

Missing data methods in longitudinal studies: a review

Peer-reviewed author version

IBRAHIM, Joseph & MOLENBERGHS, Geert (2009) Missing data methods in longitudinal studies: a review. In: TEST, 18(1). p. 1-43.

DOI: 10.1007/s11749-009-0138-x

Handle: <http://hdl.handle.net/1942/10769>

# Missing Data Methods in Longitudinal Studies: A Review\*

Joseph G. Ibrahim · Geert Molenberghs

Received: date / Accepted: date

**Abstract** Incomplete data are quite common in biomedical and other types of research, especially in longitudinal studies. During the last three decades, a vast amount of work has been done in the area. This has led, on the one hand, to a rich taxonomy of missing-data concepts, issues, and methods, and, on the other hand, to a variety of data-analytic tools. Elements of taxonomy include: missing data patterns, mechanisms, and modeling frameworks; inferential paradigms; and sensitivity analysis frameworks. These are described in detail. A variety of concrete modeling devices is presented. To make matters concrete, two case studies are considered. The first one concerns quality of life among breast cancer patients, while the second one examines data from the Muscatine children's obesity study.

**Keywords** Expectation-maximization algorithm · Incomplete data · Missing completely at random · Missing at random · Missing not at random · Pattern-mixture model · Selection model · Shared-parameter model

## 1 Introduction

In a longitudinal study, each experimental or observational unit is measured at baseline and repeatedly over time. Incomplete data are not unusual under such designs, as many subjects are not available to be measured at all time points. In addition, a subject can be missing at one follow-up time and then measured again at one of the next, resulting in nonmonotone missing data patterns. Such data present a considerable modeling challenge for the statistician.

Rubin (1976) distinguished between three important mechanisms. When missingness is unrelated to the data, missingness is termed *missing completely at random* (MCAR). When

---

\* This invited paper is discussed in the comments available at:

Joseph G. Ibrahim  
Department of Biostatistics, University of North Carolina at Chapel Hill, USA  
E-mail: [ibrahim@bios.unc.edu](mailto:ibrahim@bios.unc.edu)

Geert Molenberghs  
Center for Statistics, Hasselt University, Belgium  
E-mail: [geert.molenberghs@uhasselt.be](mailto:geert.molenberghs@uhasselt.be)

missingness depends on the observed data and, when given the observed data, it does not depend on the unobserved data, the mechanism is *missing at random* (MAR). A mechanism where missingness depends on the unobserved data, perhaps in addition to the observed data, is termed *missing not at random* (MNAR). In the likelihood and Bayesian paradigm, and when mild regularity conditions are satisfied, the MCAR and MAR mechanisms are *ignorable*, in the sense that inferences can proceed by analyzing the observed data only, without explicitly addressing a (parametric) form of the missing data mechanism. In this situation, MNAR mechanisms are *non-ignorable*. Note that frequentist inference is generally ignorable only under MCAR.

In the ignorable situation, standard longitudinal data software allowing for unbalanced data can be used. Examples include the SAS procedures MIXED, GLIMMIX, and NLMIXED, and the SPlus and R functions `lme` and `n1me`, etc... Such tools eliminate complete-case bias by incorporating all available information. However, in the nonignorable case, methods that do not model the missing data mechanism are subject to bias.

Whereas ignorable likelihood analyses and appropriate frequentist techniques, such as weighted generalized estimating equations (Robins, Rotnitzky and Zhao, 1995), provide a versatile framework, as opposed to the collection of simple methods such as complete case analysis or last observation carried forward, nonignorable missing data occur very commonly in longitudinal studies. In many cancer and AIDS clinical trials, the side effects of the treatment may affect participation, and missingness can depend on the outcome as well as the treatment covariate. In quality of life studies, compliance is not compulsory, and those with a poor prognosis may be more likely not to complete the questionnaire at every visit. In environmental studies, geographic location or environmental factors may influence the response. Examples of nonignorable missingness can also be found in longitudinal psychiatric studies (Molenberghs et al, 1997; Little and Wang, 1996).

Estimating parameters with nonignorable missing data is complex. Likelihood-based methods require specification of the joint distribution of the data and the missing data mechanism. This specification can be further classified into three types of models: selection, pattern-mixture, and shared-parameter models (Little, 1995). The selection approach models the hypothetical complete data together with the missing data process conditional on the hypothetical complete data. The pattern-mixture approach models the distribution of the data conditional on the missing data pattern. Both of these approaches will be discussed in this paper. The third approach, shared-parameter models, accounts for the dependence between the measurement and missingness processes by means of latent variables such as random effects (Wu and Bailey, 1988, 1989; Wu and Carroll, 1988; Creemers et al, 2009).

There is an enormous literature on literature missing data methods in longitudinal studies. We refer the reader to the excellent books by Diggle, Heagerty, Liang, Zeger (2002), Fitzmaurice et al (2004), Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005), Molenberghs and Verbeke (2005), Molenberghs and Kenward (2007), Daniels and Hogan (2008), Fitzmaurice, Davidian, Verbeke, Molenberghs (2008), and the many references therein. Most of the literature focuses on maximum likelihood methods of estimation with nonignorable missing longitudinal data, predominantly focusing on mixed-effects models and normally distributed outcomes. A substantial part of the literature also assumes monotone patterns of missingness, where sequences of measurements on some subjects simply terminate prematurely. Approaches using selection models include Diggle and Kenward (1994), Little (1995) and Ibrahim, Chen, Lipsitz (2001). Approaches based on pattern-mixture models include Little (1995), Little and Wang (1996), Hogan and Laird (1997), and Thijs et al (2002). Troxel, Harrington and Lipsitz (1998) and Troxel, Lipsitz and Harrington (1998) propose a selection model which is valid for nonmonotone missing data, but is intractable

---

for more than three time points. There is less literature, however, on estimating parameters for the class of generalized linear mixed models (GLMM) with nonignorable missing data. Follman and Wu (1995) consider an extension of the conditional linear model to generalized linear models. Molenberghs, Kenward and Lesaffre (1997) propose a selection model for longitudinal ordinal data with nonrandom dropout. Ekholm and Skinner (1998) discuss a pattern-mixture model for a longitudinal binary, partially incomplete data set. Ibrahim, Chen, Lipsitz (2001) propose a method for estimating parameters in the GLMM using a selection model with nonignorable missing response data, while Fitzmaurice and Laird (2000) propose a method based on generalized estimating equations for estimating parameters in the GLMM using a mixture model with nonignorable dropouts.

While other methods of estimation with nonignorable nonresponse will be considered briefly, likelihood-based frequentist methods using selection and pattern-mixture models will be the primary focus of this paper. The literature is just too enormous to review all possible inference paradigms in this paper, such as multiple imputation, Bayesian methods, and weighted estimating equations, for example. For the class of generalized linear models, Ibrahim, Chen, Lipsitz, Herring (2005) present a detailed overview and comparisons of the four main paradigms for handling missing covariate data, these being i) maximum likelihood (ML), ii) multiple imputation (MI), iii) Bayesian methods, and iv) weighted estimating equations (WEE).

The remainder of this section motivates the setting with two real longitudinal data sets with likely nonignorable missing data. Section 2 discusses types of missing data in longitudinal studies. Section 3 focuses on estimation in the normal random effects model. Section 4 discusses present methods for estimation in the GLMM. Section 5 reviews methods for handling nonignorable missing data in the GLMM. Shared-parameter models are the topic of Section 6. We give a brief discussion of Bayesian methods in Section 7, and give some concluding remarks in Section 8.

## 1.1 Motivating Examples

As previously mentioned, many longitudinal studies call for estimation methods that can handle nonignorable missing data, since the possibility of such mechanism operation is impossible to rule out. This section presents two common examples to illustrate the problem in more detail.

### *Example 1: IBCSG Data*

Consider a data set concerning the quality of life among breast cancer patients in a clinical trial comparing four different chemotherapy regimens conducted by the International Breast Cancer Study Group (IBCSG Trial VI; Ibrahim, Chen, Lipsitz, 2001). The main outcomes of the trial were time until relapse and death, but patients were also asked to complete quality of life questionnaires at baseline and at three-month intervals. Some patients did refuse, on occasion, to complete the questionnaire. However, even when they refused, the patients were asked to complete an assessment at their next follow-up visit. Thus, the structure of this trial resulted in nonmonotone patterns of missing data. One longitudinal quality of life outcome was the patient's self-assessment of her mood, measured on a continuous scale from 0 (best) to 100 (worst). The three covariates of interest included a dichotomous covariate for language (Italian or Swedish), a continuous covariate for age, and three dichotomous covariates for the treatment regimen (4 regimens). Data from the first 18 months of the study were

used, implying that each questionnaire was filled out at most seven times, i.e., at baseline plus at six follow-up visits.

There are 397 observations in the data set, and mood is missing at least one time for 71% of the cases, resulting in 116 (29%) complete cases. The amount of missing data is minimal at baseline (2%), and ranges between 24% and 31% at the other six times: 26.2% at the second, 24.2% at the third, 29% at the fourth, 24.9% at the fifth, 28.2% at the sixth, and 30.5% at the seventh occasion. Table 1 provides a summary of the missing data patterns; the overall fraction of missing measurements is 23.6%. All patients were alive at the end of 18 months, so no missingness is due to death. However, it is reasonable to conjecture that the mood of the patient affected their decision to fill out the questionnaire. In this case, the missingness would be MNAR, and an analysis that does not include the missing data mechanism would be biased. In fact, Ibrahim, Chen, Lipsitz (2001) show a slight difference in the significance of one of the treatment covariates and the age covariate between their ignorable and nonignorable models.

**Table 1** IBCSG Trial VI Patterns of Missingness

Number of Missing Components of $y_j$	Frequency	Percentage
0	116	29.2
1	116	29.2
2	62	15.6
3	35	8.8
4	30	7.6
5	38	9.6

*Source:* Ibrahim, Chen, Lipsitz (2001)

#### *Example 2: Muscatine Children's Obesity Data*

The Muscatine Coronary Risk Factor Study (MCRFS) was a longitudinal study of coronary risk factors in school children (Woolson and Clarke, 1984; Ekholm and Skinner, 1998). Five cohorts of children were measured for height and weight in 1977, 1979, and 1981. Relative weight was calculated as the ratio of a child's observed weight to the median weight for their age-sex-height group. Children with a relative weight greater than 110% of the median weight for their respective stratum were classified as obese. The analysis of this study involves binary data (1 =obese, 0 =not obese) collected at successive time points. For every cohort, each of the following seven response patterns occurs:  $(p, p, p)$ ,  $(p, p, m)$ ,  $(p, m, p)$ ,  $(m, p, p)$ ,  $(p, m, m)$ ,  $(m, p, m)$ , and  $(m, m, p)$ , where a  $p$  ( $m$ ) denotes that the child was present (missing) for the corresponding measurement. The distribution over the patterns is shown in Table 2.

The statistical problem is to estimate the obesity rate as a function of age and sex. However, as can be seen in Table 2, many data records are incomplete since many children participated in only one or two occasions of the survey. Ekholm and Skinner (1998) report that the two main reasons for nonresponse were: (i) no consent form signed by the parents was received and (ii) the child was not in school on the day of the examination. If the parent did not sign the consent form because they did not want their child to be labeled as obese, or if the child did not attend school the day of the survey because of their weight, then the

**Table 2** MCRFS Patterns of Missingness

Response Pattern	Frequencies (%) for the following boys' cohorts:					Frequencies (%) for the following girls' cohorts:				
	B6	B8	B10	B12	B14	G6	G8	G10	G12	G14
$(p, p, p)$	24	43	43	38	30	23	48	45	39	28
$(p, p, m)$	4	10	13	13	24	5	8	12	13	26
$(p, m, p)$	2	5	4	4	2	2	5	3	9	1
$(m, p, p)$	34	11	9	5	4	37	12	13	5	4
$(p, m, m)$	8	10	14	22	25	6	11	14	21	25
$(m, p, m)$	9	7	8	6	12	7	6	7	5	12
$(m, m, p)$	19	13	9	11	4	19	11	6	7	4
Sum	100	99	100	99	101	99	101	100	99	100
Number of Children	493	522	533	476	472	442	492	492	461	483

Source: Ekholm and Skinner (1998)

missingness would at least be MAR, and likely even MNAR. In the latter case, an analysis that ignores the missing data mechanism would be biased. However, since the outcome is binary, these data cannot be modeled using the normal random effects model. Instead, a general method for estimating parameters for the class of GLMM's with nonignorable missing response data is needed.

## 2 Missing Data in Longitudinal Studies

We will now formalize the ideas loosely described in the introduction. Methods for handling missing data often depend on the pattern of missingness and the mechanism that generates the missing values. To illustrate the various missingness patterns and mechanisms in a regression setting, consider a data set that consists of a univariate vector of responses  $y_i = (y_{i1}, \dots, y_{in_i})'$  that may contain missing values, and an  $n_i \times p$  matrix  $X_i = (x_{i1}, \dots, x_{in_i})'$  of completely observed explanatory variables. We first define missing data patterns and then mechanisms.

### 2.1 Patterns of Missing Data

Data follow a **monotone missing pattern** if, once a subject misses a measurement occasion, s/he is never observed again. Monotone missing data are also termed dropout. For example, missing values in the vector of responses,  $y_i$  take the dropout form if, whenever  $y_{ij}$  is missing, so are  $y_{ik}$ , for all  $k \geq j$ . Likelihood functions are easier to evaluate with monotone patterns of missing data since they can be factored in terms of conditional densities.

Data follow a **nonmonotone missing pattern** if at least some subject values are observed again after a missing value occurs. For example, if  $y_i$  contains missing values, they are intermittent and  $y_{ij}$  may be missing while  $y_{ik}$  is observed, for some  $k > j$ . Likelihoods are more difficult to evaluate with nonmonotone patterns of missing data since almost always no simple factorization applies. In the MAR case, however, where ignorability applies, conventional software tools for longitudinal data models, allowing for unbalanced data, can be used to satisfaction.

## 2.2 Classifications of Missing Data Mechanisms

We present the mechanisms, in accordance with Rubin (1987) and Little and Rubin (2002).

