white babies, but Mexican-American babies have better overall survival (Buekens et al. 2000; David 1980).

Epidemiologists recognize the excess risk associated with belonging to the residual component of the birth weight and gestational age distributions, and they have attempted to classify subjects into the high risk group by defining specific cut points. In particular, gestational age less than 37 weeks (pre-term birth) or less than 32 weeks (very pre-term birth) have been commonly used as outcomes in reproductive epidemiology. Similarly, birth weight thresholds at 1500 g (very low birth weight) and 2500 g (low birth weight) are thought to be particularly indicative of postnatal complications. However, pre-defined thresholds are somewhat arbitrary and prone to misclassify babies from the predominant component as high risk while other babies from the residual component may be incorrectly placed in the normal risk group (Wilcox et al. 2001). We consider an alternative approach using mixture distributions to identify a latent class of subjects who belong to the residual component. Our methods do not rely on pre-defined cut points but still allow us identify a group of fetuses who are more likely to deliver early. We then propose a model for examining the association of belonging to the immaturity class with several covariates.

Early identification of growth restriction is important in order to reduce the morbidity and mortality associated with this problem. During pregnancy, a fetus is considered growth restricted when ultrasound measurements deviate below a specific gestational age threshold (Maulik 2006). Similarly, at birth, a baby is defined to be growth restricted if his or her birth weight is below a specified percentile of weight for gestational age. Although there is controversy as to how to calculate the percentiles, in general a baby is identified as small for gestational age (SGA) when his or her weight falls below the 10th percentile of weight for a given gestational age at birth. Ultrasound measurements can be used to provide an accurate and sensitive method for identifying fetuses that are growing more slowly than expected and more likely to be SGA (Ott 2006). Rather than use the ultrasound measurements directly, we propose a latent variable model that aggregates multiple measurements to identify fetuses that have an underlying latent tendency to be growth restricted. In particular, we use the fetal pole length and mean gestational sac diameter obtained at exactly one time point early in the first trimester to estimate latent early growth restriction. We then examine the association of the early growth restriction latent variable with birth weight and growth restriction measured at birth.

5.3 Methods

5.3.1 Statistical Methods

Latent variables are commonly used in the social sciences as a means of quantifying an unobservable concept based on many observed variables (Bollen 1989). In our fetal development analysis, we observe four outcomes that we summarize with two latent variables. The gestational sac diameter (y_{i1}) and fetal pole length (y_{i2}) are measured during a first trimester ultrasound while birth weight (y_{i3}) and gestational age at delivery (y_{i4}) are collected at birth. We model the correlation between the outcomes using latent variables representing early growth restriction (η_{i1}) and the tendency to be born immaturely (η_{i2}). While latent variables can be difficult to interpret directly, in general subjects with small values of the "immature" latent variable (η_{i2}) will have small birth weight and early gestational ages. Similarly, subjects with a smaller than average fetal pole length and gestational sac diameter, controlling for the number of days since the last menstrual period (LMP), will have small values of the "growth restriction" latent variable, η_{i1}.

The path diagram in Figure 11 displays the relationships among observed and latent variables and modeling assumptions in our proposed model. We use the symbols of Sanchez et al. (2005) in which ovals represent latent variables and rectangles represent observed variables with (solid lines) and without (dashed lines) distributional assumptions. Arrows are used to represent associations between variables so that the lack of an arrow indicates conditional independence. Figure 11 shows that the latent growth restriction variable models the correlation among the ultrasound measurements, and we also test possible associations of early growth restriction with birth weight and gestational age at delivery. The latent immaturity variable is a continuous variable that captures the correlation between birth weight and gestational age. A subject in the residual component of the latent immaturity distribution will have low weight, early age, and a small value of the latent immaturity variable. We formalize the residual component concept using a mixture distribution model that allows the mean and variance of latent immaturity to change as a function of a categorical latent variable, immaturity class (T_2). All subjects in the residual component will be assigned to the same immaturity class, while subjects in the predominant component will be placed in one of two classes indicative of normal maturity. We then use categorical regression to estimate the association of black race, current smoking status, maternal age, parity, and education level with belonging to the residual group.

Mixture distributions can be useful for flexibly representing a wide variety of distributions and are particularly appropriate when observations are believed to have arisen from one of many distinct groups (Diebolt and Robert 1994). Gestational age at delivery follows an asymmetric statistical distribution with a long tail to the left for early births and a near truncation on the right due to labor being medically induced for very long gestations. Birth weight also follows a heavily skewed distribution as some babies are prone to be born with low weight. While a single normal distribution would not adequately capture either of these unusual distributions, Gage (2002) found that birth weight and gestational age could be modeled using two-component normal or log normal mixtures. Wilcox et al. (2001) also suggests that birth weight arises from independent predominant and residual components. A natural extension to our analysis is to use a mixture of normal distributions to flexibly model the latent immaturity variable. In particular, we use a three class mixture in which one class identifies babies belonging to the residual component that is indicative of immaturity and two classes are used to model the predominant distribution indicative of normal maturity.

All parameters are estimated simultaneously in a joint model to ensure proper statistical inference. Fokoue (2005) has proposed an EM algorithm for fitting normal latent variable mixture models with covariates, but these methods are not readily available in statistical software. We prefer a Bayesian approach that uses Gibbs sampling (Casella and George 1992) to iterate the complete conditionals given in Appendix B until convergence. Our approach for fitting finite mixture distributions is based on the work of Diebolt and Robert (1994), but is complicated by needing to model latent rather than observed variables. To model the association of latent group membership with covariates, we utilize data augmentation. In this procedure, latent data augments the observed data so we can use both the latent and observed data to calculate the posterior distribution of the parameters of interest (van Dyk and Meng 2001). Model estimation details for similar Bayesian latent variable models are given by Elliott et al. (2005) and in the third paper in this dissertation.

5.4 Application

5.4.1 Background

We use a Bayesian procedure for fitting the proposed latent variable mixture model. While a thorough treatment of a Bayesian approach to data analysis (c.f. Carlin and Louis 2000; Gelman et al. 2004) is beyond the scope of this paper, we provide a brief introduction here to aid the reader in interpreting our results. Bayesian methods are based on determining the joint distribution of all parameters in the model given the data, called the posterior distribution. We summarize the posterior distribution of a parameter by quantities including the posterior mean, 95% credible interval (CI), and posterior probability. Although different conceptually, under many circumstance these posterior summaries are very similar numerically and analogous to the more commonly-seen maximum likelihood parameter estimate, 95% confidence interval, and probability (e.g. p-value) from a frequentist analysis.

5.4.2 Dataset

We apply our methods to 1707 subjects enrolled in the Right From the Start (RFTS) study of early pregnancy (Promislow et al. 2004). As soon as possible after enrollment (average of 66 days after LMP), each mother was brought into the clinic to receive an early first trimester ultrasound. At the ultrasound, at least one of the fetal pole length or gestational sac diameter were measured. Fetal pole lengths were measured for 1693 subjects, and gestational sac diameters were available for 827 subjects. All subjects also provided information on the birth weight and gestational age of the child as well as maternal age, race, education level, parity, and smoking status covariates. The covariates were dichotomized as maternal age being greater than or equal to 35 years, race being black, education level being above high school, being multiparous, and being a current smoker.

Table 2 provides the frequencies of categorical variables as well as means and standard deviations for several continuous variables. We include all singleton, live-born infants from the RFTS-I study in this analysis. As in the general population, the distributions of birth weight and gestational age in the RFTS sample had extended lower tails, with a nearly truncated upper tail for gestational age (Figures 13 and 14). Only 5.8% of babies in this sample were classified as SGA when, by definition, 10% of the referent population that defines the norm will have SGA babies.

5.4.3 Results

To perform the analysis, we used Gibbs sampling conducted in Matlab to iterating the complete conditionals given in Appendix B (Casella and George 1992). We ran five chains from disparate starting values and monitored convergence using the CODA package for R (R Development Core Team 2004). After removing an initial burn in of 15,000 iterations, all parameters were judged to have converged by a variety of diagnostic measures. For example, all Gelman-Rubin statistics ($\hat{R}$) were found to be less than 1.01, where $\hat{R} = 1$ at convergence and values less than 1.2 are generally considered sufficient for convergence

(Gelman et al. 2004). We used the remaining 35,000 iterations for inference.

Our estimate of the distribution of the immaturity latent variable, including the predominant and residual components, is presented in Figure 10. The predominant component is represented by two normal distributions, and the residual component is modeled using one normal distribution. In our sample, 3.3% of subjects (95% CI $= [1.5\%, 5.8\%]$) belong to the residual component indicative of immature birth. To ease interpretation of the latent immaturity variable, we specify necessary identifiability restrictions so that latent immaturity has the same location and scale as gestational age at delivery. On this scale, the residual component of the distribution has a posterior mean of 227 days and standard deviation of 24 days. The predominant components, both of which are indicative of normal-term pregnancies, are less variable than the residual component and have posterior means of 264 and 279 days. A subject from the residual component also has a noticeably smaller birth weight than a subject from the predominant component. On the weight scale, a subject with a typical rate of early growth ($\eta_{i1} = 0$) and in the residual component has an expected weight of 1901 (95% CI = [1294, 2428]) grams. A subject in the predominant distribution has an expected weights of 3061 (95% CI = [2866, 3241]) grams or 3546 (95% CI = [3506, 3587]) grams.

As shown in Figure 11, we use the early growth restriction latent variable to model the correlation between the fetal pole length and gestational sac diameter as well as allowing growth restriction to influence later birth outcomes. Table 3 displays the estimated change in outcome variables for a one standard deviation increase in growth restriction. Subjects with smaller values of early growth latent variable (indicating more growth restriction) will have shorter than average fetal poles lengths and gestational sac diameters, after controlling for time since LMP. Additionally, a one standard deviation decrease in the early growth restriction latent variable is associated with a 164 (95% CI: [136, 191]) gram decrease in birth weight, controlling for the latent immaturity variable. We found no such association between early growth restriction and gestational age.

We also examined the association between various ranges of latent early growth restriction and probability of growth restriction measured at birth. Birth weight Z-scores were calculated using birth weight for gestational age population standards provided by Oken et al. (2003). Specific cut points for defining growth restriction are arbitrary, so we present results using the 5th and 10th percentile of the Z-score and several cut points for early growth restriction in Figure 12. Early growth restriction follows an approximately normal distribution with a mean of zero and variance of one, so cut points of -2, -1, -1.5, and 0 roughly correspond to the second, seventh, sixteenth and fiftieth percentiles, respectively. Using the 10th percentile of birth weight Z-score as the cut point (i.e. SGA), subjects with a value of early growth restriction of -2 or lower have an 18% posterior probability of being SGA at birth while subjects with positive values have only a 4.3% risk. In general, subjects with early growth restriction less than -1 are at an increased risk for growth restriction at birth while subjects with positive values are at decreased risk for later growth restriction. When early growth restriction is in the range $(-1, 0]$, a subject has a posterior probability of future growth restriction approximately equal to the SGA percentile cut point used. For example, at the 5th and 10th percentile cutoffs for birth weight Z-score, we would expect 5% and 10% of subjects to be growth restricted, respectively. For these cutoffs, the model estimates that subjects with early growth restriction in $(-1, 0]$ have a 5.4% and 9.9% posterior probability of being growth restricted at birth.

Our latent variable mixture distribution approach fits the observed birth weight and gestational age distributions well. Figure 13 depicts the observed and estimated cumulative distribution functions (CDFs) for gestational age at delivery while Figure 14 displays these CDFs for birth weight. Tail areas are magnified in the figures. Particularly interest lies in identifying fetuses that have a latent tendency to be born the most early and with the least weight, so we assigned these subjects to a specific latent class. To help understand the type of subjects who belong to this immaturity class, we include conditional probability plots in Figures 13 and 14. *A priori*, each subject had an equal probability of belonging to the residual distribution, but the posterior distribution of class membership is strongly related to birth weight and gestational age. For example, Figure 13 includes the cumulative probability of being assigned to the immaturity class, conditional on gestational age. All subjects born before 224 days (very pre-term) and approximately 30% of subjects born before 259 days (pre-term) are assigned to the immaturity latent class. Similarly, almost all subjects with very low birth weight ($<$ 1500g) and 45% of subjects with low birth weight ($<$ 2500 g) belong to this latent class (Figure 14).

Finally, we found several associations between covariates and belonging to the residual distribution. In unadjusted models, current smokers are 3.08 times (95% CI = [1.08,6.55]) more likely to belong to the residual distribution than former- or never-smokers. Black race (OR = 2.63, 95% CI = [1.38,4.83]) and education (OR = 2.14, 95% CI = [1.13,3.82]) were also associated with belonging to the residual distribution while the posterior odds ratio credible intervals for being multiparous (OR = 1.15, 95% CI = [0.66,2.22]) and advanced maternal age (OR = 1.93, 95% CI = [0.82,3.83]) all contain one. In a multivariate regression model that jointly includes all five covariates, the bivariate associations for smoking and black race remained significant. The parameter estimates from the multivariate model are presented in Table 4.

Table 4 also contains a comparison of our latent variable mixture model analysis with the parameter and standard error estimates from models using pre-term birth (PTB) or very PTB as the outcome of interest. All models use the probit link function to facilitate comparison. For covariates that are statistically significant, posterior means from the latent class analysis are larger than from a PTB analysis with standard deviations that are smaller than the very PTB analysis. In the multivariate latent variable model, there is a borderline significant association between achieving a high school education and belonging to the residual distribution, which does not appear in either the PTB or very PTB analysis. Finally, for parity, there is a borderline negative association with PTB and an insignificant positive association with very PTB, but little evidence in favor of any previous live births being associated with belonging to the residual distribution.

5.5 Discussion

In this paper we propose a latent variable mixture model to identify babies that belong to residual component of the distributions of birth weight and gestational age at delivery. We found that approximately 3.4% of babies in our population can be classified as belonging to the residual component, which is in the 2% to 5% range for birth weight reported by Wilcox et al. (2001). However, the RFTS study of early pregnancy is not a random sample of pregnant women from the general population, so it is possible that the residual component frequency in the general population could be different. Babies in the residual component were found to be born approximately five to seven weeks earlier and weigh 1100 to 1600 grams less than babies in the predominant distribution. We found that current smoking status and black race were associated with increased risk of belonging to the residual component in both adjusted and unadjusted models, and a possible association with education as well. Finally, growth restriction estimated using a early first trimester ultrasound measurements is associated with growth restriction at birth.

Our model is conceptually different from the traditional approach in that it allows for subject-specific probabilities of being assigned to the residual component. Other approaches assume that the high risk group membership is known by using pre-defined cut points such as 37 weeks (PTB) or 32 weeks (very PTB). If the goal is to identify the group of subjects in the residual component, using pre-defined cut points fails to recognize that there is error in measuring the outcome. In our approach, lower birth weight and lower gestational ages increase the probability of being in the residual component of the distribution (Figures 13 and 14), but we do not assume group membership is known. We model the subject-specific probability of belonging to the residual component while jointly examining the association between covariates and belonging to the residual component indicative of immaturity.

Birth weight and gestational age at delivery are routinely analyzed in reproductive health due to their strong associations with many postnatal complications. Often researchers attempt to find associations between exposures such as air pollution and partner abuse and mean changes in birth weight or gestational age using linear regression models (Curry et al. 1998; Glinianaia et al. 2004). Such models assume that shifts in the mean of birth weight or gestational age will correspond to an increase in the percentage of the population at particularly high risk for complications. However, from our perspective, changes in the mean could be due to a shift in the predominant component of the distribution, or a more concerning increase in the proportion of babies born in the residual component, or a combination of the two. For example, Mexican-American babies on average have lower birth weights compared to US non-Hispanic white babies, but without a corresponding increase in mortality. Buekens et al. (2000) explains this observation as being due to an unimportant shift in the predominant component toward lower birth weights without an increased risk of being born in the residual component of the distribution. Our model focuses on identifying variables that increase the risk of falling in the residual component.

We found an association between the early fetal growth restriction and decreased birth weight, but no association of early growth with gestational age. However, it can be difficult to accurately measure gestational age at delivery, which might influence the interpretation of early growth restriction. Variability in the time from LMP to conception can arise from both misdating the LMP and natural variability in the follicular phase distribution so that, for example, a subject developing at a normal rate but with an abnormally long time from LMP to conception would have a smaller than expected fetal pole or gestational sac. In the model, this type of error would lead to a small value of latent growth restriction, and a longer than expected gestational age at delivery. However, in another analysis of the RFTS dataset, we estimated an early fetal growth rate variable while attempting to account for the unknown time from LMP to conception (Slaughter et al. 2007). A post-hoc analysis indicates that the two early fetal growth variables estimated in these analyses are highly correlated.

Characteristic	Percent	Mean	Std Dev	Min	Max
LMP to ultrasound (days)		66	13	32	105
Birth weight (grams)		3394	582	600	5229
Gestational age (days)		274	14	170	300
Small for gestational age	5.8%				
Pre-term birth (< 37 weeks)	9.6%				
Very pre-term birth (< 32 weeks)	1.2%				
Low birth weight (< 2500 g)	5.2%				
Very low birth weigh (< 1500 g)	1.1%				
Black Race	27.7%				
Education ≤ High School	25.5%				
Maternal age ≥ 35	10.9%				
Parity ≥ 1	52.8%				
Current Smoker	5.3%				

Table 2: Summary statistics for RFTS subjects (N = 1707)

Outcome	Mean	Lower 2.5%	Upper 97.5%
Gest. sac diameter	0.22	0.20	0.24
Fetal pole length	0.34	0.31	0.36
Birth weight	163.8	136.3	191.3
Gestational age	5.5E-6	-1.4E-3	1.4E-3

Table 3: Posterior summaries characterizing the association of a one standard deviation increase in latent growth restriction with observed outcomes.

	Latent		PTB		Very PTB	
Covariate	Mean	SD	Mean	SD	Mean	SD
Intercept	-2.14	0.20	-1.43	0.07	-2.89	0.20
Black	0.37	0.14	0.23	0.09	0.61	0.18
Education	0.22	0.14	0.25	0.10	0.27	0.20
Maternal age ≥ 35	0.35	0.18	0.20	0.13	0.36	0.26
Parity > 0	-0.01	0.13	-0.15	0.08	0.22	0.20
Current Smoker	0.45	0.22	0.39	0.16	0.46	0.29

Table 4: Comparison of proposed latent class analysis with analyses that use pre-term and very pre-term birth as the outcomes. Posterior means and standard errors for covariates are presented for each outcome.

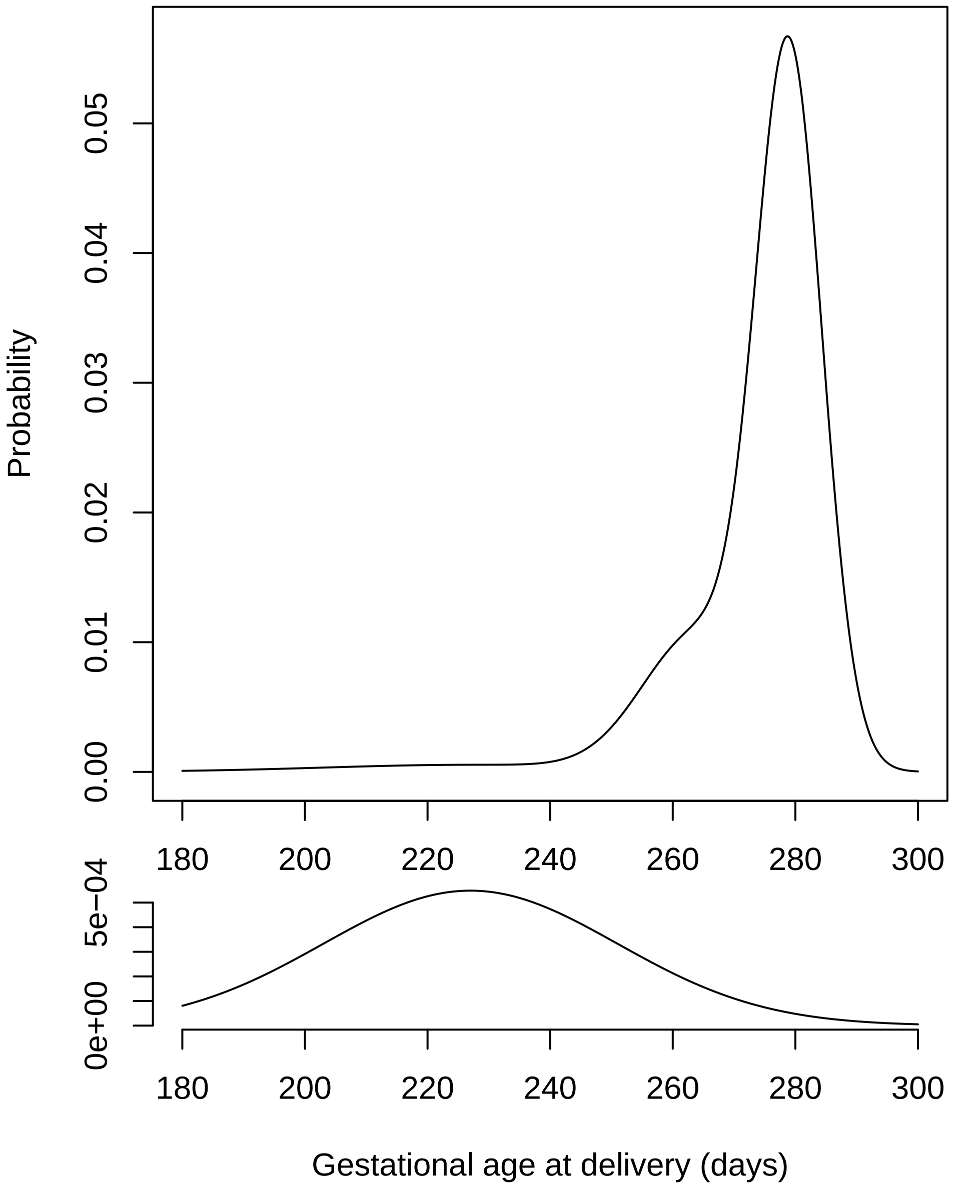

Figure 10: Estimated probability density function for the predominant and residual components of latent immaturity distribution (main plot). The lower plot magnifies the residual component.

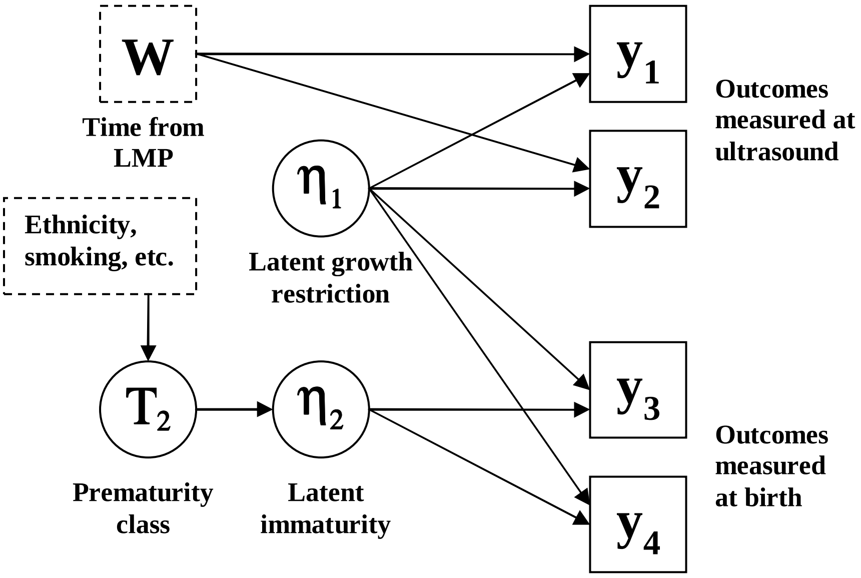

Figure 11: Path diagram representing dependencies in fetal development model. Circles represent unobserved variables and boxes observed variables both measured with error (solid lines) and without distributional assumptions (dashed lines).

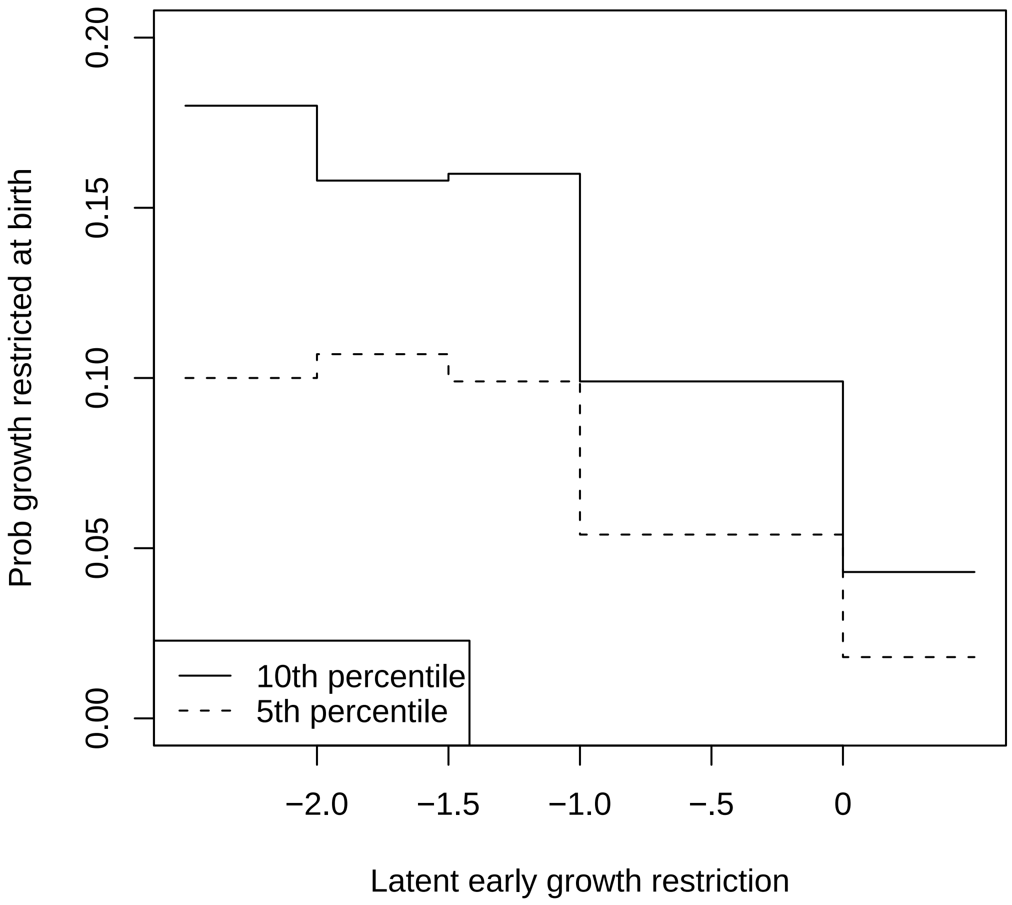

Figure 12: Probability of being growth restricted at birth for various ranges of early growth restriction. Growth restriction at birth is defined using either the 10th (solid lines) or 5th (dashed lines) percentile of birth weight Z-scores.

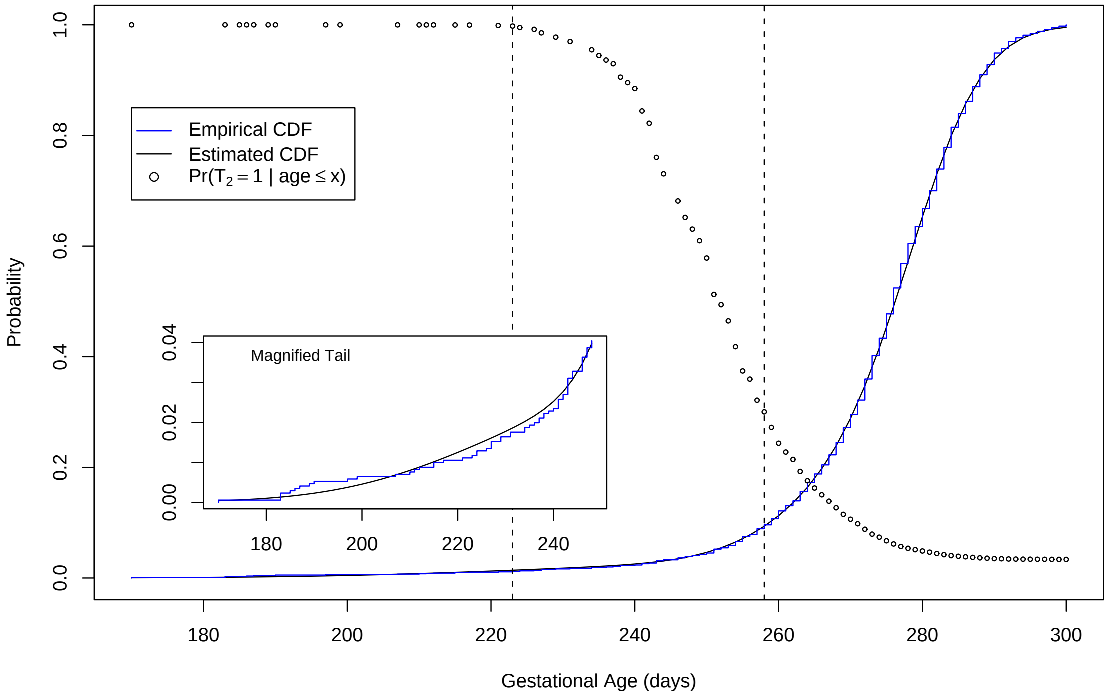

Figure 13: The solid lines represent the empirical and estimated cumulative distribution functions for gestational age at delivery. The fit of the residual distribution is magnified. Points provide the cumulative probability of being assigned to the latent class indicative of immature delivery ($T_2 = 1$), conditional on gestational age. Vertical dashed lines indicate the commonly-used very PTB and PTB cutoffs.

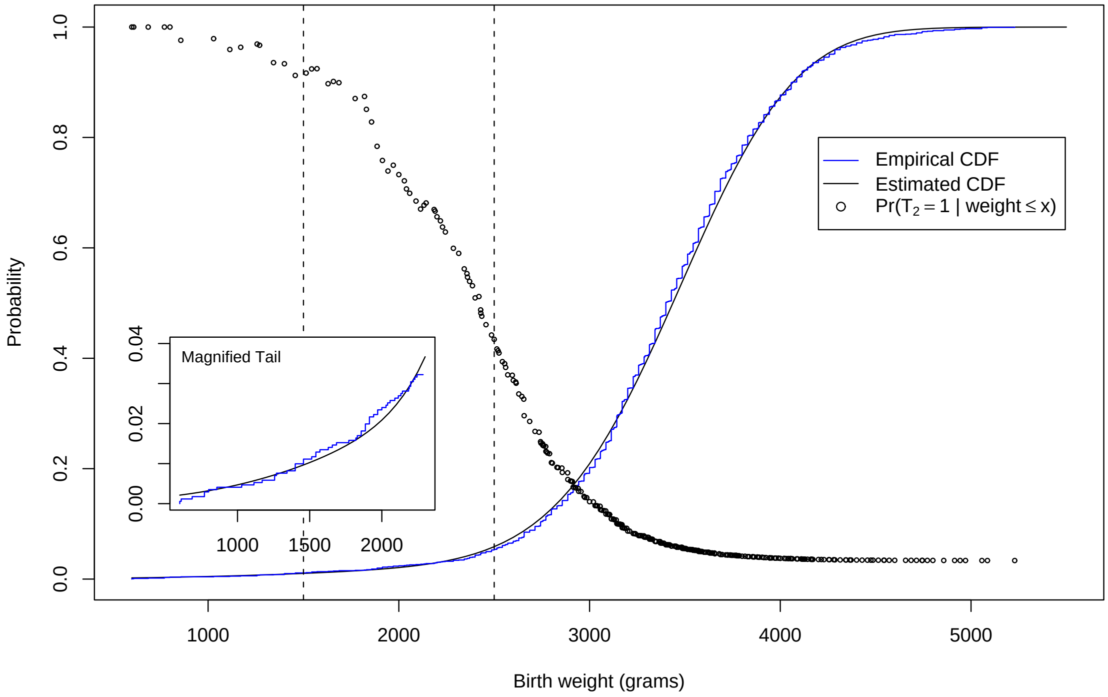

Figure 14: The solid lines provide the empirical and estimated cumulative distribution functions for birth weight. The fit of the residual distribution is magnified. Points represent the cumulative probability of being assigned to the latent class indicative of immature delivery ($(T_2 = 1)$), conditional on birth weight. Vertical dashed lines indicated the commonly-used very LBW and LBW cutoffs.

6 A Bayesian Latent Variable Mixture Model with Covariates for Fetal Growth

6.1 Abstract

Latent variables can be useful for aggregating multiple correlated outcomes that are measured on the same subject into a fundamental, underlying concept. We propose a Bayesian approach for jointly modeling several fetal outcomes, measured by ultrasound during pregnancy, and growth measurements, made at birth, using latent variables. Our procedure is more flexible than typical latent variable methods in that we allow the latent variables to follow finite mixture distributions. Using mixture distributions also permits us to identify and group individuals with similar characteristics. We apply our methods to a study of fetal development in which we use latent variables to identify latent classes of subjects who are more likely to be growth restricted during pregnancy and growth restricted at birth. We then examine the association between measured covariates and latent classes of growth restriction. Our methods are able to identify a latent class of subjects who have increased blood flow restriction and below average intrauterine size during the second trimester who are more likely to be growth restricted at birth than a class of individuals with normal size and blood flow.

6.2 Introduction

In reproductive health, birth weight and gestational age at delivery have been extensively studied because low birth weight and pre-term birth are both strongly associated with infant mortality. Birth weight and gestational age are positively correlated and follow unusual, skewed distributions with long tails in the direction of low weights and early births, respectively. Wilcox et al. (2001) has described the birth weight distribution as arising from a predominant and a residual component, with the residual component containing most of the lower tail. A higher proportion of births in the residual component is associated with an increased risk of poor pregnancy outcomes, but a shift in the mean of the predominant distribution toward lower weight may not be associated with any increased risk (Buekens et al. 2000). With this in mind, we focus on identifying a latent class of subjects who belong to the residual distribution and on finding factors that are associated with class membership.

Early identification of fetuses that are at increased risk for being growth restricted at birth or being born before they are fully mature is important to reduce infant mortality and other morbidity events. Babies are often determined to be growth restricted at birth when their weight is substantially less than would be expected for their given gestational age at delivery. For example, an individual is defined to be small for gestational age (SGA) if his or her birth weight is below the 10th percentile of weight for a given gestational age (Oken et al. 2003). To identify growth restriction during pregnancy, many different ultrasound measurements of fetal size and blood flow may be useful. In particular, the head circumference (HC), abdominal circumference (AC), biparietal diameter (BPD) or femur length (FL) can be used individually to estimate fetal size during the first and second trimester. A fetus will then be considered to be “growth restricted” when one of these ultrasound measurements falls below a specific gestational age threshold (Maulik 2006). Additionally, when a fetus does not receive enough oxygen or nutrients during pregnancy, growth may be limited. Blood flow resistance can be measured using multiple Doppler ultrasound measurements on different arteries (uterine and umbilical), at different locations (left and right), and at multiple times during the pregnancy. Women with high uterine and umbilical artery resistance have been shown to be at an increased risk for delivering a growth-restricted baby (Dugoff et al. 2005; Hugo et al. 2007).

While many ultrasound measurements are available to diagnose growth restriction during pregnancy, it is not clear how best to combine and use all of the measurements. The most common approach compares the measurements of the HC, AC, BPD, and FL to predicted sizes for a gestational age based on population studies to define second trimester growth restriction (e.g. Filly and Hadlock 2000, Dugoff et al. 2005). These analyses focus on identifying a best size measure, although there is often little to distinguish them, and do not discuss situations where different measurements disagree. Other researchers have suggested using a simple average of all of the fetal size measurements for prediction of subsequent growth restriction (Ott 1994). Instead, we propose a latent variable approach, which aggregates the multiple ultrasound measurements into underlying latent variables. We allow the latent variables to follow finite mixture distributions so that we are able to group individuals into latent classes and identify groups of subjects who, based on their ultrasound measurements during pregnancy, are more likely to be growth restricted at birth. For the outcomes measured at birth, the latent class approach also allows us to formalize Wilcox's concept of "residual" and "predominant" distribution components, as we identify a group of subjects who are more likely to have low birth weight and early gestational age. Finally, we incorporate observed covariates including race, parity, and body mass index (BMI) by allowing the covariates to predict latent class membership.

The remainder of this paper will be organized as follows. In section 6.3, we propose our general model for latent variable mixture distributions and compare our method to previous approaches. In section 6.4, we provide a Bayesian approach to fitting such a model while focusing on our fetal development application. Section 6.5 contains the results of our example analysis followed by a discussion in section 6.6.

6.3 Methods

Latent variable methods have been widely used in the social sciences, but the sensitivity of parametric latent variable models to assumptions is one factor that limits their more general use. The parametric assumptions are often difficult to evaluate, and there are no simple methods for alleviating violations of model assumptions. Our reproductive epidemiology data are further complicated by the unusual distribution of gestational age at delivery, which has a long left tail for early ages and a near truncation on the right due to medically-induced labor for longer gestations, due to risk of fetal demise among post term deliveries. The assumption of normality is particularly unlikely to hold in our example, so we propose an alternative method in which the latent variables are allowed to follow mixture distributions. Other approaches that are robust to outliers (Lee and Xia 2006) are only appropriate for symmetrically heavy-tailed distributions. Early births are particularly indicative of future problems so that cutoffs at 37 (pre-term) and 32 (very pre-term) weeks of gestation have been established in the reproductive health literature, with gestational age at delivery being treated as a binary outcome. Our mixture model approach will allow us to identify latent classes of subjects who are more likely to deliver early, without relying on these pre-defined cut points. We then propose a model for examining the association between latent classes and measured covariates.

Mixture distribution models are particularly appropriate when observations are believed to arise from one of several unobserved groups. Our Bayesian approach for fitting finite mixture distributions is based on the work of Diebolt and Robert (1994) and Richardson and Green (1997), but is complicated by needing to model latent rather than observed variables. To model the association of latent group membership with covariates, we utilize data augmentation. In this procedure, latent data augments the observed data so we can use both the latent and observed data to calculate the posterior distribution of the parameters of interest (van Dyk and Meng 2001). Alternatively, Fokoue (2005) has proposed an EM algorithm for normal latent variable mixture model with covariates, but does not model associations between the latent classes.

We assume that each of our observed outcomes measured during pregnancy, y_{ij} $(j = 1, \ldots, p_1)$, follows a normal distribution with a mean that is a function of observed covariates $\boldsymbol{W}_i(r_j \times 1)$ and latent variables $\boldsymbol{\eta}_{i1}(q_1 \times 1)$ with parameter vectors $\boldsymbol{\Gamma}_j(r_j \times 1)$ and $\boldsymbol{\Lambda}_j(q_1 \times 1)$, respectively. Outcomes measured at birth, $z_{ig}(g = 1, \ldots, p_2)$, such as birth weight and gestational age, also follow a normal distribution with a mean that is a function of latent immaturity $(\boldsymbol{\eta}_{i2})$ with parameters $\boldsymbol{\theta}_g(q_2 \times 1)$ and latent class S_i with covariates $\boldsymbol{\beta}_g = [\beta_{g,1}, \ldots, \beta_{g,K}]'$. We allow $\boldsymbol{\eta}_{i1}(q_1 \times 1)$ and $\boldsymbol{\eta}_{i2}(q_2 \times 1)$ to follow finite mixture distributions. It is often convenient to express mixture models using a missing data formulation in which each $\boldsymbol{\eta}_{i1}$ and $\boldsymbol{\eta}_{i2}$ is presumed to arise from a specific, but unknown, underlying component (Dempster et al. 1977). Specifically, for $\boldsymbol{\eta}_{i1}$ and $\boldsymbol{\eta}_{i2}$, respectively, we introduce latent class allocation variables $S_i \in \{1, \ldots, K\}$ where $\Pr(S_i = k) = \pi_{s,k}$ and $T_i \in \{1, \ldots, L\}$ where $\Pr(T_i = l) = \pi_{t,l}$. This specification is useful for computational purposes and allows us to naturally group subjects with similar latent variable characteristics. We can then jointly examine possible associations between measured covariates $\boldsymbol{x}_i$ and S_i using parameters $\boldsymbol{\omega}$ as well as associations between T_i and S_i using parameters $\boldsymbol{\alpha}$ by following Bayesian techniques for probit regression models (Albert and Chib 1993).