Missing data are **missing completely at random** if the failure to observe a value does not depend on any values of  $y_i$ , either observed or missing, or any other observed values. For example, suppose that some components of  $y_i$  are missing while  $X_i$  is completely observed. The missing values of  $y_i$  are MCAR if the probability of observing  $y_i$  is independent of  $X_i$  and the values of  $y_i$  that are observed or would have been observed. Under MCAR, the observed data is just a random sample of all the data. A complete-case analysis may lose efficiency, but no bias is introduced. Under MCAR, the missing data mechanism takes the simple form  $f(r_i|X_i, \phi)$  (where  $\phi$  is a vector of unknown parameters), i.e., the outcomes do not intervene in the model for the missing-data indicators  $R_i = (R_{i1}, \dots, R_{in_i})'$ , where  $R_{ij} = 1$  if  $Y_{ij}$  is observed and 0 otherwise.

Missing data are **missing at random** if the failure to observe a value does not depend on the values of  $y_i$  which are unobserved, given the observed ones. However, the missingness may depend on other observed values. For example, suppose, as before, that  $X_i$  is completely observed while some components of  $y_i$  may be missing. The missing values of  $y_i$  are MAR if the probability of observing  $y_i$  is independent of the values of  $y_i$  that would have been observed, but is not necessarily independent of the observed values of  $y_i$  and  $X_i$ . This is a more realistic assumption than MCAR, but now adjustments must be made because observed responses are no longer a random sample. A complete-case analysis will be both inefficient and biased. Clearly, if data is MCAR, then it is MAR. For example, in a clinical trial, if missingness depends on the treatment allocation only, which has the status of a covariate, then the mechanism is MCAR and, *a fortiori*, also MAR. Under MAR, the missing data mechanism becomes  $f(r_i|X_i, y_{obs,i}, \phi)$ , where  $y_{obs,i}$  denotes the observed components of  $y_i$ .

The missing data mechanism is said to be **missing not at random** if the failure to observe a value depends on the value that would have been observed. For example, suppose some components of  $y_i$  are missing, but that  $X_i$  is completely observed. The missing values of  $y_i$  are MNAR if the probability that  $y_i$  is missing depends on the missing values of  $y_i$ , regardless of whether it depends on the observed values of  $y_i$  or  $X_i$ . MNAR is the most general situation and is frequently encountered in longitudinal studies with repeated measures. Valid inferences generally require either specifying the correct model for the missing data mechanism, or distributional assumptions for  $y_i$ , or both. The resulting estimators and tests are typically sensitive to these assumptions. Therefore, the mechanism should play a central role within so-called sensitivity analyses (Section 5.1). Under MNAR, the missing data mechanism is fully general:  $f(r_i|X_i, y_{obs,i}, y_{mis,i}, \phi)$ .

Within the likelihood or Bayesian inferential frameworks, and when the parameters governing the measurement and missingness process are functionally independent, then MCAR and MAR mechanisms are ignorable. However, the frequentist framework generally requires the mechanism to be MCAR for ignorability to apply (Rubin, 1976).

## 3 The Normal Random-effects Model

The normal random-effects model, also known as the *Laird-Ware model* (Laird and Ware, 1982), is a special case of the generalized linear mixed model, which is the subject of the next section. The model is intended for continuous, normally distributed outcomes. Precisely, for a given individual  $i$  with  $n_i$  repeated measurements, the Laird-Ware model for outcome

vector  $y_i$  can be written as

$$y_i = X_i\beta + Z_i b_i + e_i, \quad i = 1, \dots, N, \quad (1)$$

where  $y_i$  is  $n_i \times 1$ ,  $X_i$  is an  $n_i \times p$  known matrix of fixed-effects covariates,  $\beta$  is a  $p \times 1$  vector of unknown regression parameters, commonly referred to as fixed effects,  $Z_i$  is a known  $n_i \times q$  matrix of covariates for the  $q \times 1$  vector of random effects  $b_i$ , and  $e_i$  is an  $n_i \times 1$  vector of errors. The columns of  $Z_i$  are usually a subset of  $X_i$ , allowing for fixed effects as well as random intercepts and/or slopes. It is typically assumed that the  $e_i$ 's are independent, the  $b_i$ 's are *i.i.d.*, the  $b_i$ 's are independent of the  $e_i$ 's, and

$$e_i \sim N_{n_i}(0, \sigma^2 I_{n_i}), \quad b_i \sim N_q(0, D),$$

where  $I_{n_i}$  is the  $n_i \times n_i$  identity matrix and  $N_q(\mu, \Sigma)$  denotes the  $q$ -dimensional multivariate normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$ . The positive definite matrix  $D$  is the covariance matrix of the random effects and is typically assumed to be unstructured and unknown. Under these assumptions, the so-called conditional model, where conditioning refers to the random effects, takes the form

$$y_i | \beta, \sigma^2, b_i \sim N_{n_i}(X_i\beta + Z_i b_i, \sigma^2 I_{n_i}). \quad (2)$$

The model in (2) assumes a distinct set of regression coefficients for each individual once the random effects are known. Upon integration over the random effects, the so-called marginal distribution of  $y_i$  is

$$y_i | \beta, \sigma^2, D \sim N_{n_i}(X_i\beta, Z_i D Z_i' + \sigma^2 I_{n_i}). \quad (3)$$

### 3.1 Complete-data Estimation

Maximum likelihood (ML) estimation has been extensively considered for the normal random effects model (see, for example, Laird and Ware, 1982; Jennrich and Schluchter, 1986). The standard approach is to take the first and second derivatives of the log-likelihood based on the marginal distribution of  $y_i$  and use Newton-Raphson (based on the observed information) or Fisher scoring (based on the expected information) methods as the basis for iteratively obtaining the maximum likelihood estimates. Often, a hybrid approach to this iterative method is taken, where the updated value of  $\hat{\beta}$  is used to calculate  $\hat{\theta} = (\hat{\sigma}^2, \hat{D})$ .

The method described here uses the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) for computing ML estimates. The EM algorithm is advantageous over the Newton-Raphson or Fisher scoring algorithms when formulating models with large numbers of covariance parameters. The procedure consists of two steps. The first step uses weighted least squares ideas to update  $\hat{\beta}$ , which is equivalent to maximizing the likelihood with respect to  $\beta$  while holding the covariance parameters  $\theta = (\sigma^2, D)$  fixed. In the second step,  $\hat{\theta}$  is updated using  $Y = (y_1, \dots, y_N)$  as the observed data and  $V = (y_1, b_1, \dots, y_N, b_N)$  as the complete data.

Starting out with the first step, the log-likelihood based on the observed data,  $Y$ , is

$$\begin{aligned} \ell(\beta, \sigma^2, D) &= \log \left[ \prod_{i=1}^N f(y_i | \beta, \sigma^2, D) \right] \\ &= \sum_{i=1}^N -\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^N (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta), \end{aligned}$$



where  $\Sigma_i = Z_i D Z_i' + \sigma^2 I_{n_i}$ . The score equation for  $\beta$  is given by

$$\frac{d\ell}{d\beta} = \sum_{i=1}^N X_i' \Sigma_i^{-1} y_i - \sum_{i=1}^N X_i' \Sigma_i^{-1} X_i \beta .$$

Setting this first derivative equal to zero and solving for  $\beta$  produces the ML estimate,

$$\hat{\beta} = \left( \sum_{i=1}^N X_i' \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^N X_i' \Sigma_i^{-1} y_i .$$

The second step uses the complete data log-likelihood given by

$$\begin{aligned} \ell(\beta, \sigma^2, D) &= \log \left[ \prod_{i=1}^N f(y_i, b_i | \beta, \sigma^2, D) \right] = \sum_{i=1}^N \log [f(y_i | \beta, \sigma^2, b_i)] + \sum_{i=1}^N \log [f(b_i | D)] \\ &= \sum_{i=1}^N \left( -\frac{n_i}{2} \log(2\pi) - \frac{n_i}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y_i - X_i \beta - Z_i b_i)' (y_i - X_i \beta - Z_i b_i) \right) \\ &\quad + \sum_{i=1}^N \left( -\frac{q}{2} \log(2\pi) - \frac{1}{2} \log |D| - \frac{1}{2} b_i' D^{-1} b_i \right) . \end{aligned}$$

This expression establishes that  $\sum_{i=1}^N (y_i - X_i \beta - Z_i b_i)' (y_i - X_i \beta - Z_i b_i) \equiv \sum_{i=1}^N e_i' e_i$  and  $\sum_{i=1}^N b_i b_i'$  are the complete data sufficient statistics for  $\sigma^2$  and  $D$ , respectively. The M-step is then given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N e_i' e_i}{\sum_{i=1}^N n_i} , \quad \hat{D} = \sum_{i=1}^N \frac{b_i b_i'}{N}$$

and thus

$$\hat{\Sigma}_i = Z_i \hat{D} Z_i' + \hat{\sigma}^2 I_{n_i} .$$

The E-step consists of calculating the expected value of the sufficient statistics given the observed data and the current parameter estimates:

$$\begin{aligned} E(b_i b_i' | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) &= E(b_i | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) E(b_i' | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) + \text{Var}(b_i | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) \\ &= \hat{D} Z_i' \hat{\Sigma}_i^{-1} (y_i - X_i \hat{\beta}) (y_i - X_i \hat{\beta})' \hat{\Sigma}_i^{-1} Z_i \hat{D} + \hat{D} - \hat{D} Z_i' \hat{\Sigma}_i^{-1} Z_i \hat{D} , \end{aligned}$$

$$\begin{aligned} E(e_i' e_i | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) &= \text{tr}(E(e_i e_i' | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D})) \\ &= \text{tr}(E(e_i | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) E(e_i' | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D}) + \text{Var}(e_i | y_i, \hat{\beta}, \hat{\sigma}^2, \hat{D})) \\ &= \text{tr}(\hat{\sigma}^4 \hat{\Sigma}_i^{-1} (y_i - X_i \hat{\beta}) (y_i - X_i \hat{\beta})' \hat{\Sigma}_i^{-1} + \hat{\sigma}^2 I_{n_i} - \hat{\sigma}^4 \hat{\Sigma}_i^{-1}) , \end{aligned}$$

where  $e_i = y_i - X_i \beta - Z_i b_i$ . One iterates between both steps until convergence.

Note that the EM algorithm converges linearly, in contrast to super-linear convergence of Fisher scoring and even quadratic convergence of Newton-Raphson. However, key advantages of the EM algorithm are that (1) implementation is frequently more straightforward and intuitive and (2) there is a much lower risk for divergence. Sometimes, hybrid algorithms can be used, setting out with EM and then switching to Fisher-scoring or Newton-Raphson. Alternatively, EM-acceleration methods can be used (Louis, 1982; Meilijson, 1989). Such methods are also useful when determining measures of precision.

### 3.2 Estimation With Nonignorable Missingness

When the missing data mechanism is MNAR, one needs to specify a (parametric) model for missingness alongside the aforementioned model for the outcomes and incorporate it into the complete data log-likelihood. The missing data mechanism is defined as the distribution of the  $n_i \times 1$  random vector  $R_i$ , whose  $j$ th component,  $r_{ij} = 1$  if  $y_{ij}$  is missing and 0 otherwise. The distribution of  $R_i$  is indexed by the parameter vector  $\phi$  and takes a multinomial form with  $2^{n_i}$  cell probabilities. Under the normal mixed model, the complete data density of  $(y_i, b_i, r_i)$  for subject  $i$  is then given by  $f(y_i, b_i, r_i | \beta, \sigma^2, D, \phi)$ . Little (1993, 1995) identified two ways of factoring this joint distribution. *Selection models* decompose the joint distribution as (with covariates suppressed from notation)

$$f(y_i, b_i, r_i | \beta, \sigma^2, D, \phi) = f(y_i | \beta, \sigma^2, b_i) f(b_i | D) f(r_i | \phi, y_i),$$

whereas *pattern-mixture models* employ the reverse factorization

$$f(y_i, b_i, r_i | \beta, \sigma^2, D, \phi) = f(y_i | \beta, \sigma^2, b_i, r_i) f(b_i | D) f(r_i | \phi).$$

The term ‘‘pattern-mixture’’ emphasizes that the marginal distribution of  $y = (y'_1, \dots, y'_N)'$  is a mixture of pattern-specific distributions. Most estimation methods assume that the distribution of  $r_i$  depends on  $(y_i, X_i, Z_i)$ , but not on  $b_i$ . This assumption will be addressed in the discussion of models for the missing data mechanism.

#### 3.2.1 Selection Models

*Estimation* The complete data log-likelihood for the selection model is

$$\ell(\gamma) = \log \left[ \prod_{i=1}^N f(y_i, b_i, r_i | \beta, \sigma^2, D, \phi) \right] \quad (4)$$

$$\begin{aligned} &= \sum_{i=1}^N l(\gamma; y_i, b_i, r_i) \\ &= \sum_{i=1}^N \log [f(y_i | \beta, \sigma^2, b_i)] + \sum_{i=1}^N \log [f(b_i | D)] + \sum_{i=1}^N \log [f(r_i | \phi, y_i)], \end{aligned} \quad (5)$$

where  $\gamma = (\beta, \sigma^2, D, \phi)$ . Estimation of  $(\beta, \sigma^2, D)$  is usually of interest, with often, but not always, both the random effects as well as  $\phi$  being viewed as nuisance parameters. Diggle and Kenward (1994) discuss estimation methods for selection models assuming monotone missing data. However, these methods are not easily extended to the analysis of nonmonotone missing data, where a subject may be observed after a missing value occurs. The method described next, based on the so-called EM by Method of Weights (Ibrahim, 1990), is general in that it applies to both monotone and nonmonotone missing data.