To analyze fetal growth, we propose the following latent variable mixture model with covariates

$$\begin{aligned}
y_{ij}|\boldsymbol{\eta}_{i1} &\sim N\left(\boldsymbol{W}_i'\boldsymbol{\Gamma}_j + \boldsymbol{\eta}_{i1}'\boldsymbol{\Lambda}_j, \tau_{y,j}^{-1}\right) \\
\boldsymbol{\eta}_{i1}|S_i = k &\sim N_{q_1}\left(\boldsymbol{\mu}_{1k}, \boldsymbol{\Sigma}_{1k}\right) \\
z_{ig}|\boldsymbol{\eta}_{i2} &\sim N\left(\boldsymbol{\eta}_{i2}'\boldsymbol{\Theta}_g + \sum_{k=1}^{K} \beta_{g,k} I(S_i = k), \tau_{z,g}^{-1}\right) \\
\boldsymbol{\eta}_{i2}|T_i = l &\sim N_{q_2}\left(\boldsymbol{\mu}_{2l}, \boldsymbol{\Sigma}_{2l}\right) \\
S_i|\boldsymbol{x}_i &\sim Multi(1; h(\boldsymbol{\omega}_1, \boldsymbol{x}_i), \ldots, \pi(\boldsymbol{\omega}_K, \boldsymbol{x}_i)) \\
h(\boldsymbol{\omega}_k, \boldsymbol{x}_i) &= \Phi\left(\boldsymbol{\omega}_k'\boldsymbol{x}_i\right) \\
Pr\left(T_i = 0|S_i\right) &\sim Bin(p(\alpha_k, S_i)) \\
p(\alpha_k, S_i) &= \Phi\left(\sum_{k=1}^{K} \alpha_k I(S_i = k)\right) \qquad (6.1)
\end{aligned}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function and $I(S_i = k)$ is the indicator function that takes a value of one if $S_i = k$ and zero otherwise.

Latent variable and mixture models require fixing some parameters so that others can be identified and interpreted. One common approach sets the covariance matrices $\boldsymbol{\Sigma}_{1k}$

and $\mathbf{\Sigma}_{2l}$ to be equal to the identity matrix so that all of the elements $\mathbf{\Lambda}_j$ and $\mathbf{\Theta}_g$ can be identified and interpreted as factor loadings. Instead, we estimate an equivalent model in which $\mathbf{\Sigma}_{1k}$ has diagonal elements $(\tau_{1,1,k}, \ldots, \tau_{1,q_1,k})$ and $\mathbf{\Sigma}_{2l}$ has diagonal elements $(\tau_{2,1,l}, \ldots, \tau_{2,q_2,l})$ with all covariance terms fixed to zero. This specification requires fixing q_1 factor loadings in $\mathbf{\Lambda}$ (where $\mathbf{\Lambda}$ is a stacked matrix of the $\mathbf{\Lambda}_j$) and q_2 factor loadings in $\mathbf{\Theta}$ (a stacked matrix of the $\mathbf{\Theta}_j$) to one so that each of the q_1 latent variables in $\boldsymbol{\eta}_1$ and q_2 latent variables in $\boldsymbol{\eta}_2$ will have a scale that is commensurate with a specific outcome. Such a specification aids in specifying appropriate prior distributions as well as easing interpretation of the latent variables. Elliott et al. (2005) pursues an alternative approach in which $\mathbf{\Lambda}$ is assumed to follow a known polynomial function so that $\mathbf{\Sigma}_{1k}$ can be estimated. Their approach may be more appropriate when the longitudinal observations are measured at more time points than in our application.

6.4 Application

6.4.1 Model Description

In our fetal development example analysis, we observe eighteen total ultrasound measurements, which we summarize with three underlying latent variables, $\boldsymbol{\eta}_{i1} = [\eta_{i11}, \eta_{i12}, \eta_{i13}]'$. Four fetal size measurements are obtained at two time points. Specifically, the abdominal circumference is measured at a 15-week ultrasound $(y_{i,1})$ and 24-week ultrasound $(y_{i,5})$. Head circumference $(y_{i,2}$ and $y_{i,6})$, femur length $(y_{i,3}$ and $y_{i,7})$, and biparietal diameter $(y_{i,4}$ and $y_{i,8})$ are also measured at these two time points. We model the correlation between the fetal size outcomes using the latent early fetal growth variable (η_{i11}). Additionally, the pulsatility index (PI) and systolic-diastolic (S/D) ratio are measured in the left and right uterine arteries at two time points $(y_{i,9}, \ldots, y_{i,16})$, and we capture the correlation between these measurements using a latent uterine blood flow variable, η_{i12}. Finally, we consider the S/D ratio and resistance index (RI) measurements made in the umbilical artery $(y_{i,17}, y_{i,18})$ at week 24 to be error-prone realizations of an underlying latent variable η_{i13}. Umbilical artery measurements are not made earlier because they are technically difficult to perform and thought to be biologically meaningless before 20 weeks.

The path diagram in Figure 15 displays the relationships among variables and modeling assumptions in our proposed model. Circles represent latent variables, and rectangles represent observed variables that either have distributional assumptions (solid lines) or are assumed to be measured without error (dashed lines). Arrows are used to represent associations between variables so that the lack of an arrow indicates conditional independence. For the outcomes measured at birth, we assume that birth weight and gestational age are related to an underlying immaturity variable $(\boldsymbol{\eta}_2)$. Immaturity class membership, T_i, is a latent categorical variable used to identify individual subjects who belong to the residual or predominant portion of the distribution of $\boldsymbol{\eta}_2$. For the ultrasound measurements, we allow the measurements of fetal size and blood restriction to be stochastic functions of the number of days since the last menstrual period (LMP) and model the correlation among these variables using three latent variables $(\boldsymbol{\eta}_1 = [\eta_{11}, \eta_{12}, \eta_{13}]')$. We assume that each of the $\boldsymbol{\eta}_{i1}$ arise from a specific but unknown latent class, where class membership is indicated by the categorical variable S_i. We are then interested in determining if S_i can be used to predict individuals who are more likely to be in the residual component of the birth weight or gestational age distribution, represented by the T_i latent class. We also examine if growth or blood flow restriction during the second trimester is related to growth restriction measured at birth, so we examine the association between birth weight Z-scores and S_i. The birth weight Z-score is a continuous measure of growth restriction that is calculated using birth weight for gestational age population standards (Oken et al. 2003). Finally, we model the association between S_i and black race, being multiparous, gender, height, and BMI.

6.4.2 Measurement Models

We formally express the relationship of outcomes with latent and observed variables using the measurement model

$$\begin{aligned} y_{ij} &= \gamma_{0j} + \lambda_{1j}\eta_{i11} + \gamma_{1j}W_i + \gamma_{2j}W_i^2 + \epsilon_{ij}, \quad j = 1, \ldots, 8 && (6.2) \\ y_{ij} &= \gamma_{0j} + \lambda_{1j}\eta_{i12} + \epsilon_{ij}, \quad j = 9, \ldots, 16 && (6.3) \\ y_{ij} &= \gamma_{0j} + \lambda_{1j}\eta_{i13} + \epsilon_{ij}, \quad j = 17, 18 && (6.4) \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \tau_j^{-1}), j = 1, \ldots, 18$. The ultrasound size measurements ($j = 1, \ldots, 8$) increase with time, so we allow them to be a functions of W_i, the reported time from the last menstrual period (LMP) to the ultrasound for subject i. The blood restriction measurements ($j = 9, \ldots, 18$) do not change as a function of time since LMP over the range of times observed in our study, so we do not include any $\mathbf{\Gamma}$ parameters for time covariates in (6.3) or (6.4). Additionally, in (6.2), we restrict $\lambda_{1,j} = \lambda_{1,j+4}$, $\gamma_{0,j} = \gamma_{0,j+4}$, $\gamma_{1,j} = \gamma_{1,j+4}$, $\gamma_{2,j} = \gamma_{2,j+4}$, and $\tau_j = \tau_{j+4}$, $j = 1, \ldots, 4$, which assumes separate growth curves for the HC, FL, AC, and BPD size measurements. For identifiability, we fix

$\lambda_{0,4} = \lambda_{0,16} = \lambda_{0,18} = 0$ and $\lambda_{1,4} = \lambda_{1,16} = \lambda_{1,18} = 1$ so that η_{i11}, η_{i12}, and η_{i13} will have location and scale that is commensurate with $y_{i,4}$, $y_{i,16}$, and $y_{i,18}$, respectively.

For the outcomes measured at birth, we propose the following measurement model for the continuous outcomes birth weight (z_{i1}), gestational age at delivery (z_{i2}), and birth weight Z-score (z_{i3}). The correlation among birth weight and gestational age is captured using one latent immaturity variable, $\boldsymbol{\eta}_{i2} = [\eta_{i21}]$

$$\begin{aligned} z_{i1} &= \theta_{01} + \theta_{11}\eta_{i21} + \delta_{i1}, \quad \delta_{i1} \sim N(0, \tau_{z_1}) && (6.5) \\ z_{i2} &= \theta_{02} + \theta_{12}\eta_{i21} + \delta_{i2}, \quad \delta_{i2} \sim N(0, \tau_{z_2}) && (6.6) \\ z_{i3} &= \sum_{k=1}^{K} \beta_j I(S_i = k) + \delta_{i3}, \quad \delta_{i3} \sim N(0, \tau_{z_3}) && (6.7) \end{aligned}$$

with $\theta_{02} = 0$ and $\theta_{12} = 1$ so that η_{i21} has a location and scale that is commensurate with gestational age at delivery. Birth weight Z-scores are calculated by comparing the observed birth weight and gestational age to the expected weight for age from approximately 6.7 million US births in 1999 and 2000 (Oken et al. 2003). With such a large number of births, a nearly continuous measure of birth weight for gestational age quantile can be calculated and then transformed to a corresponding Z-score. We use the Z-score in our analysis rather than applying an arbitrary cutoff (such as -1.28) that creates a binary outcome such as small for gestational age. By definition, birth weight Z-scores are independent of gestational age, so we do not allow z_{i3} to be a function of η_{i21}; η_{i21} is thus a latent variable related to the timing of delivery.

6.4.3 Mixture Distribution with Covariates

We allow the latent variables to follow finite mixture distributions

$$f(\boldsymbol{\eta}_{i1}) \sim \sum_{k=1}^{K} \pi_{1k} N_3 \left([\mu_{11k}, \mu_{12k}, \mu_{13k}]', Dg(\tau_{11k}^{-1}, \tau_{12k}^{-1}, \tau_{13k}^{-1}) \right) \tag{6.8}$$

$$f(\boldsymbol{\eta}_{i2}) \sim \sum_{l=1}^{L} \pi_{21l} N \left(\mu_{21l}, \tau_{21l}^{-1} \right) \tag{6.9}$$

where $Dg(\tau_{11k}^{-1}, \tau_{12k}^{-1}, \tau_{13k}^{-1})$ is a diagonal covariance matrix with elements τ_{1mk}^{-1}. In our example analysis, we use a two-component mixture for early growth restriction ($K = 2$) and a three-component mixture for latent immaturity ($L = 3$).

Latent class models are subject to additional identifiability complications due to the fact that the likelihood is symmetric across the possible permutations of class membership. Consequently, assignment to a particular class k during one iteration of the Gibbs sampler may not have the same meaning in terms of model structure as assignment to class k at another iteration of the Gibbs sampler. For the latent immaturity variable $\boldsymbol{\eta}_2$, we impose the identifiability constraint that $\mu_{211} < \mu_{212} < \mu_{213}$ to deal with the "label switching" problem and to ensure that subjects with the lower birth weights and earlier gestational ages will be assigned to the first latent class. We can then examine the association between class membership and observed covariates or latent variables. While ordering the means is effective in many cases including our application, this approach is not a general solution to the label switching problem (Stephens 1997). Stephens (2000) provides a decision-theoretic approach that maximizes the posterior probability that the labeling of classes is consistent across iterations of the Gibbs sampler. We utilize his post-processing technique for the labeling of the early growth restriction classes (S_i) because, unlike the classes for immaturity at birth (T_i), the ordering of the classes is not important and can be disentangled after the completion of the Gibbs sampling runs.

Gestational age at delivery follows an unusual, asymmetric distribution with a long tail to the left for early births and a near truncation on the right due to labor being medically induced for very long gestations. Using the restriction $\mu_{211} < \mu_{212} < \mu_{213}$ ensures that subjects who are more likely to be born early will be assigned to latent class $T_i = 1$. We then propose a probit model for the probability of being in this early group

$$\Pr(T_i = 1) = \Phi\left(\sum_{k=1}^{K} I(S_i = k)\alpha_k\right), \tag{6.10}$$

where Φ is the normal cumulative distribution function. We also use a probit model to examine the association between observed covariates and S_i,

$$\Pr(S_i = 1) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\omega}_k\right), \tag{6.11}$$

where $\boldsymbol{x}_i(r \times 1)$ includes covariates of interest with parameters $\boldsymbol{\omega}$ $(r \times 1)$. In the fetal development example, we consider the covariates maternal black race, parity, gender, maternal body mass index (BMI), and maternal height. To fit these probit regression models, we use the data augmentation algorithm given Albert and Chib (1993). While we are only concerned with binary outcomes in our application, our models could be extended to incorporate ordinal or multinomial outcomes if S_i takes on more than two levels (Holmes and Held 2006).

6.4.4 Model Selection

A complication in mixture distribution models with a finite number of components is the method used to select the number of mixture components. A more general model could treat the number of mixture components as parameters to be estimated (Richardson and

Green 1997) or use a Dirichlet process in which the number of components is countably infinite (Dunson 2006a), but neither of these approaches would be able to readily incorporate covariates that predict class membership. When possible, we prefer an approach for selecting the number of mixture components that is guided by the application. For example, in the reproductive epidemiology literature birth weight and gestational age have been described as arising from a predominant and residual component that is indicative of early gestational age and low birth weight (Buekens et al. 2000; Wilcox et al. 2001). We use a three component mixture for latent immaturity class (T_i) in order to identify the residual distribution, and thus identify individuals who are at increased risk for mortality and other forms of morbidity. A mixture with only two components did not identify the residual group, and a model with more than three components only improves the fit of the predominant distribution so it has a needlessly complex interpretation.

When the applied problem is not helpful in selecting the number of mixture components, some statistical tools are available. Bayesian approaches for comparing complex hierarchical models in which the number of parameters is not clearly defined include using the deviance information criterion (DIC; Spiegelhalter et al. 2002). The DIC is not without problems, and, for mixture models in particular, the DIC is thought to favor overly-complex models (Richardson, in discussion of Spiegelhalter et al. 2002). More recently, Celeux et al. (2006) discuss and compare different constructs of the DIC in the general context of missing data models, with mixture models explored in detail. Alternatively, posterior prediction intervals can help determine if too few classes are chosen, leading to a model that under-fits the data (Lynch and Western 2004).

6.4.5 Prior distributions

To complete a Bayesian specification of the model, prior distributions must be specified for each parameter. In general, we use proper but appropriately vague priors for all parameters to obtain complete conditionals that are of known form. We use conditionally conjugate priors $p\left(\boldsymbol{\Lambda}_j|\tau_{y,j}\right) \sim N\left(\boldsymbol{\mu}_{0,\Lambda_j}, \tau_{y,j}^{-1}\boldsymbol{\Sigma}_{0,\Lambda_j}\right)$, $p\left(\boldsymbol{\Gamma}_j|\tau_{y,j}\right) \sim N\left(\boldsymbol{\mu}_{0,\Gamma_j}, \tau_{y,j}^{-1}\boldsymbol{\Sigma}_{0,\Gamma_j}\right)$, $p\left(\tau_{y,j}\right) \sim \Gamma\left(\frac{c_{y,j}}{2}, \frac{d_{y,j}}{2}\right)$, $p\left(\boldsymbol{\Theta}_g|\tau_{z,g}\right) \sim N\left(\boldsymbol{\mu}_{0,\Theta_g}, \tau_{z,g}^{-1}\boldsymbol{\Sigma}_{0,\Theta_g}\right)$, $p\left(\beta_k|\tau_{z,g}\right) \sim N\left(\boldsymbol{\mu}_{0,\beta_k}, \tau_{z,g}^{-1}\boldsymbol{\Sigma}_{0,\beta_k}\right)$, and $p\left(\tau_{z,g}\right) \sim \Gamma\left(\frac{c_{z,g}}{2}, \frac{d_{z,g}}{2}\right)$. We then choose $\boldsymbol{\mu}_{0,\Lambda_j} = \boldsymbol{\mu}_{0,\Gamma_j} = \boldsymbol{\mu}_{0,\Theta_g} = \boldsymbol{\mu}_{0,\beta_k} = \mathbf{0}$, $\boldsymbol{\Sigma}_{0,\Lambda_j} = \boldsymbol{\Sigma}_{0,\Gamma_j} = \boldsymbol{\Sigma}_{0,\Theta_j} = \boldsymbol{\Sigma}_{0,\beta_k} = 1000^2\boldsymbol{I}$, and $c_{y,j} = d_{y,j} = c_{z,g} = d_{z,g} = .01$ for every j where $\mathbf{0}$ is a conforming vector of zeros and $\boldsymbol{I}$ is a conforming identity matrix. We also assume $p(\boldsymbol{\omega}) \sim N\left(\boldsymbol{\mu}_{0,\omega}, \boldsymbol{\Sigma}_{0,\omega}\right)$ and $p(\boldsymbol{\alpha}) \sim N\left(\boldsymbol{\mu}_{0,\alpha}, \boldsymbol{\Sigma}_{0,\alpha}\right)$ with $\boldsymbol{\mu}_{0,\omega} = \boldsymbol{\mu}_{0,\alpha} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{0,\omega} = \boldsymbol{\Sigma}_{0,\alpha} = 100^2\boldsymbol{I}$.

For the mixture distribution component of our model, we use a prior specification that follows the suggestions of Richardson and Green (1997). For latent immaturity and $l = 1, 2, 3$ we assume $p\left(\mu_{21l}\right) \sim N\left(\nu_{21l}, R^2\right) I\left(\mu_{21,l-1} < \mu_{21l} < \mu_{21,l+1}\right)$ where $\mu_{210} = -\infty$ and $\mu_{214} = \infty$. We choose $\nu_{211} = 245$, $\nu_{212} = \nu_{213} = 280$, and $R = 10$ so that, a priori, we expect that the residual distribution will have a mean of 245 ± 20 (days) and the predominant distribution a mean of 280 ± 20 (Gage 2002). For the mixture component means of $\boldsymbol{\eta}_1$, we use $p\left(\mu_{1mk}\right) \sim N\left(\nu_{1mk}, R_m^2\right)$ ($m = 1, 2, 3$ and $k = 1, 2$) with $R_1 = 10$, $R_2 = 2.17$, and $R_3 = 0.32$ being the observed ranges of $\boldsymbol{y}_4$ (after adjusting for time since LMP), $\boldsymbol{y}_{16}$, and $\boldsymbol{y}_{18}$.

We use a hierarchical structure for specifying the prior distribution of each τ_{1mk} and τ_{21l}. Specifically, we allow $p\left(\tau_{1mk}|b_{0,m}\right) \sim \Gamma\left(a_{0,m}, b_{0,m}\right)$ and $p\left(\tau_{21l}|b_0\right) \sim \Gamma\left(a_0, b_0\right)$ with

$b_{0,m} \sim \Gamma(g_{0,m}, h_{0,m})$ and $b_0 \sim \Gamma(g_0, h_0)$. We choose $a_{0,m} = a_0 = 2$, $g_{0,m} = g_0 = 0.2$, $h_{0,m} = 10 * R_m^{-2}$, and $h_0 = 10 * R^{-2}$ where $\Gamma(a, b)$ is the gamma distribution with mean $a \div b$ and variance $a \div b^2$. By choosing $a_{0,m} > 1 > g_{0,m}$ (and $a_0 > 1 > g_0$) we express the general belief that, for each k (and l), the τ_{1mk} (and τ_{21l}) are similar, but we have no information on their absolute size. Finally, we assume that $\boldsymbol{\pi}_s = [\pi_{s1}, \pi_{s2}]'$ and $\boldsymbol{\pi}_t = [\pi_{t1}, \pi_{t2}, \pi_{t3}]'$ follow independent, symmetric Dirichlet distributions, $p(\boldsymbol{\pi}_s) \sim D(d_1, \ldots, d_1)$ and $p(\boldsymbol{\pi}_t) \sim D(d_2, \ldots, d_2)$ and choose $d_1 = d_2 = 1$ to be appropriately vague.

6.5 Results

To perform the analysis, we used Gibbs sampling, conducted in Matlab, with the complete conditionals given in Appendix C (Casella and George 1992). We ran five chains from disparate starting values and monitored convergence using the CODA package for R (R Development Core Team 2004). After removing an initial burn in of 15,000 iterations, all parameters were judged to have converged by the Gelman-Rubin and Geweke diagnostic measures (Gelman and Rubin 1992; Geweke 1991). For example, all Gelman-Rubin statistics ($\hat{R}$) were found to be less than 1.01, where $\hat{R} = 1$ at convergence and values less than 1.2 are generally considered sufficient for convergence (Gelman et al. 2004). We used the remaining 35,000 iterations for inference, and summarize our results using posterior means and 95% credible intervals (CI) for parameters or functions of parameters (e.g. odds ratios) that are of interest.

We applied our methods to 522 subject taken from the Pregnancy, Infection, and Nutrition (PIN) cohort study of prenatal influences on pregnancy outcomes (Savitz et al.

1999). We included all singleton, live-born infants who had complete ultrasound, birth weight, gestational age, and covariate information in this analysis. Characteristics of the study subjects are presented in Table 6.6. As expected, birth weight and gestational age had skewed distributions toward early birth and low weight, respectively.

For each subject, fetal size measurements including the abdominal circumference, biparietal diameter, femur length, and head circumference were made at approximately 15 and 24 weeks gestational age using an ultrasound. Additionally, multiple blood restriction measurements were taken at week 15 and week 24 using a Doppler ultrasound. The correlation among the Doppler ultrasound blood flow measures as well as their means and standard deviations are presented in Table 6.6. The highest correlations (all $\rho \geq 0.90$) were observed among the S/D ratio and PI (or RI for umbilical artery) within a given artery, location, and time. We observed the lowest correlation between the uterine and umbilical artery measurements. In the uterine arteries, the mean resistance, as measured by either the S/D ratio or PI, did not change meaningfully from the week 15 to week 24 ultrasound. Based on this descriptive analysis, we chose to model the correlation between the uterine artery measurements using one latent variable (η_{i12}), and the correlation among the umbilical artery S/D and RI using a second latent variable (η_{i13}).

Birth weight and gestational age were collected at birth, and using this information, we calculated birth weight Z-scores using birth weight for gestational age population standards provided by Oken et al. (2003). For any given gestational age, the birth weight Z-scores follow a standard normal distribution relative to the reference population so that individuals with negative values are believed to have some degree of growth restriction at birth. In simple linear regression models, black race was associated with a 0.48 unit decrease in birth weight Z-score (95% CI: [-0.68, -0.28]) compared to all other races. Birth weight Z-scores were 0.23 units lower in women who had no previous live births (95% CI: [-0.38, -0.07]) compared with multiparous women and 0.13 units lower (95% CI: [-0.24, -0.05]) in females compared to males. Additionally, a one inch increase in maternal height (estimate = 0.06, 95%CI: [0.03, 0.09]) and a one kg/m^2 increase in maternal BMI (estimate = 0.017, 95%CI: [0.004, 0.029] were associated with increased birth weight Z-scores. In our latent variable mixture model, we examined if any of these covariates predict growth restriction latent class (S_i), with the latent class then used to predict birth weight Z-scores (Figure 15).

For early growth restriction, our latent variable mixture model identified two groups of subjects based on their multiple ultrasound measurements of fetal size and blood flow resistance. For ease of exposition, we refer to these groups as the "normal" and "restricted" groups. On average, a majority of subject belong to the normal group (posterior mean = 67%, 95% CI = [59%,74%]). Figure 16 displays a comparison of these two groups for the four measures of fetal size, the PI at various times and locations, and the RI obtained in the umbilical artery at week 24 by Doppler ultrasound. Controlling for time since LMP, the restricted group had, on average, smaller fetal size measurements and greater resistance to blood flow than the normal group. The restricted group also had significantly larger S/D ratios, indicating greater blood flow resistance, at all times and locations (results not shown). Furthermore, we found evidence of growth restriction at birth in the restricted group. The average birth weight Z-score was -0.34 (95% CI: [-0.51, -0.18]) in the restricted group and 0.11 (95% CI: [0.00, 0.22]) in the normal group.

We also examined the association of black race, parity, gender, height, and BMI with being classified into the restricted early growth group. We found a moderate association with maternal height in that a one inch increase was associated with a 0.82 (95% CI: [0.57,1.00]) fold decrease in the odds of belonging to the restricted group. Black race (posterior odds = 1.33, 95% CI: [0.77,2.14]), being nullparous (posterior odds = 1.12, 95% CI: [0.74,1.65]), and being female (posterior odds = 1.412, 95% CI: [0.91,2.12]) were not significantly associated with belonging to the restricted group. We also found no linear association between group membership and BMI, but also examined BMI by previously established categories ranging from underweight to obese (WHO Expert Committee 1995). Underweight (BMI < 18.5 kg/m^2) and obese women (BMI ≥ 30 kg/m^2) had a relatively high 0.36 (95% CI: [0.23,0.49]) and 0.38 (95% CI: [0.27,0.50]), respectively, posterior probability of belonging to the restricted group. Women with a BMI in the normal range (BMI $\in [18.5, 25)$ kg/m^2) or moderately obese women (BMI $\in [25, 30)$ kg/m^2) had lower posterior probabilities of 0.31 (95% CI: [0.23,0.39]) and 0.31 (95% CI: [0.18,0.45]), respectively.

Our latent variable mixture distribution approach fits the observed birth weight and gestational age distributions well. Figure 17 depicts the observed and estimated cumulative distribution functions (CDFs) for gestational age at delivery while Figure 18 contains these CDFs for birth weight. Tail areas are magnified in the figures. Particular interest lies in identifying fetuses that have a latent tendency to be born the very early and with low weight, so we assigned these subjects to a specific latent class. To help understand the type of subjects who belong to this immaturity class, we include conditional probability plots in Figures 17 and 18. *A priori*, each subject had an equal probability of belonging to the residual distribution, but the posterior distribution of class membership is strongly related to birth weight and gestational age. For example, Figure 17 includes the cumulative probability of being assigned to the immaturity class, conditional on gestational age. All subjects born before 224 days (very pre-term) and approximately 30% of subjects born before 259 days (pre-term) are assigned to the immaturity latent class. This compares to a 3.7% (95% CI: [1.2%, 10.4%]) marginal probability of belonging to the residual component. For birth weight, some subjects with very low weight (< 1500g) had a low probability of being assigned to the residual distribution. Such a scenario is possible when an individual has low birth weight due to slow growth during pregnancy without early gestational age. Still, about 90% of subjects with very low birth weight (< 1500g) and 40% of subjects with low birth weight (< 2500 g) belong to the latent class indicative of the residual distribution (Figure 18).

Finally, we examined the association between our two latent class variables. Using a probit regression model, individuals in the restricted latent class during the second trimester were 3.45 times (95% CI: [0.86,58.9]) more likely to belong to the residual component of the distribution at birth. The large posterior odds ratio indicates that belonging to the restricted class is potentially an important predictor of future immaturity, but the credible interval is too wide to make a definitive statement in this dataset.

In our analysis, our primary inference is about latent variables and latent classes, and we make a number of distributional assumptions, so it is important to evaluate the fit of our model. To do so, we calculated the posterior predictive distributions for observed data and then compared these distributions to our observed data graphically following Lynch and Western (2004). Figure 19 shows the posterior distributions and observed data for the S/D ratio obtained in the right uterine artery at week 15 for 90 subjects.

Our model appears to fit the observed data well. Posterior predictive plots for other ultrasound measurements, birth outcomes, and subjects also do not indicate a lack of fit.

6.6 Discussion

We develop a Bayesian approach for analyzing multiple correlated pregnancy outcomes measured during pregnancy by ultrasound and collected routinely at birth using latent variable mixture models. We found evidence in favor of the existence of a latent class of subjects who were more likely to have smaller fetal size measurements and restricted blood flow during the second trimester. Subjects in the restricted group had increased growth restricted at birth and may be at increased risk for belonging to the residual distribution of birth weight and gestational age. Finally, we found that height, BMI, black race, gender, and parity were directly related to birth weight Z-scores, but we did not find an association between these covariates and the latent class of subjects who were restricted during pregnancy in a mediation model.

Latent variables methods provide a natural way of aggregating multiple correlated outcomes to describe underlying concepts. We collected a total of eighteen ultrasound measurements on each subject, summarized them using three latent variables, and then subsequently into two latent classes. Rather than use the ultrasound measurements directly, we focus on making inference using the latent classes. *A priori*, the model assumes that each subject has an equal probability of belonging to each of the latent classes, but the posterior distribution is a mixture over all classes. Individuals with smaller fetal size and larger blood restriction have an increased posterior probability of belonging to a restricted latent class, which could be useful in early identification of individuals who are more likely to be growth restricted at birth for possible medical intervention or closer monitoring.

Our mixture distribution approach is particularly appropriate for formalizing the concept of predominant and residual components of the birth weight and gestational age distributions developed by Wilcox et al. (2001). We estimated that approximately 3.7% of births lie in the residual distribution, which is in the 2% to 5% range previously estimated for birth weight (Wilcox et al. 2001). A population with a higher proportion of births in the residual component will be at an increased risk for infant mortality and later morbidity outcomes such as diabetes (Harder et al. 2007; Shan and Ohlsson 2002). Conversely, two populations may differ in their predominant distributions without a corresponding change in outcome. Often researchers attempt to find associations between exposures such as air pollution or abuse and mean changes in birth weight or gestational age using linear regression models (Curry et al. 1998; Glinianaia et al. 2004). Such models assume that shifts in the mean of birth weight or gestational age will correspond to an increase in the percentage of the population at particularly high risk for complications. However, from our perspective, changes in the mean could be due to a shift in the predominant component of the distribution, or a more concerning increase in the proportion of babies born in the residual component, or a combination of the two. For example, Mexican-American babies on average have lower birth weights compared to US non-Hispanic white babies, but without a corresponding increase in mortality. This observation could be due to an unimportant shift in the predominant component toward lower birth weights without an increased risk of being born in the residual component of the distribution (Buekens et al. 2000). Our model focuses on identifying a latent class of subjects who, based on many second trimester ultrasound measurements, are at increased the risk of falling in the residual component.

Characteristic	Percent	Mean	Std Dev	Min	Max
Birth weight (grams)		3324	570	1118	5171
Gestational age (days)		274	13.5	203	298
Maternal height (in)		65	2.5	59	72
Maternal BMI (kg/m^2)		25	6.2	16	53
Small for gestational age	8.7%				
Pre-term birth ($<$ 37 weeks)	10.7%				
Very pre-term birth ($<$ 32 weeks)	1.3%				
Low birth weight ($<$ 2500 g)	7.1%				
Very low birth weight ($<$ 1500 g)	1.1%				
Maternal Black Race	17.5%				
Parity ≥ 1	55.1%				
Male gender	47.8%				

Table 5: Descriptive statistics for the 522 PIN subjects studied

	Week 15 Ultrasound				Week 24 Ultrasound					
	Left UA		Right UA		Left UA		Right UA		Umbilical	
	S/D	PI	S/D	PI	S/D	PI	S/D	PI	S/D	RI
15LTSD	1.00									
15LTPI	0.92	1.00								
15RTSD	0.33	0.35	1.00							
15RTPI	0.34	0.39	0.92	1.00						
24LTSD	0.46	0.43	0.32	0.31	1.00					
24LTPI	0.42	0.43	0.30	0.30	0.93	1.00				
24RTSD	0.23	0.24	0.46	0.46	0.38	0.38	1.00			
24RTPI	0.18	0.22	0.41	0.43	0.34	0.36	0.90	1.00		
24UMSD	0.10	0.10	0.04	0.05	0.09	0.06	0.12	0.07	1.00	
24UMPI	0.10	0.10	0.02	0.04	0.09	0.06	0.08	0.04	0.97	1.00
Mean	2.58	1.08	2.50	1.05	2.02	0.81	2.05	0.84	2.97	0.65
Std Dev	0.97	0.40	0.92	0.41	0.41	0.25	0.43	0.26	0.51	0.06

Table 6: Mean, standard deviation, and correlation among the blood restriction measurements including the Systolic-Diastolic ratio (SD), Pulsatility Index (PI) and Resistance Index (RI) made in the left (LT) or right (RT) uterine artery (UA) or umbilical (UM) artery during the week 15 or week 24 ultrasound.

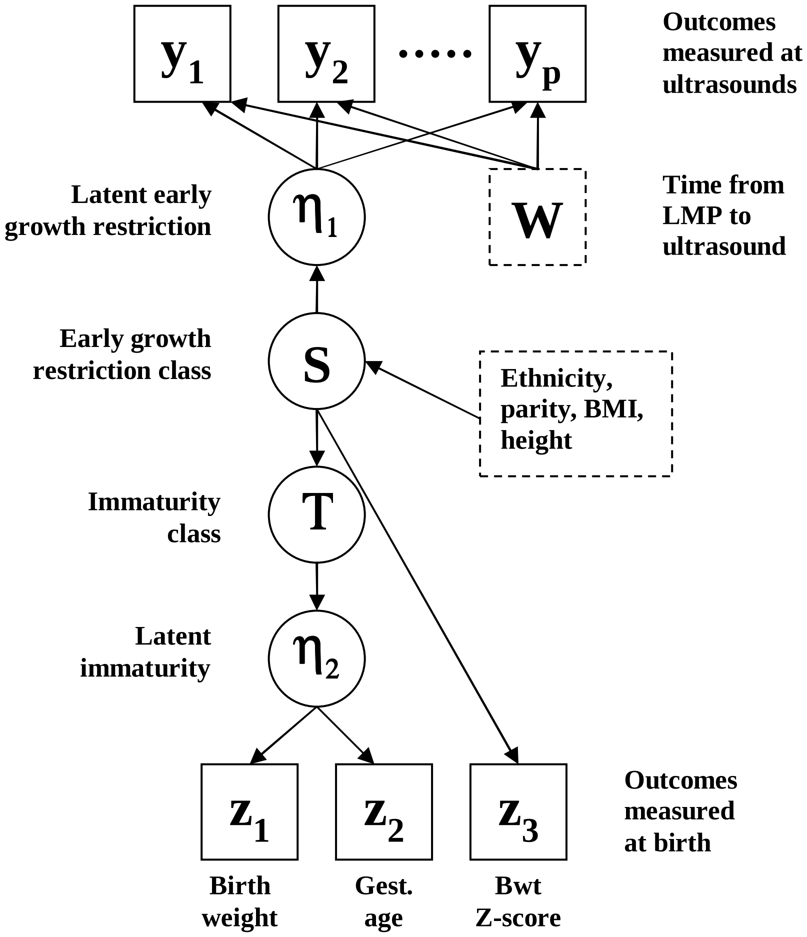
y_1
y_2
y_p
Outcomes measured at ultrasounds
Latent early growth restriction
η_1
W
Time from LMP to ultrasound
Early growth restriction class
S
Ethnicity, parity, BMI, height
Immaturity class
T
Latent immaturity
η_2
z_1
z_2
z_3
Outcomes measured at birth
Birth weight
Gest. age
Bwt Z-score

Figure 15: Path diagram illustrating the dependencies in the proposed growth restriction and latent immaturity model. Circles represent latent variables, squares indicate observed variables, and arrows show association.

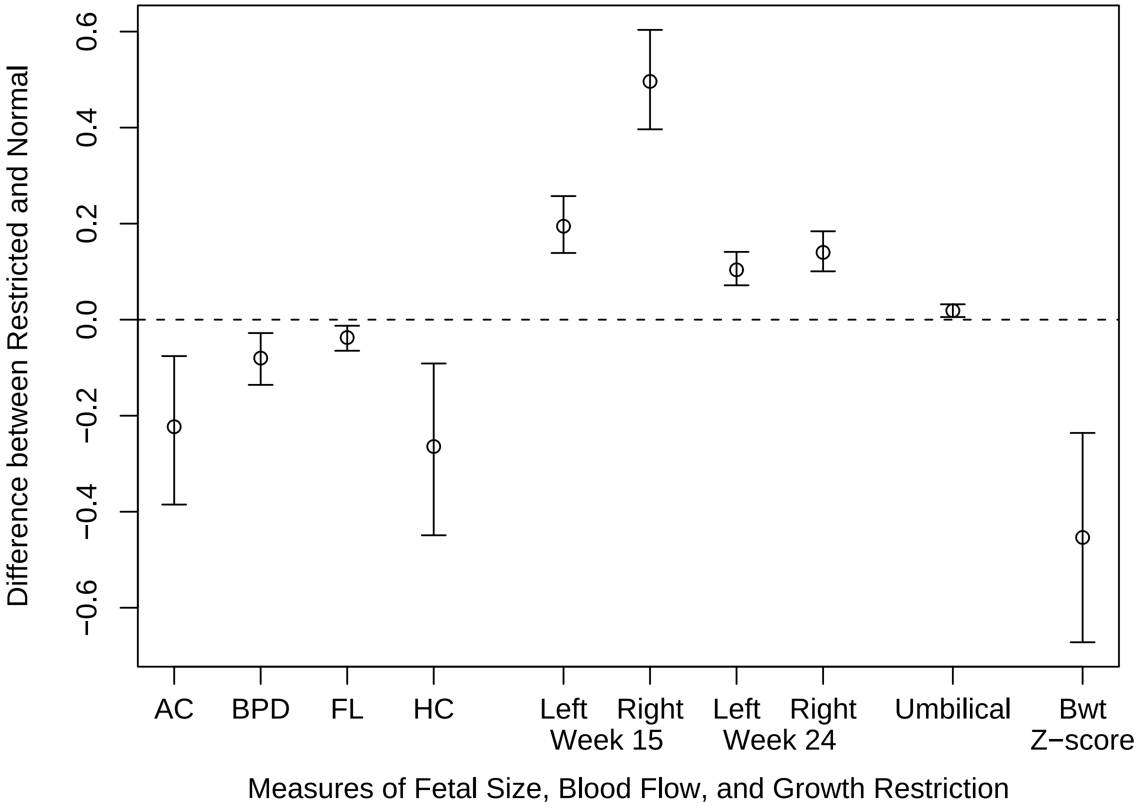

Figure 16: Differences in ultrasound measurements of fetal size, blood restriction, and birth weight Z-scores for the restricted and normal latent classes. Fetal size measurements include the abdominal circumference (AC), biparietal diameter (BPD), femur length (FL), and head circumference (HC). The pulsatility index in the left and right uterine arteries at the week 15 and week 24 ultrasound as well as the resistance index in the umbilical artery are shown to measure blood resistance.