For ease of exposition, write  $y_i = (y_{mis,i}, y_{obs,i})$ , where  $y_{mis,i}$  is the  $s_i \times 1$  vector of missing components of  $y_i$ . The Monte Carlo EM (MCEM) algorithm has been used for parametric estimation in selection models with nonignorable missing response data (Ibrahim, Chen, Lipsitz, 2001). The E-step consists of calculating the expected value of the complete data log-likelihood given the observed data and current parameter estimates. Since both  $b_i$  and  $y_{mis,i}$  are unobserved, they must be integrated over. Thus, the E-step for the  $i$ th observation at the  $(t+1)$ st iteration is

$$Q_i(\gamma | \gamma^{(t)}) = E(l(\gamma; y_i, b_i, r_i) | y_{obs,i}, r_i, \gamma^{(t)})$$

$$\begin{aligned}
&= \int \int \log [f(y_i|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, b_i)] f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) db_i dy_{mis,i} \\
&\quad + \int \int \log [f(b_i|D)] f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) db_i dy_{mis,i} \\
&\quad + \int \int \log [f(r_i|\boldsymbol{\phi}, y_i)] f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) db_i dy_{mis,i} \\
&\equiv I_1 + I_2 + I_3, \tag{6}
\end{aligned}$$

where  $\boldsymbol{\gamma}^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}, D^{(t)}, \boldsymbol{\phi}^{(t)})$  and  $f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)})$  represents the conditional distribution of the data considered ‘‘missing,’’  $(y_{mis,i}, b_i)$ , given the observed data.

To integrate out  $b_i$  from  $I_1$  and  $I_2$ , write

$$f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) = f(b_i|y_i, \boldsymbol{\gamma}^{(t)})f(y_{mis,i}|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}),$$

and note that standard conditional distribution calculations yield

$$(b_i|y_i, \boldsymbol{\gamma}^{(t)}) \sim N_q(b_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}),$$

where

$$\begin{aligned}
b_i^{(t)} &= D^{(t)}Z_i'(Z_iD^{(t)}Z_i' + \boldsymbol{\sigma}^{2(t)}I_{n_i})^{-1}(y_i - X_i\boldsymbol{\beta}^{(t)}) \\
&= \boldsymbol{\Sigma}_i^{(t)}Z_i'(y_i - X_i\boldsymbol{\beta}^{(t)})/\boldsymbol{\sigma}^{2(t)},
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}_i^{(t)} &= D^{(t)} - D^{(t)}Z_i'(Z_iD^{(t)}Z_i' + \boldsymbol{\sigma}^{2(t)}I_{n_i})^{-1}Z_iD^{(t)} \\
&= \left[ \boldsymbol{\sigma}^{-2(t)}Z_i'Z_i + (D^{(t)})^{-1} \right]^{-1}.
\end{aligned}$$

Now,  $I_1$  can be written as

$$\begin{aligned}
I_1 &= \int \int \log [f(y_i|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, b_i)] f(y_{mis,i}, b_i|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) db_i dy_{mis,i} \\
&= -\frac{n_i}{2} \log (2\pi) - \frac{n_i}{2} \log (\boldsymbol{\sigma}^2) \\
&\quad - \int \frac{1}{2\boldsymbol{\sigma}^2} \left( \int (y_i - X_i\boldsymbol{\beta} - Z_ib_i)'(y_i - X_i\boldsymbol{\beta} - Z_ib_i) f(b_i|y_i, \boldsymbol{\gamma}^{(t)}) db_i \right) \\
&\quad \times f(y_{mis,i}|y_{obs,i}, r_i, \boldsymbol{\gamma}^{(t)}) dy_{mis,i}. \tag{7}
\end{aligned}$$

To evaluate the integral with respect to  $b_i$  in (7), note that

$$\begin{aligned}
&(y_i - X_i\boldsymbol{\beta} - Z_ib_i)'(y_i - X_i\boldsymbol{\beta} - Z_ib_i) \\
&= (y_i - X_i\boldsymbol{\beta} - Z_ib_i^{(t)})'(y_i - X_i\boldsymbol{\beta} - Z_ib_i^{(t)}) - 2(y_i - X_i\boldsymbol{\beta} - Z_ib_i^{(t)})'Z_i(b_i - b_i^{(t)}) \\
&\quad + (b_i - b_i^{(t)})'(Z_i'Z_i)(b_i - b_i^{(t)}). \tag{8}
\end{aligned}$$

Substituting (8) into (7), we have

$$I_1 = -\frac{n_i}{2} \log (2\pi) - \frac{n_i}{2} \log (\boldsymbol{\sigma}^2) - \frac{1}{2\boldsymbol{\sigma}^2} \left( \text{tr}(Z_i'Z_i\boldsymbol{\Sigma}_i^{(t)}) \right)$$

$$+ \int (y_i - X_i\beta - Z_i b_i^{(t)})' (y_i - X_i\beta - Z_i b_i^{(t)}) \times f(y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}) dy_{mis,i} \Big). \quad (9)$$

Following similar logic and upon noting that  $b_i \sim N_q(0, D)$ ,  $I_2$  can be written as

$$\begin{aligned} I_2 &= -\frac{q}{2} \log(2\pi) - \frac{1}{2} \log(|D|) - \frac{1}{2} \text{tr}(D^{-1} \Sigma_i^{(t)}) \\ &\quad - \frac{1}{2} \int (b_i^{(t)'} D^{-1} b_i^{(t)}) f(y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}) dy_{mis,i}. \end{aligned} \quad (10)$$

Finally, for  $I_3$ ,  $b_i$  can be easily integrated out since  $\log[f(r_i | \phi, y_i)]$  does not depend on  $b_i$ . Therefore,  $I_3$  can be written simply as

$$I_3 = \int \log[f(r_i | \phi, y_i)] f(y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}) dy_{mis,i}. \quad (11)$$

The E-step, expressed via (9), (10), and (11) does not involve  $b_i$ . Thus, to complete the E-step, we merely need to sample from  $[y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}]$ . This distribution can be written, up to a constant of proportionality, as

$$\begin{aligned} &f(y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}) \\ &\propto \exp\left(-\frac{1}{2} (y_i - X_i\beta^{(t)})' (Z_i D^{(t)} Z_i' + \sigma^{2(t)} I_{n_i})^{-1} (y_i - X_i\beta^{(t)})\right) \\ &\quad \times f(r_i | y_{mis,i}, y_{obs,i}, \gamma^{(t)}), \end{aligned} \quad (12)$$

which has the form of a normal density times a logistic regression for the  $r_i$ 's. Thus, the distribution is from the class of concave log-densities, and Gibbs sampling from (12) is straightforward, using the adaptive rejection algorithm of Gilks and Wild (1992).

Precisely, the procedure is as follows. Let  $u_{i1}, \dots, u_{im_i}$  be a sample of size  $m_i$  from  $[y_{mis,i} | y_{obs,i}, r_i, \gamma^{(t)}]$ , obtained via the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992). Also, let  $y_i^{(k)} = (u'_{ik}, y'_{obs,i})'$  and

$$b_i^{(tk)} = \Sigma_i^{(t)} Z_i' (y_i^{(k)} - X_i\beta^{(t)}) / \sigma^{2(t)}.$$

Then, the E-step for the  $i$ th observation at the  $(t+1)$ th iteration takes the form

$$\begin{aligned} &Q_i(\gamma | \gamma^{(t)}) \\ &= -\frac{n_i}{2} \log(2\pi) - \frac{n_i}{2} \log(\sigma^2) \\ &\quad - \frac{1}{2\sigma^2} \left( \text{tr}(Z_i' Z_i \Sigma_i^{(t)}) + \frac{1}{m_i} \sum_{k=1}^{m_i} (y_i^{(k)} - X_i\beta - Z_i b_i^{(tk)})' (y_i^{(k)} - X_i\beta - Z_i b_i^{(tk)}) \right) \\ &\quad - \frac{q}{2} \log(2\pi) - \frac{1}{2} \log(|D|) - \frac{1}{2} \text{tr}(D^{-1} \Sigma_i^{(t)}) - \frac{1}{2m_i} \sum_{k=1}^{m_i} b_i^{(tk)'} D^{-1} b_i^{(tk)} \\ &\quad + \frac{1}{m_i} \sum_{k=1}^{m_i} \log[f(r_i | \phi, y_i^{(k)})]. \end{aligned}$$

Obviously, the E-step for all  $N$  observations is given by

$$Q(\gamma | \gamma^{(t)}) = \sum_{i=1}^N Q_i(\gamma | \gamma^{(t)}).$$

Stubbendick and Ibrahim (2003) extend this approach to the problem of nonignorable missing covariates and/or responses in the normal mixed model. Similar MCEM algorithms have been developed for other types of models, such as generalized linear models and survival models by Ibrahim, Lipsitz, Chen (1999a), Ibrahim, Chen, Lipsitz (1999b), Ibrahim, Lipsitz, Chen (1999a), Chen and Ibrahim (2001), Herring and Ibrahim (2001), Herring and Ibrahim (2002), Herring et al (2002), Herring, Ibrahim, Lipsitz (2004), Chen and Ibrahim (2001), Chen and Ibrahim (2006), and Chen, Zeng, Ibrahim (2007).

Let us turn to the M-step, which maximizes  $Q(\gamma|\gamma^{(t)})$ , and closed forms are available for  $(\beta, \sigma^2, D)$ . The procedure for the M-step is as follows:

(i) Find  $\phi^{(t+1)}$  to maximize

$$Q_\phi = \sum_{i=1}^N \frac{1}{m_i} \sum_{k=1}^{m_i} \log [f(r_i|\phi, y_i^{(k)})]. \quad (13)$$

(ii) Find  $D^{(t+1)}$  to maximize

$$Q_D = -\frac{N}{2} \log(|D|) - \frac{1}{2} \text{tr}(D^{-1} \sum_{i=1}^N \Sigma_i^{(t)}) - \frac{1}{2} \sum_{i=1}^N \frac{1}{m_i} \sum_{k=1}^{m_i} b_i^{(tk)'} D^{-1} b_i^{(tk)} \quad (14)$$

which yields

$$D^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \left[ \frac{1}{m_i} \sum_{k=1}^{m_i} b_i^{(tk)} b_i^{(tk)'} + \Sigma_i^{(t)} \right].$$

(iii) Find  $\beta^{(t+1)}$  to minimize

$$Q_\beta = \sum_{i=1}^N \frac{1}{m_i} \sum_{k=1}^{m_i} (y_i^{(k)} - X_i \beta - Z_i b_i^{(tk)})' (y_i^{(k)} - X_i \beta - Z_i b_i^{(tk)}), \quad (15)$$

which yields

$$\beta^{(t+1)} = \left( \sum_{i=1}^N X_i' X_i \right)^{-1} \sum_{i=1}^N \left( X_i' \frac{1}{m_i} \sum_{k=1}^{m_i} (y_i^{(k)} - Z_i b_i^{(tk)}) \right).$$

(iv) Find  $\sigma^{2(t+1)}$  to minimize

$$Q_{\sigma^2} = \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N \text{tr}(Z_i' Z_i \Sigma_i^{(t)}) + \frac{1}{2\sigma^2} \left( \sum_{i=1}^N \frac{1}{m_i} \sum_{k=1}^{m_i} (y_i^{(k)} - X_i \beta^{(t+1)} - Z_i b_i^{(tk)})' (y_i^{(k)} - X_i \beta^{(t+1)} - Z_i b_i^{(tk)}) \right), \quad (16)$$

which leads to

$$\sigma^{2(t+1)} = \frac{1}{n} \sum_{i=1}^N \left( \frac{1}{m_i} \sum_{k=1}^{m_i} (y_i^{(k)} - X_i \beta^{(t+1)} - Z_i b_i^{(tk)})' (y_i^{(k)} - X_i \beta^{(t+1)} - Z_i b_i^{(tk)}) + \text{tr}(Z_i' Z_i \Sigma_i^{(t)}) \right),$$

where  $n = \sum_{i=1}^N m_i$ .

*Models for the Missing Data Mechanism* Diggle and Kenward (1994) proposed a binomial model for the missing data mechanism under the selection modeling approach, i.e.,

$$f(r|\phi, y) = \prod_{i=1}^N \prod_{j=1}^{n_i} [P(r_{ij} = 1|\phi)]^{r_{ij}} [(1 - P(r_{ij} = 1|\phi))]^{1-r_{ij}},$$

where  $P(r_{ij} = 1|\phi)$  is modeled via a logistic regression involving all of the previous outcomes as well as the current outcome. This model takes the form

$$\text{logit} \left( P(r_{ij} = 1|\phi) \right) \equiv \log \left[ \frac{P(r_{ij} = 1|\phi)}{1 - P(r_{ij} = 1|\phi)} \right] = \phi_0 + \phi_1 y_{ij} + \sum_{k=2}^j \phi_k y_{j+1-k},$$

for  $i = 1, \dots, N$ , and  $j = 1, \dots, n_i$ . The model can be extended to permit possible relationships between the missing data process and covariates, including time, by making  $\phi_0$  a function of the covariates  $x_{qj}$  at time  $t_j$ . A linear function in the covariates could be written as

$$\phi_0 = \sum_{q=1}^r \phi_{q0} x_{qj}. \quad (17)$$

For example, for the IBCSG data, consider a logistic regression model that includes the previous and current outcome as well as treatment covariates. Such a choice would specialize (17) to

$$\text{logit} \left( P(r_{ij} = 1|\phi) \right) = \phi_0 + \phi_1 y_{i,j-1} + \phi_2 y_{ij} + \phi_3 \text{trt}_{Ai} + \phi_4 \text{trt}_{Bi} + \phi_5 \text{trt}_{Ci},$$

for  $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$ , and  $\text{trt}_{Ti}$  an indicator variable for whether subject  $i$  receives treatment  $T = A, B, C$ . Note that these models assume independence between the  $r_{ij}$ 's, in line with their conditional interpretation as probabilities of dropout *given one is still at risk for dropping out*.

A more general multinomial missing data model which incorporates a general correlation structure can be constructed by specifying the joint distribution of  $r_i = (r_{i1}, \dots, r_{in_i})$  through a sequence of one-dimensional conditional distributions (Ibrahim, Chen, Lipsitz, 2001). Consider

$$\begin{aligned} f(r_{i1}, \dots, r_{in_i} | \phi, y_i) \\ &= f(r_{in_i} | \phi_{n_i}, r_{i1}, \dots, r_{i(n_i-1)}, y_i) \cdot f(r_{i(n_i-1)} | \phi_{n_i-1}, r_{i1}, \dots, r_{i(n_i-2)}, y_i) \cdots \\ &\quad \dots f(r_{i2} | \phi_{n_2}, r_{i1}, y_i) \cdot f(r_{i1} | \phi_{n_1}, y_i), \end{aligned} \quad (18)$$

where  $\phi_a$  is a vector of indexing parameters for the  $a$ th conditional distribution and  $\phi = (\phi'_{n_1}, \dots, \phi'_{n_N})'$ . Thus,

$$f(r|\phi, y) = \prod_{i=1}^N f(r_{i1}, \dots, r_{in_i} | \phi, y_i),$$

where  $f(r_{i1}, \dots, r_{in_i} | \phi, y_i)$  is given in (18). Since  $r_{ij}$  is binary, a sequence of logistic regressions can be used for (18). This modeling strategy has the potential of reducing the number of nuisance parameters that have to be specified for the missing data mechanism, yields general correlation structures between the  $r_{ij}$ 's, and allows more flexibility in the specification of the missing data model. It also accommodates nonmonotone patterns of missing data and provides a natural way to specify the joint distribution of the missing data indicators when

knowledge about the missingness of one response affects the probability of missingness of another. One must be careful not to build too large a model for the missing data mechanism, since the model can easily become unidentifiable. Thus, caution should be taken when adding interaction terms or other higher-order terms. It is not clear how to characterize the set of all estimable parameters for this class of models given a certain choice of variables in the missing data mechanism. The parametric form of the assumed missing data mechanism is not testable from the data. Therefore, although a model may pass the tests for a certain missing data mechanism, this does not mean that one has captured the correct, and perhaps more complicated, missing data mechanism.

Also, it has been assumed throughout that  $[r_i|\phi, y_i]$  does not depend on  $b_i$ . This is a reasonable assumption in practice since autoregressive models for  $[r_i|\phi, y_i]$  can closely approximate models for the missing data mechanism that include the random effect  $b_i$ . In other words, conditional on the outcome vector  $y_i$ , which contains information on the trajectory of the outcome over time,  $r_i$  is independent of  $b_i$ . In addition, the inclusion of a random effect in the missing data model induces a correlation structure across subjects in the marginal model  $[r_i|\phi, y_i]$ . Note, however, that the correlation structure induced via a sequence of conditional distributions for  $[r_i|\phi, y_i]$  as in (18) would also provide a suitable approximation to a correlation structure induced from a random effects model for the missing data mechanism. Little (1995) suggests using a model where missingness depends on the values of the random coefficients when the probability of missingness depends on current and past values of some latent variable that the outcome variable is measuring with error. However, including a random effect in  $[r_i|\phi, y_i]$  makes the E-step substantially more computationally intensive and all closed forms would be lost. A plausible alternative to the assumption, as suggested by Little (1995), is to model the missing data mechanism using the expected values of  $y_i$ , rather than the actual values. In this case, (18) would then be written as

$$\begin{aligned} f(r_{i1}, \dots, r_{im_i}|\phi, \beta, \sigma^2, b_i) &= f(r_{im_i}|\phi_{n_i}, r_{i1}, \dots, r_{i(n_i-1)}, E(y_i|\beta, \sigma^2, b_i)) \\ &\quad \times f(r_{i(n_i-1)}|\phi_{n_i-1}, r_{i1}, \dots, r_{i(n_i-2)}, E(y_i|\beta, \sigma^2, b_i)) \dots \\ &\quad \dots f(r_{i2}|\phi_{n_2}, r_{i1}, E(y_i|\beta, \sigma^2, b_i)) f(r_{i1}|\phi_{n_1}, E(y_i|\beta, \sigma^2, b_i)) . \end{aligned}$$

Other innovations for the normal mixed model include Lipsitz et al (2000) who consider Box-Cox transformations on the response variable in the presence of missing data and Lipsitz et al (2002) who consider missing data mechanisms based on outcome dependent follow-up.

### 3.2.2 Pattern-mixture Models

Pattern-mixture models are based on an alternative factorization of  $f(y_i, b_i, r_i|\beta, \sigma^2, D, \phi)$ . The complete data log-likelihood for the pattern-mixture model is

$$\begin{aligned} \ell(\gamma) &= \log \left[ \prod_{i=1}^N f(y_i, b_i, r_i|\beta, \sigma^2, D, \phi) \right] \\ &= \sum_{i=1}^N l(\gamma; y_i, b_i, r_i) \\ &= \sum_{i=1}^N \log [f(y_i|\beta, \sigma^2, b_i, r_i)] + \sum_{i=1}^N \log [f(b_i|D)] + \sum_{i=1}^N \log [f(r_i|\phi)] , \end{aligned}$$

where  $\gamma = (\beta, \sigma^2, D, \phi)$ . Since the distribution of  $y_i$  depends on  $r_i$ , a model based on this factorization implies that the marginal distribution of  $y_i$  is a mixture of normal distributions rather than a single normal distribution as in the selection model. By conditioning on  $r_i$ , this approach essentially stratifies the sample by the observed pattern of missing data and then models different distributions of  $y_i$  over these patterns. Stratifying on the pattern is not always the most obvious way to go forward, as substantive interest usually concerns the mean and covariance matrix of  $y_i$  averaged over pattern. However, it will be shown that inference for such parameters is not precluded in pattern-mixture models.

Recall the Laird-Ware model for the outcome vector  $y_i$  in (1). If  $y_i$  contains nonignorable missing data, model (2) becomes

$$(y_i | \beta, \sigma^2, b_i, r_i = k) \sim N_{n_i}(X_i \beta^{(k)} + Z_i b_i, \sigma^{2(k)} I_{n_i})$$

under the pattern-mixture factorization. The specification of distinct fixed parameters,  $(\beta^{(k)}, \sigma^{2(k)})$ , for each pattern creates major identification problems because not all parameters of the complete data distribution of  $y_i$  are estimable from incomplete pattern data. However, assumptions about the missing data mechanism can yield additional restrictions that help to identify the models, while avoiding explicit specification of the mechanism's parametric form, such as required in the selection model approach. See also Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005) for reviews.

To illustrate this idea, consider the analysis presented in Little and Wang (1996), in which pattern-mixture models are developed for a multivariate multiple regression. Suppose we have a sample of  $N$  independent observations on  $p$  continuous outcome variables and  $q$  covariates, so that  $y_i = (y_1, \dots, y_p)'$  and  $x_i = (x_1, \dots, x_q)'$ . Assume  $x_i$  and a subset of  $p_1$  rows of  $y_i$ , denoted by  $y_{(1)i} = (y_1, \dots, y_{p_1})'$ , are observed for all  $N$  cases and that the remaining  $p_2 = (p - p_1)$  rows of  $y_i$ , denoted  $y_{(2)i} = (y_{p_1+1}, \dots, y_p)'$ , are observed for  $N_0$  cases and are missing for  $N_1 = N - N_0$  cases. The indicator variable  $r$  is defined for observation  $i$  as  $r_i = 0$  if  $y_{(2)i}$  is observed and  $r_i = 1$  if  $y_{(2)i}$  is missing. Thus, we have a monotone missing data structure which can be found in longitudinal studies where subjects are lost to follow-up at the same time point. Now, pattern-mixture models are developed for this type of data using the model

$$(y_i | \beta, \Sigma, r_i = k) \sim N_p(\beta^{(k)} x_i, \Sigma^{(k)}), k = 0, 1; \quad (19)$$

$$r_i | \phi \sim \text{Bernoulli} \left( \frac{e^{\phi' x_i}}{1 + e^{\phi' x_i}} \right) \Rightarrow \text{logit} [P(r_i = 1 | \phi)] = \phi' x_i,$$

where  $y_i$  is  $p \times 1$ ,  $x_i$  is a  $q \times 1$  vector of known covariates,  $\beta^{(k)}$  is a  $p \times q$  coefficient matrix of unknown regression parameters for pattern  $k$ ,  $\Sigma^{(k)}$  is a  $p \times p$  unknown covariance matrix for pattern  $k$ ,  $r_i$  is an indicator variable for missingness, and  $\phi$  is a  $q \times 1$  vector of unknown logistic regression parameters. Therefore, the total number of parameters to be estimated is  $2pq + p(p+1) + q$ . If we let  $\theta^{(k)} = (\beta^{(k)}, \Sigma^{(k)})$ ,  $k = 0, 1$ , and  $\theta = (\theta^{(0)}, \theta^{(1)})$ , then  $\phi$  is distinct from  $\theta$  and is estimated by standard methods for logistic regression of  $r_i$  on  $x_i$ . Note that the parameters of  $\theta^{(1)}$  cannot be directly estimated due to the missing data. However, these parameters can be identified by exploiting assumptions about the missing data mechanism. It should be noted that this model is more restrictive than the normal random-effects model of Laird and Ware (1982), which permits a distinct design matrix for each response and can incorporate random effects. It only encompasses models for repeated-measures data where the means are modeled as functions of between-subject covariates. Little (1995) considers random-effects models, but does not give any details as to how pattern-mixture models would be developed.



The important step in developing pattern-mixture models is in making an assumption about the missing data mechanism. Suppose that

$$P(r_i = 1 | y_{(1)i}, y_{(2)i}, x_i) = g(y_{(2)i}, x_i), \quad (20)$$

where  $g$  is an arbitrary function of  $y_{(2)i}$  and  $x_i$ . Since missingness depends on the value of the missing variable,  $y_{(2)i}$ , this is a nonignorable missing data mechanism. This assumption can then be converted into constraints on the parameters by factorizing the distribution of  $y_i = (y_{(1)i}, y_{(2)i})$  in pattern  $k$  as

$$f(y_{(1)i}, y_{(2)i} | \theta^{(k)}, x_i, r_i = k) = f(y_{(2)i} | \theta_2^{(k)}, x_i, r_i = k) f(y_{(1)i} | \theta_{1.2}^{(k)}, y_{(2)i}, x_i, r_i = k),$$

where  $\theta_2^{(k)} = (\beta_{x:2.x}^{(k)}, \Sigma_{22.x}^{(k)})$  consists of the  $(p_2 \times q)$  regression coefficient matrix and  $(p_2 \times p_2)$  residual covariance matrix for the regression of  $y_{(2)i}$  on  $x_i$  within pattern  $k$ , and  $\theta_{1.2}^{(k)} = (\beta_{2:1.2.x}^{(k)}, \beta_{x:1.2.x}^{(k)}, \Sigma_{11.2.x}^{(k)})$  consists of the  $(p_1 \times p_2)$  and  $(p_1 \times q)$  regression coefficient matrices and  $(p_1 \times p_1)$  residual covariance matrix for the regression of  $y_{(1)i}$  on  $y_{(2)i}$  and  $x_i$  within pattern  $k$ .

Note that assumption (20) states that  $[r_i | y_{(1)i}, y_{(2)i}, x_i] \perp y_{(1)i}$ , which implies in turn that  $[y_{(1)i} | y_{(2)i}, x_i, r_i] \perp r_i$ , where  $\perp$  indicates independence. In other words, the conditional distribution of  $y_{(1)i}$  given  $y_{(2)i}$  and  $x_i$  is the same for both patterns so that

$$\theta_{1.2}^{(0)} = \theta_{1.2}^{(1)} = \theta_{1.2}. \quad (21)$$

This yields  $p_1(p_2 + q) + \frac{p_1(p_1+1)}{2}$  restrictions that help to identify the model, and likelihood inference now depends on the relative sizes of  $p_1$  and  $p_2$ .

The log-likelihood of  $\theta$  for the model in (19) is

$$\begin{aligned} \ell(\theta^{(0)}, \theta^{(1)}) &= \log \left[ \prod_{i=1}^N f(y_i | \beta, \Sigma, r_i = k) \right] \\ &= -\frac{N_0 p_1}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{11.x}^{(0)}| \\ &\quad - \frac{1}{2} \sum_{i=1}^{N_0} (y_{(1)i} - \beta_{x:1.x}^{(0)} x_i)' \Sigma_{11.x}^{(0)-1} (y_{(1)i} - \beta_{x:1.x}^{(0)} x_i) \\ &\quad - \frac{N_1 p_1}{2} \log(2\pi) - \frac{N_1}{2} \log |\Sigma_{11.x}^{(1)}| \\ &\quad - \frac{1}{2} \sum_{i=1}^{N_1} (y_{(1)i} - \beta_{x:1.x}^{(1)} x_i)' \Sigma_{11.x}^{(1)-1} (y_{(1)i} - \beta_{x:1.x}^{(1)} x_i) \\ &\quad - \frac{N_0 p_2}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{22.1x}^{(0)}| \\ &\quad - \frac{1}{2} \sum_{i=1}^{N_0} (y_{(2)i} - \beta_{1:2.1x}^{(0)} y_{(1)i} - \beta_{x:2.1x}^{(0)} x_i)' \Sigma_{22.1x}^{(0)-1} (y_{(2)i} - \beta_{1:2.1x}^{(0)} y_{(1)i} - \beta_{x:2.1x}^{(0)} x_i). \end{aligned}$$

The model has  $2p_1 q + p_1(p_1 + 1) + 2p_2(p_1 + q) + p_2(p_2 + 1)$  parameters, but only  $2p_1 q + p_1(p_1 + 1) + p_2(p_1 + q) + \frac{p_2(p_2+1)}{2}$  can be identified from the data, namely

$$\theta_{id} = (\beta_{x:1.x}^{(0)}, \Sigma_{11.x}^{(0)}, \beta_{x:1.x}^{(1)}, \Sigma_{11.x}^{(1)}, \beta_{1:2.1x}^{(0)}, \beta_{x:2.1x}^{(0)}, \Sigma_{22.1x}^{(0)}).$$