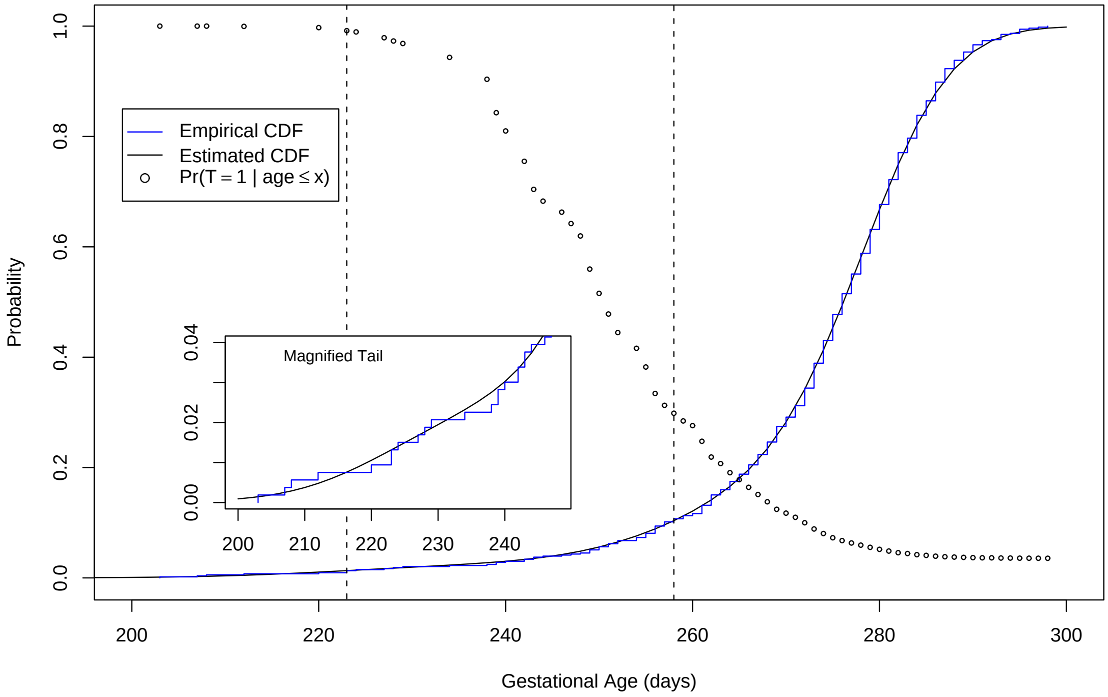

Figure 17: The solid lines represent the empirical and estimated cumulative distribution functions for gestational age at delivery. The fit of the residual distribution is magnified. Points provide the cumulative probability of being assigned to the latent class indicative of immature delivery ($T = 1$), conditional on gestational age. Vertical dashed lines indicate the commonly-used very PTB and PTB cutoffs.

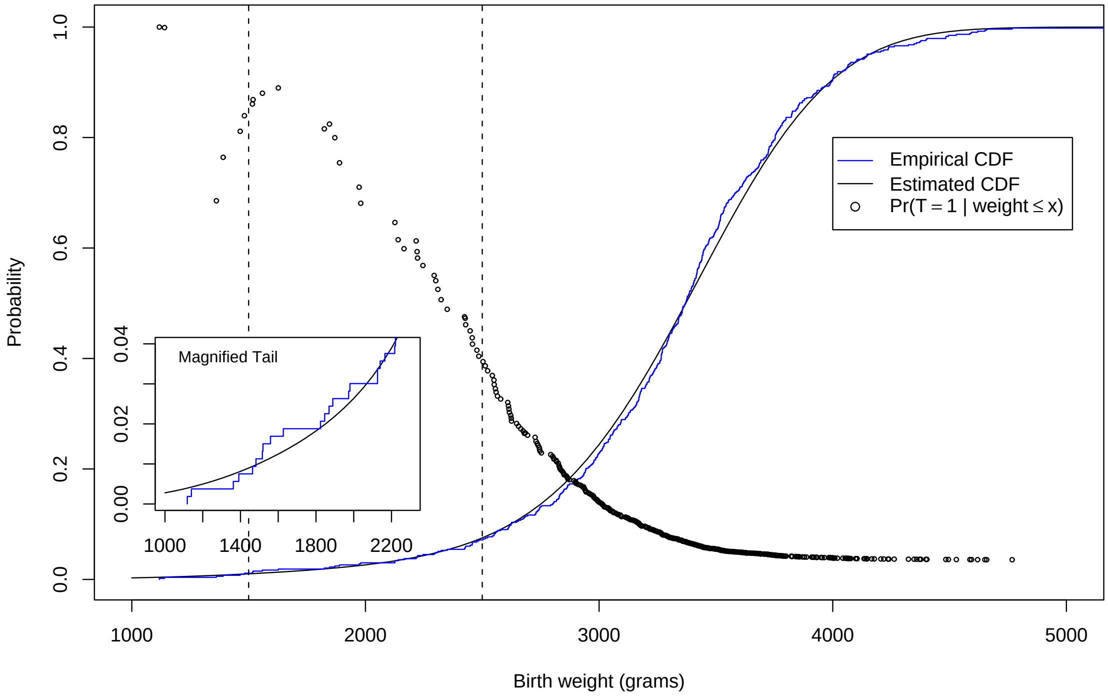

Figure 18: The solid lines provide the empirical and estimated cumulative distribution functions for birth weight. The fit of the residual distribution is magnified. Points represent the cumulative probability of being assigned to the latent class indicative of immature delivery ($T = 1$), conditional on birth weight. Vertical dashed lines indicate the commonly-used very LBW and LBW cutoffs.

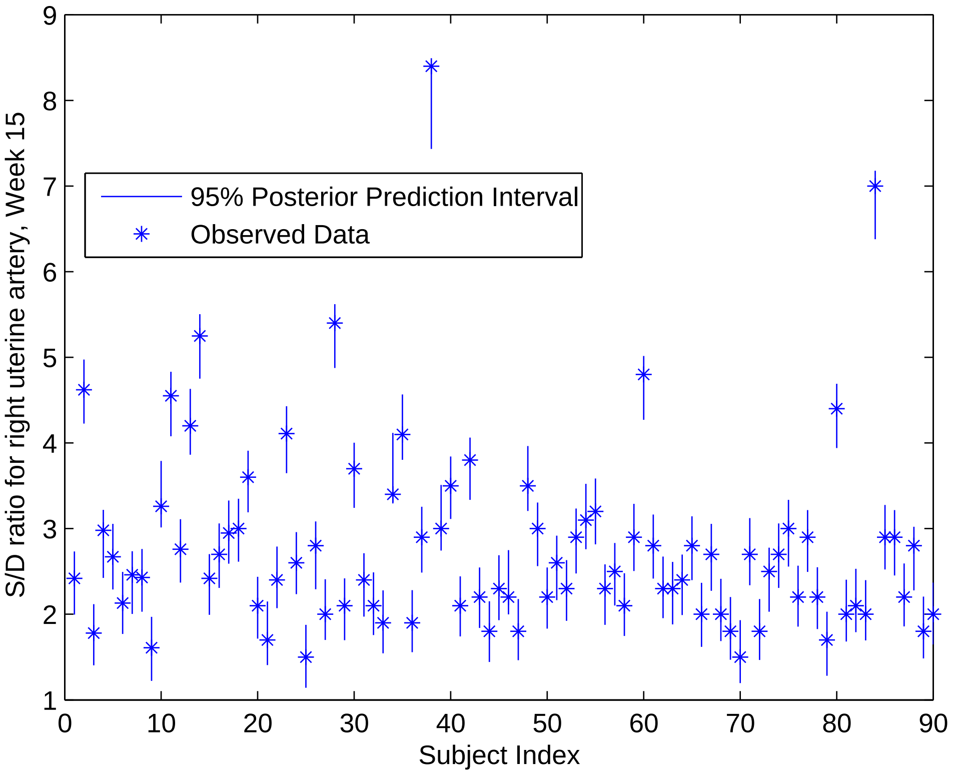

Figure 19: Posterior prediction intervals and observed S/D ratios for for ninety selected subjects.

7 Conclusions and Future Directions

Latent variable methods provide a flexible approach for complex modeling of correlation in longitudinal studies. We have discussed methods for using latent variables to aggregate multiple ultrasound measurements and model fetal growth and development during the first and second trimesters. These results are particularly important to researchers who use ultrasounds to date pregnancies while assuming that there is no measurable variability in fetal growth early in pregnancy. There is a general need to make latent variable methods more familiar to biostatisticians by applying them to research areas in public health. Furthermore, by having a solid understanding of the subject matter, the insights gained from an analysis using latent variable methods can be effectively communicated to researchers in epidemiology and clinical disciplines. To make our methods more accessible, it is possible that, with some modification, the latent variable mixture models describe in papers two and three could be estimated using commercial software such as M-Plus. Using available software would be particularly useful for journal articles intended for applied researchers in reproductive health.

In our first paper, we develop a Bayesian discrete time multistate growth model for inference from cross-sectional data with unknown initiation times. Our methods were able to identify subjects who have smaller than expected gestational sac diameters, have shorter fetal pole lengths, and transition through growth states relatively slowly. We then found evidence in favor of a previously hypothesized but unproven association between slow fetal growth early in pregnancy and increased risk of subsequent pregnancy loss. Other analysis options are also possible. In particular, we only included pregnancies that appeared to be progressing normally at the time of the ultrasound so that any pregnancy loss occurred after the ultrasound. Another model could consider losses at any time by including an additional, absorbing "loss" state that could be reached from any of the growth states considered in our analysis. Such an approach could focus on estimating transition rates into the loss state as well as determining probable developmental state prior to the loss.

The sensitivity of latent variable models to parametric assumptions are one factor that limit their use beyond the social sciences. The parametric assumptions are often difficult to evaluate and there are no simple methods for alleviating violations of model assumptions. Semi-parametric methods for latent variable models are an area in need of further development. While our first paper was based on more established methods that assume the latent variable follows a normal distribution, our second and third papers relaxed this assumption. In papers two and three, we used a latent variable mixture model to approximate the unusual distributions of birth weight and gestational age at delivery. Additionally, in paper three, we identified a latent class of subjects who had small fetal size measurements and increased blood flow restriction who were more likely to be growth restricted at birth.

Our latent variable mixture models attempted to formalize Wilcox's (2001) idea of a "residual" and "predominant" distribution and provide a new method of examining low birth weight and pre-term birth. Low birth weight can be caused by slow intrauterine growth, early gestational age, or a combination of these two factors. In estimating a latent immaturity variable and immaturity class, we characterized the residual group as having early gestational age with corresponding low birth weight. However, in reproductive health, interest also lies in identifying babies that have a relatively low weight without early gestational age. Our latent variable methods could be altered to estimate membership in this important latent class and identify covariates that are associated with class membership.

While completing these three papers, we considered several additional research areas for our specific applications, but did not address these topics in a general manner. Identifiability is a common concern in latent variable models and general rules or methods for evaluating identifiability need to be developed. For mixture distribution models, methods for insuring proper labeling of the k mixture components across iterations of the Gibbs sampler need to be developed. The currently favored approach (Stephens 2000) relies on post-processing the results at the end of the Gibbs sampling run. This solution is not helpful when the classes need to be ordered at every iteration, such as when we examining the association of covariates with class membership. We used order restriction-based conditions (i.e. on the k class means) to identify the predominant and residual components of the distribution, but restrictions are not a general solution and were only effective in our analysis because the class means were sufficiently dispersed. Model selection, including selecting the number of mixture components, is also an area of current research. More generally, computational issues associated with slow convergence of the Gibbs sampler are another concern and are particularly troublesome in large datasets. Finally, in a Bayesian analysis, the choice of priors and the robustness of latent variable models to different prior assumptions need to be evaluated.

A Appendix for paper 1

We consider the complete conditionals for a three state model where R_i is the number of time intervals between initiation and $l_i^{(1)}$ and S_i the number of intervals between $l_i^{(2)}$ and $l_i^{(1)}$. For notational convenience, let $\phi_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the Normal probability density function for the p dimensional random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ and let $\phi_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{Y}=\boldsymbol{y})$ be the Normal density evaluated at $\boldsymbol{Y}=\boldsymbol{y}$ where $\boldsymbol{\mu}$ is a function of $\boldsymbol{Y}$. Also, let $\psi(x;\alpha,\beta)\propto x^{\alpha-1}e^{-x\beta}$ be the gamma probability density function for random variable x. For matrix calculations, let $\boldsymbol{A}^{\otimes 2}=\boldsymbol{A}'\boldsymbol{A}$, $\boldsymbol{1}_n$ be a $n\times 1$ vector of ones, and $\boldsymbol{I}_n$ the identity matrix of rank n.

Step 1: Using conjugate priors, parameters $\boldsymbol{\Lambda}_k, \tau_{z_k}$, and $\xi, k=1,2$, follow Normal and Gamma forms from linear regression results

$$\begin{aligned}
[\boldsymbol{\Lambda}_k|\boldsymbol{X}_k,\boldsymbol{Z}_k,\xi,\boldsymbol{Z}^*,\tau_{z_k}] &\propto \phi_n\left(\boldsymbol{Z}_k;\boldsymbol{X}_k\boldsymbol{\Lambda}_k+\xi\boldsymbol{Z}^*,\tau_{z_k}^{-1}\boldsymbol{I}_n\right)\phi\left(\boldsymbol{\Lambda}_k;\boldsymbol{\mu}_{0,\Lambda_k},\tau_{z_k}^{-1}\boldsymbol{\Sigma}_{0,\Lambda_k}\right)\\
&\sim N\left(\boldsymbol{A}^{-1}\boldsymbol{b},\tau_{z_1}^{-1}\boldsymbol{A}^{-1}\right);\boldsymbol{A}=\boldsymbol{X}_k'\boldsymbol{X}_k+\boldsymbol{\Sigma}_{0,\Lambda_k}^{-1},\boldsymbol{b}=\boldsymbol{X}_k'\left(\boldsymbol{Z}_k-\xi\boldsymbol{Z}^*\right)+\boldsymbol{\Sigma}_{0,\Lambda_k}^{-1}\boldsymbol{\mu}_{0,\Lambda_k}\\
[\tau_{z_k}|\boldsymbol{\Lambda}_k,\boldsymbol{X}_k,\xi,\boldsymbol{Z}^*,\boldsymbol{Z}_k] &\propto \phi_n\left(\boldsymbol{Z}_k;\boldsymbol{X}_k\boldsymbol{\Lambda}_k+\xi\boldsymbol{Z}^*,\tau_{z_k}^{-1}\boldsymbol{I}_n\right)\psi\left(\tau_{z_k};\frac{\delta_{0,z_k}}{2},\frac{\lambda_{0,z_k}}{2}\right)\\
&\sim G\left(\frac{n+\delta_{0,z_k}}{2},\frac{1}{2}\left[\left(\boldsymbol{Z}_k-\boldsymbol{X}_k\boldsymbol{\Lambda}_k-\xi\boldsymbol{Z}^*\right)^{\otimes 2}+\lambda_{0,z_k}\right]\right)\\
[\xi|\boldsymbol{\Lambda}_1,\boldsymbol{X}_1,\boldsymbol{Z}_1,\boldsymbol{\Lambda}_2,\boldsymbol{X}_2,\boldsymbol{Z}_2,\boldsymbol{Z}^*] &\propto \phi_n\left(\boldsymbol{Z}_1;\boldsymbol{X}_1\boldsymbol{\Lambda}_1+\xi\boldsymbol{Z}^*,\tau_{z_1}^{-1}\boldsymbol{I}_n\right)\\
&\quad\times\phi_n\left(\boldsymbol{Z}_2;\boldsymbol{X}_2\boldsymbol{\Lambda}_2+\xi\boldsymbol{Z}^*,\tau_{z_2}^{-1}\boldsymbol{I}_n\right)\phi_1\left(\xi;0,1\right)\\
&\sim N\left(\boldsymbol{A}^{-1}\boldsymbol{b},\boldsymbol{A}^{-1}\right);A=2\boldsymbol{Z}^{*'}\boldsymbol{Z}+1,b=\boldsymbol{Z}^{*'}\left(\boldsymbol{Z}_1-\boldsymbol{X}_1\boldsymbol{\Lambda}_1\right)+\boldsymbol{Z}^{*'}\left(\boldsymbol{Z}_2-\boldsymbol{X}_2\boldsymbol{\Lambda}_2\right)
\end{aligned}$$

Step 2: For individuals in state 2 at W_i, sample the unknown interval of entry into state two

$$\begin{aligned}
&\Pr\left(l_i^{(1)} = j | T_i^{(1)} \leq W_i, T_i^{(2)} > W_i, \alpha^{(1)}, \alpha^{(2)}\right) \\
&\quad \propto \; I\left(j \leq m_i\right) \prod_{h=1}^{j-1} \left(1 - \alpha_{ih}^{(1)}\right) \alpha_{ij}^{(1)} \prod_{h=j+1}^{m_i} \left(1 - \alpha_{ih}^{(2)}\right)
\end{aligned}$$

For individuals in state three, sample both the interval of entry into state two and state three.

$$\begin{aligned}
&\Pr\left(l_i^{(1)} = j | l_i^{(2)} = s, \alpha^{(1)}, \alpha^{(2)}\right) \\
&\quad \propto \; I\left(j \leq s-1\right) \prod_{h=1}^{j-1} \left(1 - \alpha_{ih}^{(1)}\right) \alpha_{ij}^{(1)} \prod_{h=j+1}^{s-1} \left(1 - \alpha_{ih}^{(2)}\right) \alpha_{is}^{(2)} \\
&\Pr\left(l_i^{(2)} = j | l_i^{(1)} = r, \alpha^{(1)}, \alpha^{(2)}\right) \\
&\quad \propto \; I\left(r+1 \leq j \leq m_i\right) \prod_{h=1}^{r-1} \left(1 - \alpha_{ih}^{(1)}\right) \alpha_{ir}^{(1)} \prod_{h=r+1}^{j-1} \left(1 - \alpha_{ih}^{(2)}\right) \alpha_{ij}^{(2)}
\end{aligned}$$

Step 3: Generate latent outcome variables used for fitting probit regression models as outlined by Albert and Chib (1993). For the state one to two transition, let $T_{ij}^{(1)} = 1$ if the transition occurred in interval j and let $T_{ij}^{(1)} = 0$ if it did not yet occur. Similarly let $T_{ij}^{(2)} = 1$ if the state two to three transition occurred in interval j and $T_{ij}^{(2)} = 0$ otherwise and let $\text{SAB}_i = 1$ if subject i had a spontaneous abortion by week 20.

$$U_{ij} | \boldsymbol{M}_j, \boldsymbol{\omega}, Z_i^*, \gamma_1 \sim \begin{cases} N_1\left(\boldsymbol{M}_j \boldsymbol{\omega} + \gamma_1 Z_i^*, 1\right) I(U_{ij} > 0) & \text{if} \quad T_{ij}^{(1)} = 1 \\ N_1\left(\boldsymbol{M}_j \boldsymbol{\omega} + \gamma_1 Z_i^*, 1\right) I(U_{ij} < 0) & \text{if} \quad T_{ij}^{(1)} = 0 \end{cases}$$

$$V_{ij} | \nu, Z_i^*, \gamma_2 \sim \begin{cases} N_1\left(\nu + \gamma_2 Z_i^*, 1\right) I(V_{ij} > 0) & \text{if} \quad T_{ij}^{(2)} = 1 \\ N_1\left(\nu + \gamma_2 Z_i^*, 1\right) I(V_{ij} < 0) & \text{if} \quad T_{ij}^{(2)} = 0 \end{cases}$$

$$Q_i | \boldsymbol{\mu}, Z_i^* \sim \begin{cases} N_1\left(\mu_0 + \mu_1 Z_i^*, 1\right) I(Q_i > 0) & \text{if} \quad \text{SAB}_i = 1 \\ N_1\left(\mu_0 + \mu_1 Z_i^*, 1\right) I(Q_i < 0) & \text{if} \quad \text{SAB}_i = 0 \end{cases}$$

Step 4: Using the latent outcomes generated in the previous step, γ_1, $\boldsymbol{\omega}$, γ_2, ν, and $\boldsymbol{\mu}$ follow from linear regression results. Note that $\boldsymbol{U}_i$, $\boldsymbol{V}_i$, and $\boldsymbol{Z}^*$ are $R_i \times 1$, $S_i \times 1$ and $n \times 1$ vectors, respectively, while Z_i^* and Q_i are scalars.