If  $p_1 = p_2$ , then the number of restrictions in (21) equals the number of unidentified parameters and the model is just identified. Maximum likelihood (ML) estimates for the identified parameters are obtained by standard complete data methods, namely two multivariate regressions of  $y_{(1)}$  on  $x$  and one multivariate regression of  $y_{(2)}$  on  $y_{(1)}$  and  $x$ . For example,

$$\begin{aligned}\widehat{\beta}_{x:1.x}^{(0)} &= (X'X)^{-1}X'Y, \\ \widehat{\Sigma}_{11.x}^{(0)} &= \frac{1}{N_0}(Y - X\widehat{\beta}_{x:1.x}^{(0)})'(Y - X\widehat{\beta}_{x:1.x}^{(0)}),\end{aligned}$$

where  $Y$  is an  $N_0 \times p_1$  matrix of responses and  $X$  is an  $N_0 \times q$  matrix of covariates. The estimates of interest, however, are from  $[y_{(1)i}, y_{(2)i} | x_i]$ , averaged over patterns,  $(\beta_{x:1.x}, \Sigma_{11.x}, \beta_{x:2.x}, \Sigma_{22.x}, \Sigma_{21.x})$ . These can be expressed as functions of the identified parameters and  $\phi$  by applying the identifying restrictions. The following ML estimates are then obtained by substituting the ML estimates of the identified parameters and  $\phi$  into these functions,

$$\begin{aligned}\widehat{\beta}_{x:1.x} &= (1 - \widehat{p}_x)\widehat{\beta}_{x:1.x}^{(0)} + \widehat{p}_x\widehat{\beta}_{x:1.x}^{(1)}, \\ \widehat{\Sigma}_{11.x} &= (1 - \widehat{p}_x)\widehat{\Sigma}_{11.x}^{(0)} + \widehat{p}_x\widehat{\Sigma}_{11.x}^{(1)} + \widehat{p}_x(1 - \widehat{p}_x)(\widehat{\beta}_{x:1.x}^{(0)} - \widehat{\beta}_{x:1.x}^{(1)})xx'(\widehat{\beta}_{x:1.x}^{(0)} - \widehat{\beta}_{x:1.x}^{(1)})', \\ \widehat{\beta}_{x:2.x} &= \widehat{\beta}_{x:2.x}^{(0)} + (\widehat{\beta}_{2:1.2x}^{(0)})^{-1}(\widehat{\beta}_{x:1.x}^{(0)} - \widehat{\beta}_{x:1.x}^{(0)}), \\ \widehat{\Sigma}_{22.x} &= \widehat{\Sigma}_{22.x}^{(0)} + (\widehat{\beta}_{2:1.2x}^{(0)})^{-1}(\widehat{\Sigma}_{11.x} - \widehat{\Sigma}_{11.x}^{(0)})(\widehat{\beta}_{2:1.2x}^{(0)'})^{-1}, \\ \widehat{\Sigma}_{21.x} &= \widehat{\Sigma}_{21.x}^{(0)} + (\widehat{\beta}_{2:1.2x}^{(0)})^{-1}(\widehat{\Sigma}_{11.x} - \widehat{\Sigma}_{11.x}^{(0)}),\end{aligned}$$

where  $\widehat{p}_x = P(r_i = 1 | \widehat{\phi}, x_i)$ . A modification of these equations is required if the resulting covariance matrices are not positive semidefinite. Specifically, if  $(\widehat{\Sigma}_{11.x} - \widehat{\Sigma}_{11.x}^{(0)})$  is not positive semidefinite, it is replaced by  $PQP'$  where  $P$  is the orthogonal matrix of eigenvectors of  $(\widehat{\Sigma}_{11.x} - \widehat{\Sigma}_{11.x}^{(0)})$  and  $Q$  is the diagonal matrix of eigenvalues of  $(\widehat{\Sigma}_{11.x} - \widehat{\Sigma}_{11.x}^{(0)})$  with the negative elements replaced by zero.

If  $p_1 > p_2$ , then the number of restrictions in (21) exceeds the number of unidentified parameters and the model is overidentified. Explicit ML estimates cannot be obtained, and an iterative method such as the EM algorithm is required. The complete data log-likelihood of  $\theta$  is

$$\begin{aligned}l(\theta^{(0)}, \theta^{(1)}) &= \log \left[ \prod_{i=1}^N f(y_i | \beta, \Sigma, r_i = k) \right] \\ &= -\frac{N_0 p_1}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{11.x}^{(0)}| \\ &\quad - \frac{1}{2} \sum_{i=1}^{N_0} (y_{(1)i} - \beta_{x:1.x}^{(0)} x_i)' \Sigma_{11.x}^{(0)-1} (y_{(1)i} - \beta_{x:1.x}^{(0)} x_i) \\ &\quad - \frac{N_1 p_1}{2} \log(2\pi) - \frac{N_1}{2} \log |\Sigma_{11.x}^{(1)}| \\ &\quad - \frac{1}{2} \sum_{i=1}^{N_1} (y_{(1)i} - \beta_{x:1.x}^{(1)} x_i)' \Sigma_{11.x}^{(1)-1} (y_{(1)i} - \beta_{x:1.x}^{(1)} x_i) \\ &\quad - \frac{N_0 p_2}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{22.1x}^{(0)}|\end{aligned}$$

$$\begin{aligned}
& -\frac{1}{2} \sum_{i=1}^{N_0} (y_{(2)i} - \beta_{1:2 \cdot 1x}^{(0)} y_{(1)i} - \beta_{x:2 \cdot 1x}^{(0)} x_i)' \Sigma_{22 \cdot 1x}^{(0)-1} (y_{(2)i} - \beta_{1:2 \cdot 1x}^{(0)} y_{(1)i} - \beta_{x:2 \cdot 1x}^{(0)} x_i) \\
& -\frac{N_1 p_2}{2} \log(2\pi) - \frac{N_1}{2} \log |\Sigma_{22 \cdot 1x}^{(1)}| \\
& -\frac{1}{2} \sum_{i=1}^{N_1} (y_{(2)i} - \beta_{1:2 \cdot 1x}^{(1)} y_{(1)i} - \beta_{x:2 \cdot 1x}^{(1)} x_i)' \Sigma_{22 \cdot 1x}^{(1)-1} (y_{(2)i} - \beta_{1:2 \cdot 1x}^{(1)} y_{(1)i} - \beta_{x:2 \cdot 1x}^{(1)} x_i).
\end{aligned}$$

From this it can be seen that the complete data sufficient statistics involving missing data are

$$S_{12}^{(1)} = \sum_{i=1}^{N_1} y_{(1)i} y_{(2)i}', \quad S_{x2}^{(1)} = \sum_{i=1}^{N_1} x_i y_{(2)i}', \quad S_{22}^{(1)} = \sum_{i=1}^{N_1} y_{(2)i} y_{(2)i}'.$$

The E-step at each iteration replaces these statistics by their expected values given the observed data and current parameter estimates, which can be calculated from the first and second moments of  $y_{(2)i}$ :

$$\begin{aligned}
E(y_{(2)i} | \hat{\beta}, \hat{\Sigma}, y_{(1)i}, x_i, r_i = 1) &= \hat{\beta}_{1:2 \cdot 1x}^{(1)} y_{(1)i} + \hat{\beta}_{x:2 \cdot 1x}^{(1)} x_i, \\
\text{Var}(y_{(2)i} | \hat{\beta}, \hat{\Sigma}, y_{(1)i}, x_i, r_i = 1) &= \hat{\Sigma}_{22 \cdot 1x}^{(1)}.
\end{aligned}$$

The M-step computes new parameter estimates by a complete-data maximization subject to the constraints induced by the missing data assumption. Therefore, for the restrictions of (21), the likelihood function for the complete data is rewritten as

$$\begin{aligned}
& \ell(\theta^{(0)}, \theta^{(1)}) \tag{22} \\
&= \log \left[ \prod_{i=1}^N f(y_i | \beta, \Sigma, r_i = k) \right] \\
&= -\frac{N_0 p_1}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{11 \cdot 2x}| \\
&\quad -\frac{1}{2} \sum_{i=1}^{N_0} (y_{(1)i} - \beta_{2:1 \cdot 2x} y_{(2)i} - \beta_{x:1 \cdot 2x} x_i)' \Sigma_{11 \cdot 2x}^{-1} (y_{(1)i} - \beta_{2:1 \cdot 2x} y_{(2)i} - \beta_{x:1 \cdot 2x} x_i) \\
&\quad -\frac{N_0 p_2}{2} \log(2\pi) - \frac{N_0}{2} \log |\Sigma_{22 \cdot x}^{(0)}| - \frac{1}{2} \sum_{i=1}^{N_0} (y_{(2)i} - \beta_{x:2 \cdot x}^{(0)} x_i)' \Sigma_{22 \cdot x}^{(0)-1} (y_{(2)i} - \beta_{x:2 \cdot x}^{(0)} x_i) \\
&\quad -\frac{N_1 p_2}{2} \log(2\pi) - \frac{N_1}{2} \log |\Sigma_{22 \cdot x}^{(1)}| - \frac{1}{2} \sum_{i=1}^{N_1} (y_{(2)i} - \beta_{x:2 \cdot x}^{(1)} x_i)' \Sigma_{22 \cdot x}^{(1)-1} (y_{(2)i} - \beta_{x:2 \cdot x}^{(1)} x_i).
\end{aligned}$$

Note that the E-step requires the regression of  $y_{(2)i}$  on  $y_{(1)i}$  and  $x_i$  for the pattern with missing data, whereas the M-step requires the regression of  $y_{(2)i}$  on  $x_i$  for each pattern and  $y_{(1)i}$  on  $y_{(2)i}$  and  $x_i$  pooled over patterns. The sweep operator (Little and Rubin, 2002, Chap. 6) facilitates the switching of the regressions needed for the E- and M-steps. Specifically,  $\theta_{2 \cdot 1}^{(1)} = (\beta_{1:2 \cdot 1x}^{(1)}, \beta_{x:2 \cdot 1x}^{(1)}, \Sigma_{22 \cdot 1x}^{(1)})$  are obtained by sweeping on the second and third blocks of the matrix

$$D = \begin{bmatrix} D_{11}^{-1} & \beta_{x:1 \cdot 2x}' & D_{12}^{-1} \\ \beta_{x:1 \cdot 2x} & \Sigma_{11 \cdot 2x} & \beta_{2:1 \cdot 2x} \\ D_{12}^{-1} & \beta_{2:1 \cdot 2x}' & D_{22}^{-1} \end{bmatrix}$$

where

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = -N_1 \begin{bmatrix} S_{xx}^{(1)} & S_{x_2}^{(1)'} \\ S_{2x}^{(1)} & S_{22}^{(1)} \end{bmatrix}^{-1}.$$

Note that the elements of  $D$  are calculated using the parameter estimates from the previous M-step. Once the E-step is completed, the missing parts of the following matrix,  $A_1$ , can be filled in, leading to

$$A_1 = \frac{1}{N_1} \begin{bmatrix} S_{xx}^{(1)} & S_{x_1}^{(1)} & S_{x_2}^{(1)} \\ S_{x_1}^{(1)'} & S_{11}^{(1)} & S_{21}^{(1)'} \\ S_{x_2}^{(1)'} & S_{12}^{(1)} & S_{22}^{(1)} \end{bmatrix}.$$

If we let

$$A_0 = \frac{1}{N_0} \begin{bmatrix} S_{xx}^{(0)} & S_{x_1}^{(0)} & S_{x_2}^{(0)} \\ S_{x_1}^{(0)'} & S_{11}^{(0)} & S_{21}^{(0)'} \\ S_{x_2}^{(0)'} & S_{12}^{(0)} & S_{22}^{(0)} \end{bmatrix}$$

and  $A = (N_0 A_0 + N_1 A_1)/N$ , then the M-step is completed by sweeping on the first block of  $A_0$  and  $A_1$  to obtain  $\theta_2^{(0)} = (\beta_{x:2;x}^{(0)}, \Sigma_{22;x}^{(0)})$  and  $\theta_2^{(1)} = (\beta_{x:2;x}^{(1)}, \Sigma_{22;x}^{(1)})$ , respectively. The values of  $\theta_{1,2} = (\beta_{2:1;2x}, \beta_{x:1;2x}, \Sigma_{11;2x})$  are obtained by sweeping on the first and third blocks of  $A$ . Notice that  $\theta_2^{(0)}$  is not affected by the E-step and so does not need iteration. This process yields ML estimates of  $(\beta_{x:2;x}^{(0)}, \Sigma_{22;x}^{(0)}, \beta_{x:2;x}^{(1)}, \Sigma_{22;x}^{(1)})$ . ML estimates of  $(\beta_{x:1;x}^{(0)}, \Sigma_{11;x}^{(0)}, \beta_{x:1;x}^{(1)}, \Sigma_{11;x}^{(1)})$  can be obtained from the regression of  $y_{(1)i}$  on  $x_i$  for each pattern, and ML estimates of  $\phi$  can be obtained from a logistic regression of  $r_i$  on  $x_i$ . The following functions of these ML estimates yield ML estimates of the parameters of interest:

$$\begin{aligned} \widehat{\beta}_{x:1;x} &= (1 - \widehat{p}_x) \widehat{\beta}_{x:1;x}^{(0)} + \widehat{p}_x \widehat{\beta}_{x:1;x}^{(1)}, \\ \widehat{\Sigma}_{11;x} &= (1 - \widehat{p}_x) \widehat{\Sigma}_{11;x}^{(0)} + \widehat{p}_x \widehat{\Sigma}_{11;x}^{(1)} + \widehat{p}_x (1 - \widehat{p}_x) (\widehat{\beta}_{x:1;x}^{(0)} - \widehat{\beta}_{x:1;x}^{(1)})_{xx'} (\widehat{\beta}_{x:1;x}^{(0)} - \widehat{\beta}_{x:1;x}^{(1)})', \\ \widehat{\beta}_{x:2;x} &= (1 - \widehat{p}_x) \widehat{\beta}_{x:2;x}^{(0)} + \widehat{p}_x \widehat{\beta}_{x:2;x}^{(1)}, \\ \widehat{\Sigma}_{22;x} &= (1 - \widehat{p}_x) \widehat{\Sigma}_{22;x}^{(0)} + \widehat{p}_x \widehat{\Sigma}_{22;x}^{(1)} + \widehat{p}_x (1 - \widehat{p}_x) (\widehat{\beta}_{x:2;x}^{(0)} - \widehat{\beta}_{x:2;x}^{(1)})_{xx'} (\widehat{\beta}_{x:2;x}^{(0)} - \widehat{\beta}_{x:2;x}^{(1)})', \\ \widehat{\Sigma}_{21;x} &= (1 - \widehat{p}_x) \widehat{\Sigma}_{21;x}^{(0)} + \widehat{p}_x \widehat{\Sigma}_{21;x}^{(1)} + \widehat{p}_x (1 - \widehat{p}_x) (\widehat{\beta}_{x:1;x}^{(0)} - \widehat{\beta}_{x:1;x}^{(1)})_{xx'} (\widehat{\beta}_{x:2;x}^{(0)} - \widehat{\beta}_{x:2;x}^{(1)})', \end{aligned}$$

where  $\widehat{p}_x = P(r_i = 1 | \widehat{\phi}, x_i)$ .

If  $p_1 < p_2$ , then the model remains underidentified and additional restrictive assumptions are needed to identify the model parameters. Little and Wang (1996) suggest assuming

$$P(r_i = 1 | y_{(1)i}, y_{(2)i}, x_i) = g(y_{(2s)i}, x_i),$$

where  $y_{(2s)i}$  is a subset of the variables  $y_{(2)i}$  with dimension  $p_{(2s)} \leq p_1$ . Using this approach, inference follows directly from the two scenarios previously described ( $p_1 = p_2$  case when  $p_1 = p_{(2s)}$  and  $p_1 > p_2$  case when  $p_1 > p_{(2s)}$ ). The choice of subset variables is important to the success of the model and reasons for dropout should be determined.

### 3.2.3 Discussion of Selection and Pattern-Mixture Models

All likelihood-based methods for handling nonignorable missing data must make some unverifiable assumptions, since the missing data mechanism included in the model depends on unobserved responses. Such a model is essentially nonidentifiable unless some unverifiable constraints are imposed. Inferences are only possible once these assumptions have been made, and the following aspects of the model need to be carefully considered: the bias and efficiency of parameter estimates, sensitivity to model specification, computational expense, and ease of implementation and interpretation. Selection and pattern-mixture models represent two different methods for handling nonignorable missing longitudinal data; each has its advantages and disadvantages.

Selection models directly model the distribution of primary interest, that is, the marginal distribution of the longitudinal outcomes. Thus, this method is more intuitive to most investigators. Selection models allow for a more natural way to model the missing data process, and since the missing data mechanism is modeled conditional on the repeated outcomes, it is very easy to formulate hypotheses about the missing data mechanism. However, to ensure identifiability, the set of outcomes is usually restricted in some way and arbitrary constraints must be applied to the missing data model. It is unclear how these restrictions on the missing data mechanism translate into assumptions about the distribution of the unobserved outcomes. Sensitivity of parameter estimates to model assumptions need to be considered, as well as the complexity of the computational algorithms required to fit the models.

Pattern-mixture models make specific assumptions about the distribution of the unobserved outcomes, and therefore, it may be easier to explore the sensitivity of results to model specification. By modeling the outcomes separately for each pattern, problems of identifiability are made explicit. Model identifiability is more obscure in the selection modeling approach, and in this case, one needs to characterize identifiability theoretically. Chen, Ibrahim, Shao (2004), Chen, Ibrahim, Shao (2006), and Chen, Ibrahim, Shao (2009) have carried out such investigations. The main drawback of pattern-mixture models is that the parameters of interest are not immediately available. The primary focus of inference is on the marginal distribution of the outcomes, which can only be obtained by averaging over patterns. Hence, one cannot examine the effects of the individual covariates on the marginal distribution of the outcomes in terms of the regression coefficients. Also, as shown in the previous section, the computations needed for a simple multivariate multiple regression with just one pattern of missing data are complex. It is possible that pattern-mixture models may be computationally intractable for random-effects models or more general patterns of incomplete data.

### 3.2.4 Conditional Linear Models

Several methods have been proposed for dealing with series of measurements that may be right censored due to death or withdrawal. The right censoring is termed informative if the censoring probabilities depend on an individual subject's underlying rate of change (slope) of the outcome variable. Thus, informative censoring is a special type of nonignorable missing data, and the class of joint models for longitudinal data and a nonignorable censoring process represent a specific case of the selection model. Wu and Carroll (1988) combine the normal random effects model with a probit model for the censoring process. They derive pseudo-maximum likelihood estimates and refer to their procedure as probit pseudo-maximum likelihood estimation (PPMLE). Wu and Bailey (1989) prove that under the probit model, the expectation of the slope for subject  $i$  is a monotonic increasing (decreasing)

function of the censoring time, and instead of modeling the censoring process, they propose a conditional linear model for the individual least squares estimated slope. This method can be described as an approximation to account for the informative right censoring when estimating and comparing changes of a continuous outcome variable.

Consider the following general framework. Assume that in a longitudinal study,  $n$  measurements on the outcome variable are planned to be made for each participant and that the participants are to be allocated into two equal sized treatment groups. Let  $y_i = (y_{i1}, \dots, y_{ij_i})$  be the observed outcome vector of serial measurements for subject  $i$ , where  $j_i \leq n$ . The repeated measurements of  $y_i$  are assumed to follow linear functions of time with normally distributed errors. Let  $\beta_i = (\beta_{i0}, \beta_{i1})'$  be the unobserved vector representing the true intercept and slope of the outcome variable for the  $i$ th subject, and let  $(\hat{\beta}_i | j_i) = \hat{\beta}_i$  be the usual least squares estimate of  $\beta_i$  based on the  $j_i$  observations. Furthermore, assume that when the  $i$ th subject belongs to the  $k$ th group,  $k = 1, 2$ ,  $\hat{\beta}_i$  follows a bivariate normal distribution. Thus

$$y_i = X_i \beta_i + e_i,$$

where

$$e_i \sim N_{n_i}(0, \sigma^2 I_{j_i}), \quad \beta_i \sim N_2(\beta_k, D),$$

and

$$\hat{\beta}_i | j_i = (X_i' X_i)^{-1} X_i' y_i,$$

where

$$X_i' = \begin{bmatrix} 1 & \dots & 1 \\ t_1 & \dots & t_{j_i} \end{bmatrix}.$$

The conditional linear model approach writes the slope as a linear function of the censoring time with normal errors. Specifically,

$$(\hat{\beta}_{i1} | j_i = j) = \gamma_0 k + \gamma_1 t_j + e_{kj}, \quad (23)$$

where  $E(e_{kj}) = 0$  and  $\text{Var}(e_{kj}) = \sigma_{kj}^2$ . Two methods to estimate the expected slopes,  $\beta_{k1}$ , were proposed by Wu and Bailey (1988, 1989). The linear minimum variance unbiased (LMVUB) procedure estimates  $\gamma_0 k$  and  $\gamma_1$  by weighted least squares so that

$$\text{LMVUB}(\beta_{k1}) = \hat{\gamma}_0 k + \hat{\gamma}_1 E_{ik}(t_{j_i}),$$

where  $E_{ik}(t_{j_i})$  is the expected value of the censoring time for the  $k$ th group (i.e., the sample mean for the  $k$ th group). The linear minimum mean squared error (LMMSE) estimate is a linear combination of the individual least squares slope estimates with the weights,  $W_{kj}$ , chosen to minimize the mean squared error under the linear model of (23) so that

$$\text{LMMSE}(\beta_{k1}) = \sum_{j=2}^n W_{kj} (\hat{\beta}_{k1} | j),$$

where

$$\hat{\beta}_{k1} | j = \frac{\sum_{i \in k} (\hat{\beta}_i | j_i = j)}{n_{kj}}$$

with  $n_{kj}$  denoting the number of subjects censored after  $j$  measurements were taken in the  $k$ th group. Wu and Bailey (1988) review PPMLE, LMVUB, and LMMSE, and compare these approaches together with the weighted and unweighted least squares estimates in the presence of informative censoring. Schluchter (1992) proposes a log-normal survival model which is a generalization of the conditional linear model that allows staggered patient entry and uses the exact censoring times of each individual.

#### 4 Generalized Linear Mixed Models

The generalized linear mixed model (GLMM) is the generalized linear model (GLM) extension of the normal linear random effects model. It is defined as follows. Suppose the sampling distribution of  $y_{ij}$ ,  $j = 1, \dots, n_i$ ,  $i = 1, \dots, N$ , is from an exponential family, so that

$$f(y_{ij}|\theta_{ij}, \tau) = \exp\{\tau[y_{ij}\theta_{ij} - a(\theta_{ij})] + c(y_{ij}, \tau)\}, \quad (24)$$

where  $\tau$  is a scalar dispersion parameter. Except for the normal random effects model, it will be assumed that  $\tau = \tau_0$ , where  $\tau_0$  is known, since  $\tau_0 = 1$  in the logistic and Poisson regression models. The  $y_{ij}$ 's are assumed to be independent given the random effects, and each  $y_{ij}$  has canonical parameter  $\theta_{ij}$ , which is related to the covariates by  $\theta(\eta_{ij})$ , where  $\eta_{ij} = x'_{ij}\beta + z'_{ij}b_i$  and  $x'_{ij}$  is a  $1 \times p$  vector denoting the  $j$ th row of  $X_i$  while  $z'_{ij}$  is a  $1 \times q$  vector denoting the  $j$ th row of  $Z_i$ . The link function,  $\theta(\cdot)$ , is a monotonic differentiable function. When  $\theta_{ij} = \eta_{ij}$ , the link is said to be canonical. Note that the GLMM has similarity with the normal random effects model in that we assume that conditional on the random effects,  $b_i$ , the repeated observations on subject  $i$  are independent. Letting  $y = (y_{11}, \dots, y_{Nn_N})'$ ,  $X = (X'_1, \dots, X'_N)'$ ,  $Z = \text{diag}(Z_1, \dots, Z_N)$ , and  $b = (b'_1, \dots, b'_N)'$ , the full likelihood based on  $N$  subjects for the GLMM is given by

$$f(y, b|\beta, D) = \prod_{i=1}^N \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i) f(b_i|D),$$

where  $f(b_i|D)$  is the distribution of  $b_i$ . As usual, it is assumed that  $b_i \sim N_q(0, D)$ , so that

$$f(b_i|D) = (2\pi)^{-q/2} |D|^{-1/2} \exp\left\{-\frac{1}{2} b'_i D^{-1} b_i\right\}.$$

To induce a correlation structure on the responses, inference is based on the marginal likelihood of  $\beta$  and  $D$  with the random effects integrated out. This is given by

$$f(y|\beta, D) = \int_{R^{Nq}} f(y, b|\beta, D) db, \quad (25)$$

where  $R^{Nq}$  denotes the  $Nq$  dimensional Euclidean space.

##### 4.1 Complete-data Estimation

If  $y$  is completely observed, then the likelihood function based on the observed data is given by (25). Note that

$$\begin{aligned} f(y|\beta, D) &= \int_{R^{Nq}} f(y, b|\beta, D) db \\ &= \int_{R^q} \dots \int_{R^q} \left[ \prod_{i=1}^N \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i) f(b_i|D) db_i \right] \\ &= \prod_{i=1}^N \left[ \int_{R^q} \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i) f(b_i|D) db_i \right]. \end{aligned} \quad (26)$$

Thus, the marginal likelihood involves evaluating  $Nq$ -dimensional integrals. For the general class of GLMM's, these integrals do not have a closed form and are very difficult to evaluate. This problem led to the development of quasi-likelihood based methods. Quasi-likelihood

was first introduced for the generalized linear model by Wedderburn (1974) who defined the quasi-likelihood function as follows. Suppose  $y_i, i = 1, \dots, N$ , is a set of observations with expectation  $E(y_i|\beta) = \mu_i$  and variance  $\text{Var}(y_i|\beta) = a(\tau)V(\mu_i)$ , where  $V(\mu_i)$  is some known function. The quasi-likelihood function,  $Q(y_i, \mu_i)$ , is defined by the relation

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}.$$

The log-likelihood is a special case of the quasi-likelihood function, but Wedderburn (1974) showed that one can use any function  $Q(y_i, \mu_i)$  that satisfies the above definition as a basis for defining a GLM and obtaining estimates of the  $\beta$ 's. In other words, GLM's can be used for any random variable as long as the mean, the mean function, the variance function, and the scale parameter are known.