$$[\gamma_1, \boldsymbol{\omega} | \boldsymbol{M}, \boldsymbol{Z}^*, \boldsymbol{U}] \propto \prod_{i=1}^{n} \phi_{R_i} \left(\boldsymbol{U}_i; \boldsymbol{M}\boldsymbol{\omega} + \gamma_1 \boldsymbol{Z}_i^*, \boldsymbol{I}_{R_i} \right) \phi_6 \left(\gamma_1, \boldsymbol{\omega}; \boldsymbol{\mu}_{0,\gamma_1\omega}, \boldsymbol{\Sigma}_{0,\gamma_1\omega} \right)$$

$$\sim \; N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1} \right)$$

$$\boldsymbol{A} = \sum_{i=1}^{n} [\boldsymbol{M} \quad \boldsymbol{1}_{R_i} \boldsymbol{Z}_i^*]^{\otimes 2} + \boldsymbol{\Sigma}_{0,\gamma_1\omega}^{-1}; \boldsymbol{b} = \sum_{i=1}^{n} [\boldsymbol{M} \quad \boldsymbol{1}_{R_i} Z_i^*]' \boldsymbol{U}_i + \boldsymbol{\Sigma}_{0,\gamma_1\omega}^{-1} \boldsymbol{\mu}_{0,\gamma_1\omega}$$

$$[\gamma_2, \nu | \boldsymbol{Z}^*, \boldsymbol{V}] \propto \prod_{i=1}^{n} \phi_{S_i} \left(\boldsymbol{V}_i; \nu \boldsymbol{1}_{S_i} + \gamma_2 Z_i^* \boldsymbol{1}_{S_i}, \boldsymbol{I}_{S_i} \right) \phi_2 \left(\gamma_2, \nu; \boldsymbol{\mu}_{0,\gamma_2\nu} \boldsymbol{\Sigma}_{0,\gamma_2\nu} \right)$$

$$\sim \; N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1} \right)$$

$$\boldsymbol{A} = \sum_{i=1}^{n} S_i \left(\nu + \gamma_2 Z_i^* \right)^2 + \boldsymbol{\Sigma}_{0,\gamma_2\nu}^{-1}; \boldsymbol{b} = \sum_{i=1}^{n} \left(\nu + \gamma_2 Z_i^* \right) \boldsymbol{1}_{S_i}' \boldsymbol{V}_i + \boldsymbol{\Sigma}_{0,\gamma_2\nu}^{-1} \boldsymbol{\mu}_{0,\gamma_2\nu}$$

$$[\boldsymbol{\mu} | \boldsymbol{Z}^*, \boldsymbol{Q},] \sim N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1} \right)$$

$$\boldsymbol{A} = [\boldsymbol{1}_n \quad \boldsymbol{Z}^*]^{\otimes 2} + \boldsymbol{\Sigma}_{0,\gamma_1\boldsymbol{\omega}}^{-1}; \boldsymbol{b} = [\boldsymbol{1}_n \quad \boldsymbol{Z}^*]' \boldsymbol{Q} + \boldsymbol{\Sigma}_{0,\mu}^{-1} \boldsymbol{\mu}_{0,\mu}$$

Step 5: We use day-specific conception prior probabilities taken from Table 1 of Wilcox et al. (2001) to specify the same multinomial priors, $\pi\left(t_i^I\right)$, on each t_i^I. The complete conditional for t_i^I will follow a multinomial distribution with probabilities proportional to

$$\Pr\left(t_i^I = k | Z_{1i}, Z_{2i}, \tau_{z_1}, \tau_{z_2}, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, Z_i^*, \boldsymbol{X}_1, \boldsymbol{X}_2 \right), k = 1, \ldots, 80$$

$$\propto \; \pi\left(t_i^I\right) \phi_1 \left(Z_{1i}; \boldsymbol{X}_{1i} \boldsymbol{\Lambda}_1, \tau_{z_1}^{-1} | t_i^I = k \right) \phi_1 \left(Z_{2i}; \boldsymbol{X}_{2i} \boldsymbol{\Lambda}_2, \tau_{z_2}^{-1} | t_i^I = k \right)$$

where t_i^I appears in design matrices $\boldsymbol{X}_{1i}$ and $\boldsymbol{X}_{2i}$ for subject i.

Step 6: The complete conditional for Z_i^* is the product of Normal densities from the latent outcomes variable generated in step 3 and Normal densities for Z_{1i} and Z_{2i}. Each Z_i^* can be sampled from a Normal distribution with subject-specific means and variances.

Letting $k_{1i} = Z_{1i} - (\lambda_{01} + \lambda_{21} W_i + \lambda_{31} W_i^2)$ and $k_{2i} = Z_{2i} - (\lambda_{02} + \lambda_{22} W_i + \lambda_{32} W_i^2)$

$$\begin{aligned}
& [Z_i^* | Z_{1i}, Z_{2i}, \tau_{z_1}, \tau_{z_2}, \tau_{z^*}, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\beta}, W_i, \boldsymbol{U}_i, \boldsymbol{V}_i, Q_i, \boldsymbol{\mu}] \\
\propto \quad & \phi_1 \left(Z_i^*; \boldsymbol{X}_i \boldsymbol{\beta}, \tau_{z^*}^{-1} \right) \phi_1 \left(Z_{1i}; \boldsymbol{X}_{1i} \boldsymbol{\Lambda}_1, \tau_{z_1}^{-1} \right) \phi_1 \left(Z_{2i}; \boldsymbol{X}_{2i} \boldsymbol{\Lambda}_2, \tau_{z_2}^{-1} \right) \\
& \times \phi_{R_i} \left(\boldsymbol{U}_i; \boldsymbol{M\omega} + \gamma_1 Z_i^* \mathbf{1}_{R_i}, \boldsymbol{I}_{R_i} \right) \phi_{S_i} \left(\boldsymbol{V}_i; \nu + \gamma_2 Z_i^* \mathbf{1}_{S_i}, \boldsymbol{I}_{S_i} \right) \\
& \times \phi_1 \left(Q_i; \mu_0 + \mu_1 Z_i^*, \tau_m \right) \\
\sim \quad & N_1 \left(\frac{b_i}{a_i}, \frac{1}{a_i} \right) \qquad a_i = \tau_{z^*} + \tau_{z_1} \lambda_{11}^2 + \tau_{z_2} \lambda_{12}^2 + R_i \gamma_1^2 + S_i \gamma_2^2 + \mu_1^2 \\
& b_i = \tau_{z^*} \boldsymbol{X\beta} + \tau_{z_1} \lambda_{11} k_{1i} + \tau_{z_2} \lambda_{12} k_{2i} + \mathbf{1}'_{R_i} \left(\boldsymbol{U}_i - \boldsymbol{M\omega} - \mathbf{1}_{R_i} \gamma_1 \right) \gamma_1 \\
& + \mathbf{1}'_{S_i} \left(\boldsymbol{V}_i - \mathbf{1}_{S_i} \nu - \mathbf{1}_{S_i} \gamma_2 \right) \gamma_2 + \mu_1 \left(Q_i - \mu_0 \right)
\end{aligned}$$

Step 7: Conditional on $\boldsymbol{Z}^*$, sample $\boldsymbol{\beta}$ and τ_{z^*} according to (4.6) using linear regression results.

$$\begin{aligned}
[\boldsymbol{\beta} | \boldsymbol{X}, \tau_{z^*}] \quad & \sim \quad N \left(\boldsymbol{A}^{-1} \boldsymbol{b}, \tau_{z_1}^{-1} \boldsymbol{A}^{-1} \right) \\
& \qquad \boldsymbol{A} = \boldsymbol{X}' \boldsymbol{X} + \boldsymbol{\Sigma}_{0,\beta}^{-1}; \boldsymbol{b} = \boldsymbol{X}' \boldsymbol{Z}^* + \boldsymbol{\Sigma}_{0,\beta}^{-1} \boldsymbol{\mu}_{0,\beta} \\
[\tau_{z^*} | \boldsymbol{\beta}, \boldsymbol{X}] \quad & \sim \quad G \left(\frac{n + \delta_{0,z^*}}{2}, \frac{1}{2} \left[\left(\boldsymbol{Z}^* - \boldsymbol{X\beta} \right)^{\otimes 2} + \lambda_{0,u} \right] \right)
\end{aligned}$$

B Appendix for paper 2

Measurement Model

We express the relationship between observed outcomes and latent variables in the following measurement model

$$y_{i1} = \lambda_{01} + \lambda_{11}\eta_{i1} + \gamma_{11}W_i + \gamma_{21}W_i^2 + \epsilon_{i1}, \quad \epsilon_{i1} \sim N(0, \tau_1^{-1}) \tag{B.1}$$

$$y_{i2} = \lambda_{02} + \lambda_{12}\eta_{i1} + \gamma_{12}W_i + \gamma_{22}W_i^2 + \epsilon_{i2}, \quad \epsilon_{i2} \sim N(0, \tau_2^{-1}) \tag{B.2}$$

$$y_{i3} = \lambda_{03} + \lambda_{13}\eta_{i1} + \lambda_{23}\eta_{i2} + \epsilon_{i3}, \quad \epsilon_{i3} \sim N(0, \tau_3^{-1}) \tag{B.3}$$

$$y_{i4} = \lambda_{04} + \lambda_{14}\eta_{i1} + \lambda_{24}\eta_{i2} + \epsilon_{i4}, \quad \epsilon_{i4} \sim N(0, \tau_4^{-1}) \tag{B.4}$$

In our analysis, we allow for a quadratic effect for W_i, the reported time from the last menstrual period (LMP) to the ultrasound for subject i (Hadlock et al. 1992).

Mixture Distribution Model

The mixture distributions are formally specified by

$$f(\eta_{i1}|\boldsymbol{\mu}_1, \boldsymbol{\tau}_1) \sim N(\mu_{11}, \tau_{11}^{-1}) \tag{B.5}$$

$$f(\eta_{i2}|\boldsymbol{\mu}_2, \boldsymbol{\tau}_2, \boldsymbol{\pi}_2) \sim \pi_{21}N(\mu_{21}, \tau_{21}^{-1}) + \pi_{22}N(\mu_{22}, \tau_{22}^{-1}) + \pi_{23}N(\mu_{23}, \tau_{23}^{-1}) \tag{B.6}$$

where $N(\mu, \tau^{-1})$ is the normal distribution with mean μ and variance τ^{-1} and, for η_{i2}, the mixing proportions $\boldsymbol{\pi}_2 = [\pi_{21}, \pi_{22}, \pi_{23}]'$ are fixed to sum to one.

It is convenient to express mixture models using a missing data formulation in which individual latent growth restriction is presumed to arise from a specific, but unknown, underlying component (Dempster et al. 1977). Specifically, we introduce latent class allocation variables $T_{i2} \in \{1, 2, 3\}$ where $\Pr(T_{i2} = k_2) = \pi_{2,k_2}$. This specification is useful for computational purposes because, if we additionally condition on T_{i2}, (B.6) becomes

$f(\eta_{i2}|\boldsymbol{\mu}_2, \boldsymbol{\tau}_2, \boldsymbol{\pi}_2, T_{i2} = k_2) \sim N(\mu_{2,k_2}, \tau_{2,k_2}^{-1})$. The missing data formulation also allows us to naturally group subjects with similar latent variable characteristics. We add the restriction $\mu_{21} < \mu_{22} < \mu_{23}$ to ensure that subjects who are more likely to be born early will be assigned to latent class $T_{i2} = 1$. We then probit regression to estimate the association of black race, current smoking status, maternal age being greater than or equal to 35 years, parity, and education level with the probability of belonging to the immaturity group. Our main interest is in modeling these associations.

Identifiability Restrictions

Latent variable models require fixing some parameters so that the model will be identifiable (e.g. Bollen 1989). Different choices can be made regarding the parameters to constrain, which will change the interpretations but will not impact the overall model fit. In our analysis, with four observed outcomes $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_4)$, we are limited to three or fewer latent variables $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. We demonstrate two methods of placing restrictions to insure model identifiability while providing readily interpretable results. For $\boldsymbol{\eta}_1$, we fix the mean at zero ($\mu_{11} = 0$) so that the intercepts $\lambda_{0j}, j = 1, \ldots, 4$ can be identified. Using the additional restrictions $\lambda_{11} > 0$ and $\lambda_{12} > 0$ then implies that subjects with positive values of η_{i1} are developing relatively quickly while individuals with negative values are slow growers. We fix the variance parameter for $\boldsymbol{\eta}_1$, $\tau_{11}^{-1} = 1$, to easily interpret the scale of $\boldsymbol{\eta}_1$ and insure that the factor loadings λ_{1j} are identifiable. For $\boldsymbol{\eta}_2$, we estimate the mixture component means and variances while fixing $\lambda_{04} = 0$ and $\lambda_{24} = 1$. Using this restriction, the latent $\boldsymbol{\eta}_2$ has location and scale commensurate with $\boldsymbol{y}_4$, which aids in interpretation and specifying appropriate prior distributions for μ_{2k} and $\tau_{2k}, k = 1, 2, 3$. In the model for $\boldsymbol{y}_3$, we estimate the factor loading λ_{23} to allow $\boldsymbol{\eta}_2$ to change scale, and the intercept, λ_{03}, is estimated for a shift in location.

Mixture distribution models are prone to their own identifiability problems, which we will consider in the context of our pregnancy analysis. Let $\boldsymbol{s}_i$ be a 3-dimensional vector indicating group membership for η_{i2}, such that $s_{ij} \in 0, 1$ and $\sum_{j=1}^{3} s_{ij} = 1$. The complete data likelihood for η_{i2} is $\prod_{j=1}^{3} \pi_{2j}^{s_{ij}} f(\eta_{i2}; \mu_j, \tau_j)^{s_{ij}}$ so that it is not possible to discriminate between the 3! possible permutations of group membership. Previously, we specified the restriction $\mu_{21} < \mu_{22} < \mu_{23}$ to order class membership by the mean; this restriction also removes the label switching problem. Additionally, with mixture models there is probability $(1 - \pi_{2j})^n$ that there are no observations from the jth component which leads to both identifiability and computational problems. We contend with this issue by assuming group membership is known for each of two subjects with the lowest, median, and highest values of observed gestational age.

Prior Specification

To complete a Bayesian specification of the model, prior distributions must be specified for each parameter. In general, we use proper but appropriately vague priors for all parameters. For notational simplicity, let $\boldsymbol{\alpha}_1 = [\lambda_{01}, \lambda_{11}, \gamma_{11}, \gamma_{21}]'$, $\boldsymbol{\alpha}_2 = [\lambda_{02}, \lambda_{12}, \gamma_{12}, \gamma_{22}]'$, $\boldsymbol{\alpha}_3 = [\lambda_{03}, \lambda_{13}, \lambda_{23}]'$, and $\boldsymbol{\alpha}_4 = [\lambda_{04}, \lambda_{14}, \lambda_{24}]'$ be $p_j \times 1$ vectors $(j = 1, \ldots, 4)$ with $n \times p_j$ design matrices $\boldsymbol{X}_1 = \boldsymbol{X}_2 = \left[\mathbf{1}, \boldsymbol{\eta}_1, \boldsymbol{W}, \boldsymbol{W}^2\right]$, and $\boldsymbol{X}_3 = \boldsymbol{X}_4 = [\mathbf{1}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2]$. The measurement model equations given in (B.1) - (B.4) can then be expressed in vector form as $\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_j$. We use conditionally conjugate priors $p\left(\boldsymbol{\alpha}_j | \tau_j\right) \sim N_{p_j}\left(\boldsymbol{\mu}_{0,\alpha_j}, \tau_j^{-1} \boldsymbol{\Sigma}_{0,\alpha_j}\right)$ and $p\left(\tau_j\right) \sim \Gamma\left(\frac{c_j}{2}, \frac{d_j}{2}\right)$ with $\boldsymbol{\mu}_{0,\alpha_j} = \mathbf{0}$, $\boldsymbol{\Sigma}_{0,\alpha_j} = 1000^2 \boldsymbol{I}_{p_j}$, and $c_j = d_j = 1$ for every j where $\mathbf{0}$ is a conforming vector of zeros and $\boldsymbol{I}_p$ is the identity matrix of rank p. Similarly, we assume $p(\boldsymbol{\beta}) \sim N_r\left(\boldsymbol{\mu}_{0,\beta}, \boldsymbol{\Sigma}_{0,\beta}\right)$ with $\boldsymbol{\mu}_{0,\beta} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{0,\beta} = 100^2 \boldsymbol{I}_r$.

For the mixture distribution component of our model, we use a prior specification that follows the suggestions of Richardson and Green (1997). For $k = 1, 2, 3$ we assume $p\left(\mu_{2k}\right) \sim N\left(\nu_k, R^2\right) I\left(\mu_{2,k-1} < \mu_{2k} < \mu_{2,k+1}\right)$ where $\mu_{20} = -\infty$ and $\mu_{24} = \infty$. We choose $\nu_1 = 245$, $\nu_2 = \nu_3 = 280$, and $R = 10$ so that, *a priori*, we expect that the residual distribution will have a mean of 245 ± 20 (days) and the predominant distribution a mean of 280 ± 20. We use a hierarchical structure for specifying the prior distribution of each τ_{2k}. Specifically, we allow $p\left(\tau_{2k}|b_0\right) \sim \Gamma\left(a_0, b_0\right)$ and $b_0 \sim \Gamma\left(g_0, h_0\right)$ with $a_0 = 2$, $g_0 = 0.2$, and $h_0 = 10 * R^{-2}$ where $\Gamma(a, b)$ is the gamma distribution with mean $a \div b$ and variance $a \div b^2$. By choosing $a_0 > 1 > g_0$ we express the general belief that for each k, the τ_{2k} are similar, but we have no information on their absolute size. Finally, we assume that $\boldsymbol{\pi}_2 = [\pi_{21}, \pi_{22}, \pi_{23}]'$ follows a symmetric Dirichlet distribution, $p\left(\boldsymbol{\pi}_2\right) \sim D(\delta, \ldots, \delta)$ and choose $\delta = 1$ to be appropriately vague.

Complete Conditionals

We consider the complete conditionals for the fetal development model specified in sections 3 and 4. For notational convenience, let $\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal probability density function for the p dimensional random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Also, let $\psi(x; \alpha, \beta) \propto x^{\alpha-1}e^{-x\beta}$ be the gamma probability density function for random variable x. Gibbs sampling proceeds by iterating the following complete conditionals until convergence. For matrix calculations, let $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}'\boldsymbol{A}$, $\boldsymbol{1}_n$ be a $n \times 1$ vector of ones and $\boldsymbol{I}_n$ the identity matrix of rank n.