In the GLMM, the conditional distribution of  $[y|\beta, b]$  plays the same role as the distribution of  $[y|\beta]$  in the fixed-effects GLM, and the joint quasi-likelihood function is the sum of the quasi-likelihoods of  $[y|\beta, b]$  and  $[b|D]$ . Since inference is based on the marginal likelihood of  $\beta$  and  $D$  with the random effects integrated out, an integrated quasi-likelihood function is used to estimate  $\theta = (\beta, D)$ . This is defined by

$$\exp(Q(y, \mu|b)) \propto |I_N \otimes D|^{-1} \int \exp \left\{ \left( -\frac{1}{2} \sum_{i=1}^N dev_i \right) - \frac{1}{2} b'(I_N \otimes D)^{-1} b \right\} db,$$

where  $dev_i$  denotes the deviance measure of fit for subject  $i$ ,  $b$  is the  $Nq \times 1$  vector of the  $b_i$ 's,  $I_N \otimes D$  is the  $Nq \times Nq$  covariance matrix of  $b$ , and the scalar dispersion parameter is assumed to equal one. Breslow and Clayton (1993) apply Laplace's method to approximate this function and show that

$$Q(y, \mu|b) \approx [y'\theta(\eta) - J'a(\theta(\eta))] - \frac{1}{2} b'(I_N \otimes D)^{-1} b,$$

where  $y$  is the  $n \times 1$  ( $n = \sum_{i=1}^N n_i$ ) vector of the  $y_{ij}$ 's,  $\theta(\eta)$  is the  $n \times 1$  vector of the  $\theta(\eta_{ij})$ 's,  $J$  is a column vector of ones, and  $a(\theta(\eta))$  is the  $n \times 1$  vector of the  $a(\theta(\eta_{ij}))$ 's. Differentiation with respect to  $\beta$  and  $b$  leads to score equations for these parameters and solutions can be obtained via Fisher scoring by iteratively solving

$$\begin{bmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + (I_N \otimes D)^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X'WY^* \\ Z'WY^* \end{bmatrix},$$

where  $W = GR^{-1}G$ ,  $Y^* = \hat{\eta} + (y - \hat{\mu})G^{-1}$ ,  $G = \frac{d\mu}{d\eta}$ , and  $R = \text{Var}(y|\beta, b)$ . Substitution of  $\hat{\beta}$  and  $\hat{b}$  into the approximated quasi-likelihood function and evaluation of  $W$  at  $\hat{\beta}$  and  $\hat{b}$  generates an approximate profile quasi-likelihood function for inference on  $D$ . Breslow and Clayton (1993) show that differentiating a REML version of this function with respect to the components of  $D$  yields the following estimating equations for the variance parameters:

$$-\frac{1}{2} \left[ (Y^* - X\hat{\beta})'V^{-1} \frac{\partial V}{\partial D_{ij}} V^{-1} (Y^* - X\hat{\beta}) - \text{tr} \left( P \frac{\partial V}{\partial D_{ij}} \right) \right] = 0,$$

where  $V = W^{-1} + Z(I_N \otimes D)Z'$  and  $P = V^{-1} - V^{-1}X[(X'V^{-1}X)^{-1}X'V^{-1}]$ . Breslow and Clayton (1993) call their procedure penalized quasi-likelihood (PQL) and assume that the scale parameter  $\tau$ , equals one. Wolfinger and O'Connell (1993) developed a refinement of



PQL called pseudo-likelihood (PL) which assumes  $\tau$  is unknown, and PQL is simply a special case of PL when  $\tau = 1$ . The method is implemented in the SAS procedure GLIMMIX, which recently has been augmented with Laplace approximation and numerical quadrature as well. Other software packages, such as R and MLwiN, have functions and procedures for PQL estimation, too, such as the R function `glmPQL`. Note that MLwiN allows for second-order PQL.

Alternatively, indeed, numerical integration methods have been proposed, based on so-called non-adaptive or adaptive Gaussian quadrature. The first of these methods implements a conventional quadrature rule. The second one makes use of the bell-shaped form of the conditional likelihood function, focusing attention on the portion with highest mass. While more accurate than the PQL and PL methods, numerical integration can be computationally intensive and very sensitive to starting values. It has been implemented in the SAS procedures GLIMMIX and NLMIXED.

## 4.2 Estimation with Nonignorable Missing Data

When some components of  $y$  are nonignorably missing, the estimation problem based on the observed data likelihood in (26) becomes more complicated since another integral over the missing data and the missing data mechanism would be introduced. Ibrahim, Chen, Lipsitz (2001) have developed a Monte Carlo EM algorithm for the selection model that facilitates straightforward estimation of  $\beta$  and  $D$ . Less work has been done in estimating parameters for the GLMM with nonignorable missing data using a pattern-mixture modeling approach. Fitzmaurice and Laird (2000) propose a method based on generalized estimating equations (Liang and Zeger, 1986), but theirs rather is an extension of Wu and Bailey's conditional linear model (1988; 1989) than a pattern-mixture model as described by Little (1993, 1995).

### 4.2.1 Selection Models

Recall that the complete data log-likelihood for the selection model is given by (5), where now  $f(y_i|\beta, b_i)$  is the GLMM given in (24). Assume  $y_i$  contains arbitrary and nonmonotone patterns of missingness so that some permutation of the indices of  $y_i$  can be written as  $y_i = (y_{mis,i}, y_{obs,i})$ . Ibrahim, Chen, Lipsitz (2001) use the Monte Carlo version of the EM algorithm for parameter estimation in the GLMM selection model with nonignorable missing response data. They write the E-step for an arbitrary GLMM in a weighted complete data form by using the general form of the EM by the Method of Weights (Ibrahim, 1990). Recall further that the E-step for the  $i$ th observation at the  $(t+1)$ st iteration can be written as (6), where  $\gamma^{(t)} = (\beta^{(t)}, D^{(t)}, \phi^{(t)})$  and  $f(y_{mis,i}, b_i|y_{obs,i}, r_i, \gamma^{(t)})$  represents the conditional distribution of the "missing" data,  $(y_{mis,i}, b_i)$ , given the observed data. The Monte Carlo EM algorithm given by Wei and Tanner (1990) requires generating a sample from

$$[y_{mis,i}, b_i|y_{obs,i}, r_i, \gamma^{(t)}]$$

for each  $i$ . This can be done via the Gibbs sampler by sampling from the complete conditionals  $[y_{mis,i}|y_{obs,i}, b_i, r_i, \gamma^{(t)}]$  and  $[b_i|y_{mis,i}, y_{obs,i}, r_i, \gamma^{(t)}]$ . Note that

$$f(y_{mis,i}|y_{obs,i}, b_i, r_i, \gamma^{(t)}) \propto f(y_i|b_i, \gamma^{(t)}) f(r_i|y_i, \gamma^{(t)}), \quad (27)$$

and

$$f(b_i|y_{mis,i}, y_{obs,i}, r_i, \gamma^{(t)}) \propto f(y_i|b_i, \gamma^{(t)}) f(b_i|\gamma^{(t)}). \quad (28)$$

The products on the right side of (27) and (28) are log-concave for the class of GLMM's in (24). This is true since  $f(y_i|b_i, \gamma^{(t)})$  is log-concave in the components of  $y_i$  and  $f(r_i|y_i, \gamma^{(t)})$  will be log-concave in the  $y_i$ 's if each  $[r_i|y_i, \gamma^{(t)}]$  is taken to be a logistic regression model. Also,  $f(y_i|b_i, \gamma^{(t)})$  and  $f(b_i|\gamma^{(t)})$  are both log-concave in the components of  $b_i$ . Since the sum of the logarithms of log-concave densities is a concave function, the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992) can be used to sample from

$$f(y_{mis,i\ell}|y_{mis,i,j}, j \neq \ell, y_{obs,i}, b_i, r_i, \gamma^{(t)}) \quad (29)$$

and

$$f(b_{i\ell}|b_{i,j}, j \neq \ell, y_{mis,i}, y_{obs,i}, r_i, \gamma^{(t)}), \quad (30)$$

where  $y_{mis,i\ell}$  denotes the  $\ell$ th component of  $y_{mis,i}$  ( $s_i \times 1$ ), and  $b_{i\ell}$  denotes the  $\ell$ th component of  $b_i$  ( $q \times 1$ ).

Suppose for the  $i$ th observation, a sample of size  $m_i, v_{i1}, \dots, v_{im_i}$ , is taken from the joint distribution of  $[y_{mis,i}, b_i|y_{obs,i}, r_i, \gamma^{(t)}]$  via the Gibbs sampler described by (29) and (30) in conjunction with the adaptive rejection algorithm as discussed above. Note that each  $v_{ik}$  will be an  $(s_i + q) \times 1$  vector for  $k = 1, \dots, m_i$  and that each  $v_{ik}$  depends on the iteration number which is suppressed. The E-step for the  $i$ th observation at the  $(t + 1)$ st iteration can now be written as

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) &= \frac{1}{m_i} \sum_{k=1}^{m_i} l(\gamma; y_{obs,i}, v_{ik}, r_i) \\ &= \frac{1}{m_i} \sum_{k=1}^{m_i} f(y_i|\beta, b_i) + \frac{1}{m_i} \sum_{k=1}^{m_i} f(b_i|D) + \frac{1}{m_i} \sum_{k=1}^{m_i} f(r_i|\phi, y_i). \end{aligned} \quad (31)$$

Note, this E-step takes a complete data weighted form in which each  $(y_{mis,i}, b_i)$  gets filled in by a set of  $m_i$  values, each contributing a weight of  $1/m_i$ . The E-step for all of the observations is given by

$$Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^N \sum_{k=1}^{m_i} \frac{1}{m_i} l(\gamma; y_{obs,i}, v_{ik}, r_i).$$

The resulting M-step is like one of complete data for the the GLMM and can be obtained as follows. Let

$$\dot{Q}(\gamma|\gamma^{(t)}) = \left( \dot{Q}^{(1)}(\beta|\gamma^{(t)}), \dot{Q}^{(2)}(D|\gamma^{(t)}), \dot{Q}^{(3)}(\phi|\gamma^{(t)}) \right)'$$

denote the score vector of  $Q(\gamma|\gamma^{(t)})$  so that

$$\dot{Q}(\gamma|\gamma^{(t)}) \equiv \sum_{i=1}^N \dot{Q}_i(\gamma|\gamma^{(t)}) = \sum_{i=1}^N \sum_{k=1}^{m_i} \frac{1}{m_i} \frac{\partial l(\gamma; y_{obs,i}, v_{ik}, r_i)}{\partial \gamma}.$$

Also, let

$$\ddot{Q}(\gamma|\gamma^{(t)}) \equiv \frac{\partial^2 Q(\gamma|\gamma^{(t)})}{\partial \gamma \partial \gamma'}$$

denote the Hessian matrix. Since  $\beta, D$ , and  $\phi$  are distinct, derivatives of  $l(\gamma; y_{obs,i}, v_{ik}, r_i)$  are straightforward to compute and  $\ddot{Q}(\gamma|\gamma^{(t)})$  is block diagonal in  $\beta, D$ , and  $\phi$ . Computation of the asymptotic covariance matrix of  $\hat{\gamma}$  can be done using Louis's (1982) method. The estimated observed information matrix of  $\gamma$  based on the observed data is given by

$$\mathcal{I}(\hat{\gamma}) = -\ddot{Q}(\hat{\gamma}|\hat{\gamma}) - \left\{ \left[ \sum_{i=1}^N \sum_{k=1}^{m_i} \frac{1}{m_i} S_i(\hat{\gamma}; y_{obs,i}, v_{ik}, r_i) S_i(\hat{\gamma}; y_{obs,i}, v_{ik}, r_i)' \right] \right\}$$

$$-\sum_{i=1}^N \dot{Q}_i(\hat{\gamma}|\hat{\gamma}) \dot{Q}_i(\hat{\gamma}|\hat{\gamma})' \Big\}, \quad (32)$$

where  $\hat{\gamma}$  is the estimate of  $\gamma$  at EM convergence and

$$S_i(\hat{\gamma}; y_{obs,i}, v_{ik}, r_i) = \left[ \frac{\partial l(\gamma; y_{obs,i}, v_{ik}, r_i)}{\partial \gamma} \right]_{\gamma=\hat{\gamma}}.$$

The quantities in (32) are easily computed since both  $\ddot{Q}_i(\hat{\gamma}|\hat{\gamma})$  and  $\dot{Q}_i(\hat{\gamma}|\hat{\gamma})$  are obtained from the M-step and  $S_i(\hat{\gamma}; y_{obs,i}, v_{ik}, r_i)$  is easily obtained outside of the EM algorithm.

The method described here is valid for arbitrary patterns of missing data in the response variable. The complexity of the estimate of  $D$  in the M-step depends on the structure of  $D$ . In any case, the estimation of  $D$  corresponds to estimation from a problem of complete data, and one can use any existing complete data software to estimate  $D$ . Also, note that models for the missing data mechanism in GLMM's do not change from the normal random-effects model.

#### 4.2.2 Pattern-mixture Models

Recall that pattern-mixture models stratify the incomplete data by the pattern of missing values and formulate distinct models within each stratum. Thus, the complete data log-likelihood is written as

$$l(\gamma) = \sum_{i=1}^N \log [f(y_i|\beta, b_i, r_i)] + \sum_{i=1}^N \log [f(b_i|D)] + \sum_{i=1}^N \log [f(r_i|\phi)].$$

Little work has been done using pattern-mixture models for GLMM's with nonignorable missing data. Ekholm and Skinner (1998) analyze longitudinal binary data using a pattern-mixture model, but do not generalize their method to the GLMM. Fitzmaurice and Laird (2000) develop a model for the GLMM with nonignorable dropout which they consider to be a *mixture* model based on Wu and Bailey's conditional linear model (1988; 1989), since dropout time is used as a covariate. This is the method that will be described here.

Consider the following notation. Assume that  $N$  subjects are to be observed at the same set of  $n$  occasions,  $\{t_1, \dots, t_n\}$ . Let  $y_i^c = (y_{i1}, \dots, y_{in})'$  denote the complete response vector for subject  $i$ , and let  $X_i$  denote the  $n \times p$  matrix of covariates for  $y_i^c$ . Each subject also has an event time,  $r_i$ , denoting the dropout time, which is thought to be related to  $y_i^c$ . Note that dropout implies that no subsequent repeated measures are made, so if  $r_i \leq t_n$ , then the  $i$ th subject is a dropout.  $r_i$  is considered to be discrete and occurring at  $t_{j+1}$  if the response at  $t_{j+1}$  is not observed. Let  $\phi_{ij} = P(r_i = t_j)$  and assume that  $\phi_{i1} = 0$  for all  $i$ . An additional category,  $\phi_{i(n+1)}$ , is included for the completers. The observed data for each subject consists of  $(y_i, X_i, r_i)$ .

Consider models for  $y_i$ , conditional of the time of dropout, that are of the following general form:

$$g(E(y_{ij}|\beta, r_i)) = z'_{ij}\beta, \quad (33)$$

where  $g(\cdot)$  is a known link function and the design vector,  $z_{ij}$ , includes the dropout time, the covariates, and their interactions. The parameters in this model have an unappealing interpretation due to the stratification by pattern of dropout, which may depend on the outcome.

Therefore, the parameter of interest is not  $\beta$ , but the marginal expectation of the repeated outcome averaged over the distribution of dropout times,

$$E(y_{ij}|\beta) = \mu_{ij} = \sum_{l=2}^{n+1} \phi_{il} g^{-1}(z'_{ij}\beta),$$

where  $z_{ij}$  includes the dropout time and  $x_{ij}$  and  $\phi_{il}$  depends on  $X_i$ . Since this estimate has been averaged over the distribution of the dropout times, the marginal mean will not, in general, follow the link function model assumed in (33). Therefore, the  $z_{ij}$  should be saturated in any covariate effects of interest so that comparisons can be made in terms of the marginal means.