Step 1: Using conjugate priors, $\boldsymbol{\alpha}_j$ and τ_j have normal and Gamma posterior distributions.

$$\begin{aligned}
& f\left(\boldsymbol{\alpha}_j|\boldsymbol{X}_j,\tau_j,\boldsymbol{y}_j\right) \propto \phi_n\left(\boldsymbol{y}_j;\boldsymbol{X}_j\boldsymbol{\alpha}_j,\tau_j^{-1}\boldsymbol{I}_n\right)\phi_{p_j}\left(\boldsymbol{\alpha}_j;\boldsymbol{\mu}_{0,\alpha_j},\boldsymbol{\Sigma}_{0,\alpha_j}\right) \\
& \sim \quad N_{p_j}\left(\boldsymbol{A}^{-1}\boldsymbol{b},\tau_j^{-1}\boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{X}_j'\boldsymbol{X}_j + \boldsymbol{\Sigma}_{0,\alpha_j}^{-1}, \boldsymbol{b} = \boldsymbol{X}_j'\boldsymbol{y}_j + \boldsymbol{\Sigma}_{0,\alpha_j}^{-1}\boldsymbol{\mu}_{0,\alpha_j} \\
& f\left(\tau_j|\boldsymbol{X}_j,\boldsymbol{\alpha}_j,\boldsymbol{y}_j\right) \propto \phi_n\left(\boldsymbol{y}_j;\boldsymbol{X}_j\boldsymbol{\alpha}_j,\tau_j^{-1}\boldsymbol{I}_n\right)\psi\left(\tau_j;\frac{c_j}{2},\frac{d_j}{2}\right) \\
& \sim \quad G\left(\frac{n+c_j}{2},\frac{1}{2}\left[\left(\boldsymbol{y}_j-\boldsymbol{X}_j\boldsymbol{\alpha}_j\right)^{\otimes 2}+d_j\right]\right)
\end{aligned}$$

Step 2: A Dirichlet prior for the selection probabilities gives a Dirichelt posterior distribution. Letting $n_k = \#\{i : T_{i2} = k\}, k = 1,2,3$ and $\boldsymbol{\pi}_2 = [\pi_{21},\pi_{22},\pi_{23}]'$

$$f\left(\boldsymbol{\pi}_2|n_1,n_2,n_3\right) \propto \pi_{21}^{\delta-1}\pi_{22}^{\delta-1}\pi_{23}^{\delta-1} \times \pi_{21}^{n_1}\pi_{22}^{n_2}\pi_{23}^{n_3} \sim \quad D(n_1+\delta,n_2+\delta,n_3+\delta)$$

Step 3: Sample latent group membership, $\boldsymbol{T}_2 = [T_{12},T_{22},\ldots,T_{n2}]'$ for each subject from a multinomial distribution with probabilities

$$\Pr\left(T_{i2}=k|\eta_{i2},\mu_{2k},\tau_{2k}\right) \propto \pi_{2k}\phi_1\left(\eta_{2i};\mu_{2k},\tau_{2k}^{-1}\right)$$

Step 4: Sample the mixture components means and variances, repeating for $k = 1,2,3$. Also sample the hyperparameter b_0 for a Gamma posterior.

$$\begin{aligned}
& f\left(\mu_{2k}|\boldsymbol{\eta}_2,\tau_{2k},\boldsymbol{T}_2\right) \propto \phi_1\left(\mu_{2k};\mu_{0,k},R^2\right)\prod_{i:T_{i2}=k}\phi_1\left(\eta_{i2};\mu_{2k},\tau_{2k}^{-1}\right) \\
& \sim \quad N_1\left(a^{-1}b,a^{-1}\right); a = R^{-2}+\tau_{2k}n_k, b = \mu_{0,k}R^{-2}+\tau_{2k}\sum_{i:T_{i2}=k}\eta_{i2} \\
& f\left(\tau_{2k}|\boldsymbol{\eta}_2,\mu_{2k},\boldsymbol{T}_2,b_0\right) \propto \psi\left(\tau_{2k};a_0,b_0\right)\prod_{i:T_{i2}=k}\phi_1\left(\eta_{i2};\mu_{2k},\tau_{2k}^{-1}\right) \\
& \propto \quad \Gamma\left(\frac{2a_0+n_k}{2},b_0+\frac{1}{2}\sum_{i:T_{i2}=k}\left(\eta_{i2}-\mu_{2k}\right)^2\right) \\
& f\left(b_0|\tau_{21},\tau_{22},\tau_{23}\right) \propto \psi\left(b_0;g_0,h_0\right)\psi\left(\tau_{21};a_0,b_0\right)\psi\left(\tau_{22};a_0,b_0\right)\psi\left(\tau_{23};a_0,b_0\right) \\
& \sim \quad \Gamma(3a_0+g_0,h_0+\tau_{21}+\tau_{22}+\tau_{23})
\end{aligned}$$

Step 5: Generate latent outcome variables used for fitting probit regression models as outlined by Albert and Chib. If any of $\boldsymbol{y}_j$ are binary, this procedure could be used to generate $\boldsymbol{y}$. In our application, we generate the latent outcomes $\boldsymbol{T}_2^* = [T_{12}^*, T_{22}^*, \ldots, T_{n2}^*]'$ from a truncated normal distribution to then sample $\boldsymbol{\beta}$ using linear regression results.

$$f\left(T_{i2}^*|\boldsymbol{X}, \boldsymbol{\beta}\right) \sim \begin{cases} N_1\left(\boldsymbol{X\beta}, 1\right) I(T_{i2}^* > 0) & \text{if} \quad T_{i2} > 1 \\ N_1\left(\boldsymbol{X\beta}, 1\right) I(T_{i2}^* < 0) & \text{if} \quad T_{i2} = 1 \end{cases}$$

$$[\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{T}_2^*] \propto \phi_n\left(\boldsymbol{T}_2^*; \boldsymbol{X\beta}, 1\right)\phi_r\left(\boldsymbol{\beta}; \boldsymbol{\mu}_{0,\beta}, \boldsymbol{\Sigma}_{0,\beta}\right)$$

$$\sim \quad N_r\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Sigma}_{0,\beta}^{-1}; \boldsymbol{b} = \boldsymbol{X}'\boldsymbol{T}_2^* + \boldsymbol{\Sigma}_{0,\beta}^{-1}\boldsymbol{\mu}_{0,\beta}$$

Step 6: Letting $k_{ij} = y_{ij} - \boldsymbol{\alpha}_j\boldsymbol{X}_j + \lambda_{1j}\eta_{i1}, j = 1, \ldots, 4$ and using the constraints needed for identifiability, generate the latent predictor variables for each $i = 1, \ldots, n$.

$$\begin{aligned} f\left(\eta_{i1}|-\right) &\propto \phi_1\left(\eta_{i1}; 0, 1\right)\prod_{j=1}^4 \phi_1\left(y_{ij}; \boldsymbol{X}_j\boldsymbol{\alpha}_j, \tau_j^{-1}\right) \\ &\sim N\left(a_i^{-1}b_i, a_i^{-1}\right); a_i = 1 + \sum_{j=1}^4 \lambda_{1j}^2\tau_j, b_i = \sum_{j=1}^4 \lambda_{1j}k_{ij}\tau_j \\ f\left(\eta_{i2}|-, T_i = k\right) &\propto \phi_1\left(\eta_{i2}; \mu_{2k}, \tau_{2k}^{-1}\right)\phi_1\left(y_{i3}; \boldsymbol{X}_3\boldsymbol{\alpha}_3, \tau_3^{-1}\right)\phi_1\left(y_{i4}; \boldsymbol{X}_4\boldsymbol{\alpha}_4, \tau_4^{-1}\right) \\ &\sim N\left(a_i^{-1}b_i, a_i^{-1}\right); a_i = \tau_{2k} + \lambda_{23}^2\tau_3 + \tau_4 \\ & \quad\; b_i = \mu_{2k}\tau_{2k} + \lambda_{23}\tau_3\left(y_{i3} - \lambda_{03} - \lambda_{13}\eta_{i1}\right) + \tau_4\left(y_{i4} - \lambda_{14}\eta_{i1}\right) \end{aligned}$$

C Appendix for paper 3

We consider the complete conditionals for the model specified in sections 2 and 3. For notational convenience, let $\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal probability density function for the p dimensional random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Also, let $\psi(x; \alpha, \beta) \propto x^{\alpha-1}e^{-x\beta}$ be the gamma probability density function for random variable x. Gibbs sampling proceeds by iterating the complete conditionals until convergence criteria are met. For matrix calculations, let $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}'\boldsymbol{A}$, $\mathbf{1}_n$ be a $n \times 1$ vector of ones and $\boldsymbol{I}_n$ the identity matrix of rank n.

Step 1: Let $\boldsymbol{y}_j = [y_{i1}, \ldots, y_{in}]'_{n\times 1}$, $\boldsymbol{W}_j = [W_{i1}, \ldots, W_{in}]'_{n\times 1}$, $\boldsymbol{\eta}_1 = [\boldsymbol{\eta}_{11}, \ldots, \boldsymbol{\eta}_{n1}]'_{n\times 3}$, and $\boldsymbol{X}_j = [\mathbf{1}_n, \boldsymbol{W}_j, \boldsymbol{W}_j^2]$ for $j = 1, \ldots, 8$ and $\boldsymbol{X}_j = \mathbf{1}_n$ for $j = 9, \ldots, 18$. Using conjugate priors, $\boldsymbol{\Lambda}_j$, $\boldsymbol{\Gamma}_j$ and $\tau_{y,j}$ have normal and gamma posterior distributions.

$$
\begin{aligned}
f\left(\boldsymbol{\Lambda}_j|\boldsymbol{\Gamma}_j, \boldsymbol{\eta}_1, \boldsymbol{\Lambda}_j, \tau_{y,j}, \boldsymbol{y}_j\right) &\propto \phi_n\left(\boldsymbol{y}_j; \boldsymbol{X}_j\boldsymbol{\Gamma}_j + \boldsymbol{\eta}_1\boldsymbol{\Lambda}_j, \tau_{y,j}^{-1}\boldsymbol{I}_n\right)\phi\left(\boldsymbol{\Lambda}_j; \boldsymbol{\mu}_{0,\Lambda_j}, \tau_{y,j}^{-1}\boldsymbol{\Sigma}_{0,\Lambda_j}\right)\\
&\sim \; N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \tau_{y,j}^{-1}\boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{\eta}_1'\boldsymbol{\eta}_1 + \boldsymbol{\Sigma}_{0,\Lambda_j}^{-1}, \boldsymbol{b} = \boldsymbol{\eta}_1'(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma}_j) + \boldsymbol{\Sigma}_{0,\Lambda_j}^{-1}\boldsymbol{\mu}_{0,\Lambda_j}\\
f\left(\boldsymbol{\Gamma}_j|\boldsymbol{\Gamma}_j, \boldsymbol{\eta}_1, \boldsymbol{\Lambda}_j, \tau_{y,j}, \boldsymbol{y}_j\right) &\propto \phi_n\left(\boldsymbol{y}_j; \boldsymbol{X}_j\boldsymbol{\Gamma}_j + \boldsymbol{\eta}_1\boldsymbol{\Lambda}_j, \tau_{y,j}^{-1}\boldsymbol{I}_n\right)\phi\left(\boldsymbol{\Gamma}_j; \boldsymbol{\mu}_{0,\Gamma_j}, \tau_{y,j}^{-1}\boldsymbol{\Sigma}_{0,\Gamma_j}\right)\\
&\sim \; N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \tau_{y,j}^{-1}\boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{X}_j'\boldsymbol{X}_j + \boldsymbol{\Sigma}_{0,\Gamma_j}^{-1}, \boldsymbol{b} = \boldsymbol{X}_j'(\boldsymbol{y}_j - \boldsymbol{\eta}_1\boldsymbol{\Lambda}_j) + \boldsymbol{\Sigma}_{0,\alpha_j}^{-1}\boldsymbol{\mu}_{0,\alpha_j}\\
f\left(\tau_{y,j}|\boldsymbol{\Gamma}_j, \boldsymbol{\eta}_1, \boldsymbol{\Lambda}_j, \boldsymbol{y}_j\right) &\propto \phi_n\left(\boldsymbol{y}_j; \boldsymbol{X}_j\boldsymbol{\Gamma}_j + \boldsymbol{\eta}_1\boldsymbol{\Lambda}_j, \tau_{y,j}^{-1}\boldsymbol{I}_n\right)\psi\left(\tau_j; \frac{c_{y,j}}{2}, \frac{d_{y,j}}{2}\right)\\
&\sim \; G\left(\frac{n + c_{y,j}}{2}, \frac{1}{2}\left[\left(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma}_j - \boldsymbol{\eta}_1\boldsymbol{\Lambda}_j\right)^{\otimes 2} + d_{y,j}\right]\right)
\end{aligned}
$$

Step 2: Let $\boldsymbol{S} = [S_1, \ldots, S_n]'_{n\times 1}$, $\boldsymbol{z}_g = [z_{1g}, \ldots, z_{ng}]'_{n\times 1}$, $\boldsymbol{\eta}_2 = [\eta_{i21}, \ldots, \eta_{n2}]'_{n\times 1}$, and $\boldsymbol{X}_z = [\mathbf{1}_n, \boldsymbol{\eta}_2]_{n\times 2}$. Also, let $\boldsymbol{X}_s = [I(\boldsymbol{S} = 1), I(\boldsymbol{S} = 2)]'_{n\times 2}$ where $I(\boldsymbol{S} = k)$ is a $n \times 1$ indicator vector with row i equal to one if $S_i = k$ and zero otherwise. Using conjugate priors, $\boldsymbol{\Theta}_g$, $\boldsymbol{\beta}$ and $\tau_{z,g}$ have normal and gamma posterior distributions.

$$\begin{aligned}
f\left(\boldsymbol{\Theta}_g|\boldsymbol{X}_z, \tau_{z,g}, \boldsymbol{z}_g\right) &\propto \phi_n\left(\boldsymbol{z}_g; \boldsymbol{X}_z\boldsymbol{\Theta}_g, \tau_{z,g}^{-1}\boldsymbol{I}_n\right)\phi\left(\boldsymbol{\Theta}_g; \boldsymbol{\mu}_{0,\Theta_g}, \tau_{z,g}^{-1}\boldsymbol{\Sigma}_{0,\Theta_g}\right), g = 1,2\\
&\sim\ N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \tau_{z,g}^{-1}\boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{X}'_z\boldsymbol{X}_z + \boldsymbol{\Sigma}_{0,\Theta_g}^{-1}, \boldsymbol{b} = \boldsymbol{X}'_z\boldsymbol{z}_g + \boldsymbol{\Sigma}_{0,\Theta_g}^{-1}\boldsymbol{\mu}_{0,\Theta_g}\\
f\left(\tau_{z,g}|\boldsymbol{X}_z, \boldsymbol{\Theta}_g, \boldsymbol{z}_g\right) &\propto \phi_n\left(\boldsymbol{z}_g; \boldsymbol{X}_z\boldsymbol{\Theta}_g, \tau_{z,g}^{-1}\boldsymbol{I}_n\right)\psi\left(\tau_{z,g}; \frac{c_{z,g}}{2}, \frac{d_{z,g}}{2}\right), g = 1,2\\
&\sim\ G\left(\frac{n + c_{z,g}}{2}, \frac{1}{2}\left[\left(\boldsymbol{z}_g - \boldsymbol{X}_z\boldsymbol{\Theta}_g\right)^{\otimes 2} + d_{z,g}\right]\right)\\
f\left(\boldsymbol{\beta}|\tau_{z,3}, \boldsymbol{z}_3, \boldsymbol{X}_s\right) &\propto \phi_n\left(\boldsymbol{z}_3; \boldsymbol{X}_s\boldsymbol{\beta}, \tau_{z,3}^{-1}\boldsymbol{I}_n\right)\phi\left(\boldsymbol{\beta}; \boldsymbol{\mu}_{0,\beta}, \tau_{z,g}^{-1}\boldsymbol{\Sigma}_{0,\beta}\right)\\
&\sim\ N\left(\boldsymbol{A}^{-1}\boldsymbol{b}, \tau_{y,j}^{-1}\boldsymbol{A}^{-1}\right); \boldsymbol{A} = \boldsymbol{X}'_s\boldsymbol{X}_s + \boldsymbol{\Sigma}_{0,\beta}^{-1}, \boldsymbol{b} = \boldsymbol{X}'_s\boldsymbol{z}_3 + \boldsymbol{\Sigma}_{0,\beta}^{-1}\boldsymbol{\mu}_{0,\beta}\\
f\left(\tau_{z,3}|\boldsymbol{X}_s, \boldsymbol{\beta}, \boldsymbol{z}_3\right) &\propto \phi_n\left(\boldsymbol{z}_3; \boldsymbol{X}_s\boldsymbol{\beta}, \tau_{z,3}^{-1}\boldsymbol{I}_n\right)\psi\left(\tau_{z,3}; \frac{c_{z,3}}{2}, \frac{d_{z,3}}{2}\right)\\
&\sim\ G\left(\frac{n + c_{z,3}}{2}, \frac{1}{2}\left[\left(\boldsymbol{z}_3 - \boldsymbol{X}_s\boldsymbol{\beta}\right)^{\otimes 2} + d_{z,3}\right]\right)
\end{aligned}$$

Step 3: A Dirichlet prior for the selection probabilities gives a Dirichelt posterior distribution. Letting $n_{sk} = \#\{i : S_i = k\}, k = 1,2$ and $n_{tl} = \#\{i : T_i = l\}, l = 1,2,3$ with $\boldsymbol{\pi}_s = [\pi_{s1}, \pi_{s2}, \pi_{s3}]'$ and $\boldsymbol{\pi}_t = [\pi_{t1}, \pi_{t2}, \pi_{t3}]'$

$$f\left(\boldsymbol{\pi}_s|n_{s1}, n_{s2}\right) \propto \pi_{s1}^{d_1-1}\pi_{s2}^{d_1-1} \times \pi_{t1}^{n_{s1}}\pi_{t2}^{n_{s2}} \quad \sim \quad D(n_{s1} + d_1, n_{s2} + d_1)$$

$$f\left(\boldsymbol{\pi}_t|n_{t1}, n_{t2}, n_{t3}\right) \propto \pi_{t1}^{d_2-1}\pi_{t2}^{d_2-1}\pi_{t3}^{d_2-1} \times \pi_{t1}^{n_{t1}}\pi_{t2}^{n_{t2}}\pi_{t3}^{n_{t3}} \quad \sim \quad D(n_{t1} + d_2, n_{t2} + d_2, n_{t3} + d_2)$$

Step 4: Letting $\boldsymbol{\mu}_{1k} = [\mu_{11k}, \mu_{12k}, \mu 13k]$ and $\boldsymbol{\tau}_{1k} = [\tau_{11k}, \tau_{12k}, \tau_{13k}]$, sample latent group membership, S_i and T_i, for each subject from a multinomial distribution with probabilities

$$\begin{aligned}
\Pr\left(S_i = k|\boldsymbol{\eta}_{i1}, \boldsymbol{\mu}_{1k}, \boldsymbol{\tau}_{1k}\right) &\propto \pi_{tk}\phi_3\left(\boldsymbol{\eta}_{i1}; \boldsymbol{\mu}_{1k}, Dg\left(\boldsymbol{\tau}_{1k}^{-1}\right)\right)\phi_1\left(T_i^*; \boldsymbol{X}_s\boldsymbol{\alpha}, 1\right)\phi_1\left(z_{i3}; \boldsymbol{X}_s\boldsymbol{\beta}, \tau_{z,3}^{-1}\right)\\
\Pr\left(T_i = l|\eta_{i21}, \mu_{21l}, \tau_{21l}\right) &\propto \pi_{21l}\phi_1\left(\eta_{i21}; \mu_{21l}, \tau_{21l}^{-1}\right)
\end{aligned}$$

Step 4: Sample the mixture components means and variances, repeating for $k = 1,2$ and $l = 1,2,3$. Also sample the hyperparameters $b_{0,m}(m = 1,2,3)$ and b_0 from gamma posteriors.

$$
\begin{aligned}
f\left(\mu_{1mk}|\boldsymbol{\eta}_1,\tau_{1mk},\boldsymbol{S}\right) &\propto \phi_1\left(\mu_{1mk};\mu_{0,mk},R_m^2\right)\prod_{i:S_i=l}\phi_1\left(\eta_{i1m};\mu_{1mk},\tau_{1mk}^{-1}\right)\\
&\sim\ N_1\left(a^{-1}b,a^{-1}\right); a=R_m^{-2}+\tau_{1mk}n_{sk}, b=\mu_{0,mk}R_m^{-2}+\tau_{1mk}\sum_{i:S_i=k}\boldsymbol{\eta}_{i1}\\
f\left(\tau_{1mk}|\boldsymbol{\eta}_1,\mu_{1mk},\boldsymbol{S},b_{0,m}\right) &\propto \psi\left(\tau_{1mk};a_{0,m},b_{0,m}\right)\prod_{i:S_i=k}\phi_1\left(\eta_{i1m};\mu_{1mk},\tau_{1mk}^{-1}\right)\\
&\propto\ \Gamma\left(\frac{2a_{0,m}+n_{sk}}{2},b_{0,m}+\frac{1}{2}\sum_{i:S_i=k}\left(\eta_{ijk}-\mu_{1mk}\right)^2\right)\\
f\left(b_{0,m}|\tau_{1m1},\tau_{1m2}\right) &\propto \psi\left(b_{0,m};g_{0,m},h_{0,m}\right)\psi\left(\tau_{1m1};a_{0,m},b_{0,m}\right)\psi\left(\tau_{1m2};a_{0,m},b_{0,m}\right)\\
&\sim\ \Gamma(2a_{0,m}+g_{0,m},h_{0,m}+\tau_{1m1}+\tau_{1m2})\\
f\left(\mu_{21l}|\boldsymbol{\eta}_2,\tau_{21l},\boldsymbol{T}\right) &\propto \phi_1\left(\mu_{21l};\mu_{0,l},R^2\right)\prod_{i:T_i=l}\phi_1\left(\eta_{i21};\mu_{21l},\tau_{21l}^{-1}\right)\\
&\sim\ N_1\left(a^{-1}b,a^{-1}\right); a=R^{-2}+\tau_{21l}n_{tl}, b=\mu_{0,l}R^{-2}+\tau_{21l}\sum_{i:T_i=l}\eta_{i21}\\
f\left(\tau_{21l}|\boldsymbol{\eta}_2,\mu_{21l},\boldsymbol{T}_2,b_0\right) &\propto \psi\left(\tau_{21l};a_0,b_0\right)\prod_{i:T_i=l}\phi_1\left(\eta_{i21};\mu_{21l},\tau_{21l}^{-1}\right)\\
&\propto\ \Gamma\left(\frac{2a_0+n_{tl}}{2},b_0+\frac{1}{2}\sum_{i:T_i=l}\left(\eta_{i21}-\mu_{21l}\right)^2\right)\\
f\left(b_0|\tau_{211},\tau_{212},\tau_{213}\right) &\propto \psi\left(b_0;g_0,h_0\right)\psi\left(\tau_{211};a_0,b_0\right)\psi\left(\tau_{212};a_0,b_0\right)\psi\left(\tau_{213};a_0,b_0\right)\\
&\sim\ \Gamma(3a_0+g_0,h_0+\tau_{211}+\tau_{212}+\tau_{213})
\end{aligned}
$$

Step 5: Generate latent outcome variables used for fitting probit regression models as outlined by Albert and Chib. If any of $\boldsymbol{y}_j$ are binary, this procedure could be used to generate $\boldsymbol{y}$. In our application, we generate the latent outcomes $\boldsymbol{S}^*=[S_1^*,S_2^*,\ldots,S_n^*]'_{n\times 1}$ and $\boldsymbol{T}^*=[T_1^*,T_2^*,\ldots,T_n^*]'_{n\times 1}$ from a truncated normal distribution to then sample $\boldsymbol{\alpha}$ and

$\boldsymbol{\omega}$ using linear regression results.

$$f\left(S_i^*|\boldsymbol{x},\boldsymbol{\omega}\right) \sim \begin{cases} N_1\left(\boldsymbol{x}\boldsymbol{\omega},1\right)I(S_i^*>0) & \text{if} \quad S_i=2 \\ N_1\left(\boldsymbol{x}\boldsymbol{\omega},1\right)I(S_i^*<0) & \text{if} \quad S_i=1 \end{cases}$$

$$[\boldsymbol{\omega}|\boldsymbol{x},\boldsymbol{S}^*] \propto \phi_n\left(\boldsymbol{S}^*;\boldsymbol{x}\boldsymbol{\omega},1\right)\phi_r\left(\boldsymbol{\omega};\boldsymbol{\mu}_{0,\omega},\boldsymbol{\Sigma}_{0,\omega}\right)$$

$$\sim \quad N_r\left(\boldsymbol{A}^{-1}\boldsymbol{b},\boldsymbol{A}^{-1}\right);\boldsymbol{A}=\boldsymbol{x}'\boldsymbol{x}+\boldsymbol{\Sigma}_{0,\omega}^{-1};\boldsymbol{b}=\boldsymbol{x}'\boldsymbol{S}^*+\boldsymbol{\Sigma}_{0,\omega}^{-1}\boldsymbol{\mu}_{0,\omega}$$

$$f\left(T_i^*|\boldsymbol{X}_s,\boldsymbol{\alpha}\right) \sim \begin{cases} N_1\left(\boldsymbol{X}_s\boldsymbol{\alpha},1\right)I(T_i^*>0) & \text{if} \quad T_i>1 \\ N_1\left(\boldsymbol{X}_s\boldsymbol{\alpha},1\right)I(T_i^*<0) & \text{if} \quad T_i=1 \end{cases}$$

$$[\boldsymbol{\alpha}|\boldsymbol{X}_s,\boldsymbol{T}^*] \propto \phi_n\left(\boldsymbol{T}^*;\boldsymbol{X}_s\boldsymbol{\alpha},1\right)\phi\left(\boldsymbol{\alpha};\boldsymbol{\mu}_{0,\alpha},\boldsymbol{\Sigma}_{0,\alpha}\right)$$

$$\sim \quad N\left(\boldsymbol{A}^{-1}\boldsymbol{b},\boldsymbol{A}^{-1}\right);\boldsymbol{A}=\boldsymbol{X}_s'\boldsymbol{X}_s+\boldsymbol{\Sigma}_{0,\alpha}^{-1};\boldsymbol{b}=\boldsymbol{X}_s'\boldsymbol{T}^*+\boldsymbol{\Sigma}_{0,\alpha}^{-1}\boldsymbol{\mu}_{0,\alpha}$$

Step 6: Letting $k_{ij}=y_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\Gamma}_j$ and using the constraints needed for identifiability, generate the latent predictor variables for each $i=1,\ldots,n$.

$$\begin{aligned}
f\left(\eta_{i11}|-,S_i=k\right) &\propto \phi_1\left(\eta_{i11};\mu_{11k},\tau_{11k}\right)\prod_{j=1}^{8}\phi_1\left(y_{ij};\boldsymbol{X}_j\boldsymbol{\Gamma}_j+\lambda_{1j}\eta_{i11},\tau_{y,j}^{-1}\right) \\
&\sim N\left(a_i^{-1}b_i,a_i^{-1}\right);a_i=\tau_{11k}+\sum_{j=1}^{8}\lambda_{1j}^2\tau_{y,j},b_i=\sum_{j=1}^{8}\lambda_{1j}k_{ij}\tau_{y,j}+\tau_{11k}\mu_{11k} \\
f\left(\eta_{i12}|-,S_i=k\right) &\propto \phi_1\left(\eta_{i12};\mu_{12k},\tau_{12k}\right)\prod_{j=9}^{16}\phi_1\left(y_{ij};\boldsymbol{X}_j\boldsymbol{\Gamma}_j+\lambda_{1j}\eta_{i12},\tau_{y,j}^{-1}\right) \\
&\sim N\left(a_i^{-1}b_i,a_i^{-1}\right);a_i=\tau_{12k}+\sum_{j=9}^{16}\lambda_{1j}^2\tau_{y,j},b_i=\sum_{j=9}^{16}\lambda_{1j}k_{ij}\tau_{y,j}+\tau_{12k}\mu_{12k} \\
f\left(\eta_{i13}|-,S_i=k\right) &\propto \phi_1\left(\eta_{i13};\mu_{13k},\tau_{13k}\right)\prod_{j=17}^{18}\phi_1\left(y_{ij};\boldsymbol{X}_j\boldsymbol{\Gamma}_j+\lambda_{1j}\eta_{i13},\tau_{y,j}^{-1}\right) \\
&\sim N\left(a_i^{-1}b_i,a_i^{-1}\right);a_i=\tau_{13k}+\sum_{j=17}^{18}\lambda_{1j}^2\tau_{y,j},b_i=\sum_{j=17}^{18}\lambda_{1j}k_{ij}\tau_{y,j}+\tau_{13k}\mu_{13k} \\
f\left(\eta_{i21}|-,T_i=l\right) &\propto \phi_1\left(\eta_{i21};\mu_{21l},\tau_{21l}^{-1}\right)\phi_1\left(z_{i1};\theta_{01}+\theta_{11}\eta_{i21},\tau_{z,1}^{-1}\right)\phi_1\left(z_{i2};\theta_{02}+\theta_{12}\eta_{i21},\tau_{z,2}^{-1}\right) \\
&\sim N\left(a_i^{-1}b_i,a_i^{-1}\right);a_i=\tau_{21l}+\theta_{11}^2\tau_{z,1}+\theta_{12}^2\tau_{z,2} \\
& b_i=\mu_{21l}\tau_{21l}+\theta_{11}\tau_{z,1}\left(z_{i1}-\theta_{01}\right)+\theta_{12}\tau_{z,2}\left(z_{i2}-\theta_{02}\right)
\end{aligned}$$

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Albert, P. S. and Shih, J. H. (2003). Modeling tumor growth with random onset. *Biometrics*, 59:897–906.

Ansari, A. and Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, 65(4):475–496.

Arminger, G. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63(3):271–300.

Basu, A. P. (1983). *Encyclopedia of Statistical Sciences*, chapter 4, pages 2–6. Wiley Interscience, New York. S Kotz and N L Johnson, editors.

Bauwens, L. (1984). *Bayesian full information analysis of simultaneous equation models using integration by Monte Carlo.* Springer-Verlag, New York.

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley, New York.

Bracken, M. B., editor (1984). *Perinatal epidemiology.* Oxford University Press, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Satistical Association*, 88:9–25.

Brizot, M. L., Watanabe, L. C., Izzo, C. P. M., Pereira, P. P., and Miyadahira, S. (2001). *Ultrasound features in predicting early pregnancy loss*, chapter 15. Parthenon Publishing Group, New York. In *The Embryo as a Patient.* A Kurjak, FA Chervenak, and JM Carrera, editors.

Buekens, P., Notzon, F., Kotelchuck, M., and Wilcox, A. (2000). Why do Mexican Americans give birth to few low-birth-weight infants? *American Journal of Epidemiology*, 152(4):347 – 351.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall/CRC Press.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Catalano, P. and Ryan, L. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674.

Chib, S. and Carlin, B. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, 9(1):17–26.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–367.

Craig, B. A., Fryback, D. G., Klein, R., and Klein, B. E. K. (1999). A Bayesian approach to modeling the natural history of a chronic condition from observations with intervention. *Statistics in Medicine*, 18:1355–1371.

Curry, M., Perrin, N., and Wall, E. (1998). Effects of abuse on maternal complications and birth weight in adult and adolescent women. *Obstetrics and Gynecology*, 92(4):530 – 534.

David, R. J. (1980). The quality and completeness of birth weight and gestational age data in computerized birth files. *American Journal of Public Health*, 70(9):964 – 973.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B*, 41(1):1–31.

DeGruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with applications to AIDS. *Biometrics*, 45(1):1–11.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38.

Dewanji, A. and Kalbfleisch, J. D. (1986). Nonparametric methods for survival/sacrifice experiments. *Biometrics*, 42(2):325–341.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through

Bayesian sampling. *Journal of the Royal Statistical Society Series B*, 56(2):363 – 375.

Diggle, P. J., Lian, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Duffy, S. W., Chen, H. H., Tabar, L., and Day, N. E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*, 14(14):1531–1543.

Dugoff, L., Lynch, A., Cioffi-Ragan, D., Hobbins, J., Schultz, L., Malone, F., and D'alton, M. (2005). First trimester uterine artery doppler abnormalities predict subsequent intrauterine growth restriction. *American Journal of Obstetrics and Gynecology*, 193(3):1208 – 1212.

Dunson, D. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society Series B*, 62:355–366.

Dunson, D. B. (2006a). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7(4):551–568.

Dunson, D. B. (2006b). Bayesian methods for latent trait modeling of longitudinal data. *under review*.

Dunson, D. B. and Baird, D. D. (2002). Bayesian modeling of incidence and progression of disease from cross-sectional data. *Biometrics*, 58:813–822.

Elliott, M. R., Gallo, J. J., Have, T. R. T., Bogner, H. R., and Katz, I. R. (2005). Using a Bayesian latet growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics*, 6(1):119 – 143.

Filly, R. A. and Hadlock, F. P. (2000). *Ultrasonography in Obstetrics and Gynecology: Sonographic determination of menstrual age*, chapter 6, pages 146–170. WB Saunders Co.

Fokoue, E. (2005). Mixtures of factor analyzers: An extension with covariates. *Journal of Multivariate Analysis*, 95(2):370 – 384.

Gage, T. B. (2002). Modeling birthweight and gestational age distributions: Additive vs. multiplicative processes. *American Journal of Human Biology*, 14(6):728 – 734.

Gelfand, A., Sahu, S., and Carlin, B. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman and Hall, 2nd edition.

Gelman, A., Hwang, Z., van Dyk, D. A., and Boscardin, W. J. (2003). Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models. Technical report, Columbia University, www.stat.columbia.edu/∼gelman/.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments*. Clarendon Press. In *Bayesian Statistics 4*. JM Bernado, JO Berger, AP Dawid and AFM Smith, editors.

Glinianaia, S., Rankin, J., Bell, R., Pless-Mulloli, T., and Howel, D. (2004). Particulate air pollution and fetal health a systematic review of the epidemiologic evidence. *Epidemiology*, 15(1):36 – 45.

Guo, J., Wall, M., and Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1):145–163.

Hadlock, F. P., Shah, Y. P., Kanon, D. J., and Lindsey, J. V. (1992). Fetal crown-rump length: Reevaulation of relation to menstrual age (5-18 weeks) with high-resolution real-time US. *Radiology*, 182:501–505.

Harder, T., Rodekamp, E., Schellong, K., Dudenhausen, J. W., and Plagemann, A. (2007). Birth weight and subsequent risk of type 2 diabetes: A meta-analysis. *American*

Journal of Epidemiology, 165:849–857.

Hastings, W. (1970). Monte-carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

Hugo, E., Odendaal, H., and Grove, D. (2007). Evaluation of the use of umbilical artery doppler flow studies and outcome of pregnancies at a secondary hospital. *Journal of Maternal-Fetal and Neonatal Medicine*, 20(3):233 – 239.

Imai, K. and Van Dyk, D. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124(2):311–334.

Jewell, N. P. (2005). Correspondences between regression models for complex binary outcomes and those for structured multivariate survival analyses. *U.C. Berkeley Division of Biostatistics Working Paper Series*. http://www.bepress.com/ucbbiostat/paper195.

Joreskog, K. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251.

Joreskog, K. and Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351):631–639.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Interscience, New York.

Kalbfleish, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871.

Kiely, M., editor (1991). *Reproductive and perinatal epidemiology*. CRC Press, New York.

Kramer, M., Mclean, F., Boyd, M., and Usher, R. (1988). The validity of gestational-age estimation by menstrual dating in term, preterm, and postterm gestations. *Journal of the American Medical Association*, 260(22):3306–3308.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lee, S. and Xia, Y. (2006). Maximum likelihood methods in treating outliers and symmetrically heavy-tailed distributions for nonlinear structural equation models with missing data. *Psychometrika*, 71(3):565 – 585.

Lee, S. Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, 46:153–160.

Lee, S. Y. (1992). Bayesian-analysis of stochastic constraints in structural equation models. *British Journal of Mathematical and Statistical Psychology*, 45:93–107.

Lee, S. Y. and Song, X. Y. (2004). Bayesian model comparison of non-linear structural equation models with missing continuous and ordinal data. *British Journal of Mathematical and Statistical Psychology*, 57:131–150.

Lin, H., Mcculloch, C., Turnbull, B., Slate, E., and Clark, L. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, 19(10):1303–1318.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770.

Lynch, S. M. and Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods and Research*, 32:301–335.

Mantoni, M. and Pedersen, J. F. (1982). Fetal growth dealy in threatened abortion: an ultrasound study. *British Journal of Obstetrics and Gynecology*, 89:525–527.

Martin, J. and McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases. *Psychometrika*, 40:505–517.

MathWorks (2006). *Matlab*. Natick, Massachusetts.

Maulik, D. (2006). Fetal growth compromise: Definitions, standards, and classification. *Clinical Obstetrics and Gynecology*, 49(2):214 – 218.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, New York, 2nd edition.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240.

Miglioretti, D. (2003). Latent transition regression for mixed outcomes. *Biometrics*, 59(3):710–720.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3):391–411.

Muthen, B. (1983). Latent variable structural equation modeling with categorical-data. *Journal of Econometrics*, 22(1-2):43–65.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Muthen, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2):463–469.

Muthen, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29(1):81–117.

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8(3):229–242.

Nyberg, D. A., Mack, L. A., Laing, F. C., and Patten, R. M. (1987). Distinguishing normal from abnormal gestational sac growth in early pregnancy. *Journal of Ultrasounds in Medicine*, 6:23–27.

Oken, E., Kleinman, K. P., Rich-Edwards, J., and Gillman, M. W. (2003). A nearly continuous measure of birth weight for gestational age using a United States national reference. *BMC Pediatrics*, pages 1–10.

Ott, W. (1994). Accurate gestational dating - revisited. *American Journal of Perinatology*, 11(6):404 – 408.

Ott, W. J. (2006). Sonographic diagnosis of fetal growth restriction. *Clinical Obstetrics and Gynecology*, 49(2):295–307.

Palomo, J., Dunson, D. B., and Bollen, K. A. (2007). *Bayesian Structural Equation Modeling*. Elsevier, New York. In *Handbook of Latent Variable and Related Models*, S-Y Lee, editor.

Promislow, J. H. E., Makarushka, C. M., Gorman, J. R., Howards, P. P., Savitz, D. A., and Hartmann, K. E. (2004). Recruitment for a community-based study of early pregnancy: the Right From The Start study. *Pediatric and Perinatal Epidemiology*, 18(2):143–152.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B*, 59(4):731–758.

Robinson, P. M. (1974). Identification, estimation, and large sample theory for regressions containing unobservable variables. *International Economic Review*, 15:680–692.

Roy, J., Lin, X., and Ryan, L. (2003). Scaled marginal models for multiple continuous outcomes. *Biostatistics*, 4(3):371–383.

Ryan, L. M. and Orav, E. J. (1988). On the use of covariate for rodent bioassay and screening experiments. *Biometrika*, 75(4):631–7.

Sammel, M., Ryan, L., and Legler, J. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society Series B*, 59(3):667–678.

Sanchez, B., Budtz-Jorgensen, E., Ryan, L., and Hu, H. (2005). Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association*, 100(472):1443–1455.

Satten, G. A. and Sternberg, M. R. (1999). Fitting semi-Markov model to interval censored data with unknown initiation times. *Biometrics*, 55:507–513.

Savitz, D., Dole, N., Williams, J., Thorp, J., McDonald, T., Carter, A., and Eucker, B. (1999). Determinants of participation in an epidemiological study of preterm delivery. *Paediatric and Perinatal Epidemiology*, 13(1):114 – 125.

Savitz, D., Terry, J., Dole, N., Thorp, J., Siega-Riz, A., and Herring, A. (2002). Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. *American Journal of Obstetrics and Gynecology*, 187(6):1660–1666.

Shan, P. and Ohlsson, A. (2002). Literature review of low birth weight, including small for gestational age and preterm birth. Technical report, Toronto Public Health, www.toronto.ca/health/.

Slaughter, J. C., Herring, A. H., and Hartmann, K. E. (2007). Bayesian modeling of embryonic growth using latent variables. *Biostatistics*. Revise and resubmit.

Song, X. and Lee, S. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 57:29–52.

Stapelton, D. C. (1977). *Sociological Methodology 1978*, pages 52–74. Josey-Bass, San

Francisco. Analyzing political participation data with a MIMIC model.

Stephens, M. (1997). *Bayesian methods for mixtures of Normal distributions.* PhD thesis, Magdalen College, University of Oxford.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74.

Sternberg, M. R. and Satten, G. A. (1999). Discrete time nonparametric estimation for chain of events data subject to interval censoring and truncation. *Biometrics*, 55:507–513.

Tanner, M. and Wong, H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

van Dyk, D. and Meng, X. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

WHO Expert Committee (1995). Physical status: The use and interpretation of anthropometry. Technical Report 854, World Health Organization.

Wilcox, A., Weinberg, C. R., O'Connor, J., Baird, D., Schlatterer, J., Canfield, R., Armstrong, E., and Nisula, B. (1988). Incidence of early loss of pregnancy. *New England Journal of Medicine*, 319:189–194.

Wilcox, A. J., Dunson, D. B., Weinberg, C. R., Trussell, J., and Baird, D. D. (2001). Likelihood of conception with a single act of intercourse: Providing benchmark rates for assessment of post-coital contraceptives. *Contraception*, 63:211–215.

Wilcox, A. J., Weinberg, C. R., and Baird, D. D. (1995). Timing of sexual intercourse in relation to ovulation. *New England Journal of Medicine*, 333:1517–1521.

Zhang, J. and Bowes, W. A. (1995). Birth-weight-for-gestational-age patterns by race, sex and parity in the United States population. *Obstetrics and Gynecology*, 86(2):200–208.

Zhou, H. (2006). Statistical models for human fecundability. *Statistical Methods in Medical Research*, 15(2):181–194.