Unlike the normal random effects model, it is difficult to account for the covariance among the repeated outcomes when the response variable is categorical, ordinal, or count data. Generalized estimating equations (GEE's) (see Liang and Zeger, 1986; Zeger and Liang, 1986) represent a general method for incorporating within-subject correlation in the GLM without having to completely specify the joint distribution of  $y_i$ . Only the forms of the first and second moments are required. Note that the GEE approach can accommodate any intermediate MCAR missingness in the outcome since each subject is allowed a distinct set of measurement times. The estimating equations for  $\beta$  with nonignorable missing data are given by

$$U(\beta) = \sum_{i=1}^N G'_i V_i^{-1} [y_i - E(y_i|\beta, r_i)] = 0,$$

where  $y_i$  is the  $n_i \times 1$  vector of observed responses,  $G_i = \frac{\partial E(y_i|\beta, r_i)}{\partial \beta}$ , and  $V_i$  is the  $n_i \times n_i$  working covariance matrix of  $y_i$ . Note that  $V_i$  depends on the marginal means,  $E(y_{ij}|\beta, r_i)$ , and a set of association parameters,  $\rho$ . Typically  $\rho$  is unknown and can be estimated with another set of estimating equations. It can be shown that  $N^{1/2}(\hat{\beta} - \beta)$  has an asymptotic normal distribution with mean 0 and covariance matrix

$$V_{\hat{\beta}} = \lim_{N \rightarrow \infty} N \left[ \sum_{i=1}^N G'_i V_i^{-1} G_i \right]^{-1} \left[ \sum_{i=1}^N G'_i V_i^{-1} \text{cov}(Y_i) V_i^{-1} G_i \right] \left[ \sum_{i=1}^N G'_i V_i^{-1} G_i \right]^{-1}.$$

Estimation of the dropout probabilities also needs to be considered. With a small number of discrete covariates, the dropout probabilities,  $\phi_{ij}$ , can be estimated as the sample proportion with each dropout time stratified by covariate pattern. The asymptotic covariance matrix of  $N^{1/2}(\hat{\phi} - \phi)$  is then given by

$$V_{\hat{\phi}} = \text{diag}(\phi) - \phi \phi'.$$

When the number of dropout times or covariates is large, then parametric models such as a multinomial log-linear regression model can be used to estimate  $\phi$ .

The appealing aspect of the mixture model presented above is that the SAS procedure GENMOD, or any other statistical software for GEE's, can be used to estimate  $\beta$ . The dropout times and their interactions with the other covariates are simply included as additional covariates in the model. The marginal means at times  $t_j$  can then be estimated by

$$\hat{\mu}_{ij} = \sum_{l=2}^{n+1} \hat{\phi}_{il} g^{-1}(z'_{ij}\hat{\beta}).$$

### 4.2.3 Semiparametric Methods

Robins, Rotnitzky and Zhao (1995) develop a class of estimators for generalized linear mixed models that are based on inverse probability weighted estimating equations when the data are MAR. Rotnitzky, Robins and Scharfstein (1998) extend this methodology to account for nonignorable nonresponse in the outcomes. Their conditional mean model of  $y_{it}$ ,  $t = 1, \dots, T$ , given the  $T \times p$  covariate matrix  $X_i$  follows the regression model

$$E(y_{it}|X_i) = k_t(X_i; \beta),$$

where  $k_t(X_i; \beta)$  is a known smooth function of  $\beta$ ,  $t = 1, \dots, T$ , and  $\beta$  is a  $p \times 1$  vector of unknown parameters on which inferences are to be made. Note that this model places no restrictions on the conditional mean of  $y_{it}$  given  $X_i$  at any time  $t$  and so is referred to as nonparametric.

Consider the following notation. Let  $v_{it}$  be a vector of time-dependent covariates that are not of interest. Define  $w_{it} = (v'_{it}, y_{it})'$ ,  $t = 1, \dots, T$ ,  $w_{i0} = (x'_{i0}, v'_{i0}, y_{i0})'$ ,  $w_i = (w'_{i0}, w'_{i1}, \dots, w'_{iT})'$ , and let  $r_{it}$  be an indicator variable for time  $t$ ,  $t = 1, \dots, T$ , that takes the value 1 if  $w_{it}$  is observed and 0 otherwise. Let  $\pi_i(1) = P(r_i = 1' | w_i)$  be the conditional probability of observing the full data,  $w_i$ , for the  $i$ th subject given  $w_i$ . In addition, suppose given  $w_i$ ,  $r_i$  is a vector of possibly correlated binary variables taking values in the set  $\{r = (r_1, \dots, r_T) : r = 0 \text{ or } 1, 1 \leq t \leq T\}$ . Letting  $\bar{r}_i = (r_{i1}, \dots, r_{i(t-1)})'$  and defining  $\bar{r}_1 = 1$ , the conditional distribution of  $r_i$  given  $w_i$  is

$$P(r_i = r | w_i) = \prod_{t=1}^T P(r_{it} = 1 | \bar{r}_i, w_i)^{r_t} P(r_{it} = 0 | \bar{r}_i, w_i)^{1-r_t} = \pi_i(r),$$

where  $P(r_{it} = 1 | \bar{r}_i, w_i)$  follow parametric models known up to a  $q \times 1$  parameter vector  $\alpha$ . That is, letting  $\lambda_{it} = P(r_{it} = 1 | \bar{r}_i, w_i)$ , assume that

$$\lambda_{it} = \lambda_{it}(\alpha),$$

where

$$\text{logit } \lambda_{it}(\alpha) = h_t(\bar{r}_i, w_i; \alpha),$$

and  $h_t(\bar{r}_i, w_i; \alpha)$  are known functions. This definition implies that  $\pi_i(1) = \pi_i(1; \alpha)$  and that  $\pi_i(r) = \pi_i(r; \alpha)$ .

In the complete data case, parameter estimates  $\hat{\beta}$  are found by solving the estimating equations

$$\sum_{i=1}^N d(X_i; \beta) [y_i - g(X_i; \beta)] = 0,$$

where  $d(X_i; \beta)$  is a  $p \times T$  matrix of fixed functions of  $X_i$  and  $\beta$  and

$$E(d(X_i; \beta) [y_i - g(X_i; \beta)]) = 0.$$

In the incomplete data case with unknown response probabilities, the parameters  $\beta$  and  $\alpha$  can be jointly estimated from solutions to a simultaneous set of  $p + q$  estimating equations,

$$\sum_{i=1}^N \frac{I(r_i = 1')}{\pi_i(1; \alpha)} \begin{bmatrix} d^{(1)}(X_i; \beta) \\ d^{(2)}(X_i; \beta) \end{bmatrix} [y_i - g(X_i; \beta)] - \begin{bmatrix} A_i^{(1)}(\alpha) \\ A_i^{(2)}(\alpha) \end{bmatrix},$$

where  $d^{(1)}(X_i; \beta)$  and  $d^{(2)}(X_i; \beta)$  are  $p \times T$  and  $q \times T$  fixed functions of  $X_i$  and  $\beta$ , and  $A_i^{(1)}(\alpha)$  and  $A_i^{(2)}(\alpha)$  are defined as

$$A_i^{(j)}(\alpha) = \sum_{r \neq 1} \left[ I(r_i = r) - \frac{I(r_i = 1')}{\pi_i(1; \alpha)} \pi_i(r; \alpha) \right] f_r^{(j)}(w_{ri}),$$

such that

$$\sum_{i=1}^N A_i^{(j)}(\alpha) = 0.$$

Rotnitzky, Robins and Scharfstein (1998) show that the solution  $\hat{\beta}$  to these equations is consistent and asymptotically normally distributed, provided that the conditional mean model and the model for the response probabilities are correctly specified. The variance of  $\hat{\beta}$  depends on the choice of functions  $d^{(j)}(X_i; \beta)$  and  $f_r^{(j)}(w_{ri})$ , and optimal choices for these functions are discussed in the paper. This approach is extended to semiparametric models for the dropout mechanism by Scharfstein, Rotnitzky and Robins (1999).

## 5 Nonignorable Missing Covariates and Responses in the GLMM

Lipsitz, Ibrahim, Fitzmaurice (1999) consider maximum likelihood estimation for the special case of nonignorable missing responses and MAR categorical covariates in longitudinal binary data. More generally, the work of Ibrahim, Chen, Lipsitz (2001) involving missing nonignorable responses in GLMM's was extended to include both nonignorable missing responses and/or covariates for the normal mixed model in Stubbendick and Ibrahim (2003) and for the multivariate probit model by Stubbendick and Ibrahim (2006). Following Stubbendick and Ibrahim (2003), the E-step for the  $i^{\text{th}}$  observation at the  $(t+1)^{\text{st}}$  iteration for the normal mixed model is

$$\begin{aligned} Q_i(\gamma | \gamma^{(t)}) &= E(l(\gamma; y_i, X_i, b_i, r_i) | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) \\ &= \int \int \int \log [f(y_i | \beta, \sigma^2, X_i, b_i)] f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) db_i dX_{mis,i} dy_{mis,i} \\ &\quad + \int \int \int \log [f(X_i | \alpha)] f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) db_i dX_{mis,i} dy_{mis,i} \\ &\quad + \int \int \int \log [f(b_i | D)] f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) db_i dX_{mis,i} dy_{mis,i} \\ &\quad + \int \int \int \log [f(r_i | \phi, y_i, X_i)] f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) db_i dX_{mis,i} dy_{mis,i} \\ &\equiv I_1 + I_2 + I_3 + I_4, \end{aligned}$$

where  $\gamma^{(t)} = (\beta^{(t)}, \sigma^{2(t)}, D^{(t)}, \phi^{(t)})$  and  $f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)})$  represents the conditional distribution of the "missing" data,  $(y_{mis,i}, X_{mis,i}, b_i)$ , given the observed data.

To integrate out  $b_i$  from  $I_1$  and  $I_3$ , write

$$f(y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) = f(b_i | y_i, \gamma^{(t)}) f(y_{mis,i}, X_{mis,i} | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}),$$

where

$$(b_i|y_i, \gamma^{(t)}) \sim N_q \left( \Sigma_i^{(t)} Z_i' (y_i - X_i \beta^{(t)}) / \sigma^{2(t)}, \left[ \sigma^{-2(t)} Z_i' Z_i + (D^{(t)})^{-1} \right]^{-1} \right).$$

Then, to complete the E-step, samples only need to be taken from

$$[y_{mis,i}, X_{mis,i} | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}].$$

This distribution can be written up to a constant of proportionality as

$$\begin{aligned} & f(y_{mis,i}, X_{mis,i} | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}) \\ & \propto \exp \left( -\frac{1}{2} (y_i - X_i \beta^{(t)})' (Z_i D^{(t)} Z_i' + \sigma^{2(t)} I_{n_i})^{-1} (y_i - X_i \beta^{(t)}) \right) \\ & \quad \times f(r_i | y_{mis,i}, y_{obs,i}, X_{mis,i}, X_{obs,i}, \gamma^{(t)}) f(X_{mis,i} | X_{obs,i}, \gamma^{(t)}), \end{aligned} \quad (34)$$

which has the form of a normal density times a logistic regression for the  $r_i$ 's times some sort of regression for the  $X_{mis,i}$ 's. If this distribution is from the class of concave log-densities, then Gibbs sampling from (34) is straightforward using the adaptive rejection algorithm of Gilks and Wild (1992).

This methodology has been extended to the GLMM by Stubbendick and Ibrahim (2006). Thus, an MCEM sample must be generated from  $[y_{mis,i}, X_{mis,i}, b_i | y_{obs,i}, X_{obs,i}, r_i, \gamma^{(t)}]$  for each  $i$ . This can be done using the Gibbs sampler by sampling from the complete conditionals,  $[y_{mis,i} | y_{obs,i}, X_{mis,i}, X_{obs,i}, b_i, r_i, \gamma^{(t)}]$ ,  $[X_{mis,i} | y_{mis,i}, y_{obs,i}, X_{obs,i}, b_i, r_i, \gamma^{(t)}]$ , and  $[b_i | y_{mis,i}, y_{obs,i}, X_{mis,i}, X_{obs,i}, r_i, \gamma^{(t)}]$ . Note that

$$\begin{aligned} & f(y_{mis,i} | y_{obs,i}, X_{mis,i}, X_{obs,i}, b_i, r_i, \gamma^{(t)}) \\ & \propto f(y_i | X_i, b_i, \gamma^{(t)}) f(r_i | y_i, X_i, \gamma^{(t)}), \end{aligned} \quad (35)$$

$$\begin{aligned} & f(X_{mis,i} | y_{mis,i}, y_{obs,i}, X_{obs,i}, b_i, r_i, \gamma^{(t)}) \\ & \propto f(y_i | X_i, b_i, \gamma^{(t)}) f(r_i | y_i, X_i, \gamma^{(t)}) f(X_{mis,i} | X_{obs,i}, \gamma^{(t)}), \end{aligned} \quad (36)$$

$$\begin{aligned} & f(b_i | y_{mis,i}, y_{obs,i}, X_{mis,i}, X_{obs,i}, r_i, \gamma^{(t)}) \\ & \propto f(y_i | X_i, b_i, \gamma^{(t)}) f(b_i | \gamma^{(t)}). \end{aligned} \quad (37)$$

When the products on the right hand side of (35)–(37) are log-concave for the class of GLMM's, then the Gibbs sampler along with adaptive rejection algorithm of Gilks and Wild (1992) can be used to sample from the complete conditionals.

Allowing for nonignorable missing responses *and* covariates presents several additional modeling and computational challenges compared to just the missing response situation. First, a covariate distribution needs to be specified and its parameters estimated. This is done by specifying the covariate distribution via a sequence of one-dimensional conditional distributions as

$$\begin{aligned} f(x_{ij1}, \dots, x_{ijp} | \alpha) &= f(x_{ijp} | x_{ij1}, \dots, x_{ij(p-1)}, \alpha_p) \\ & \quad \times f(x_{ij(p-1)} | x_{ij1}, \dots, x_{ij(p-2)}, \alpha_{(p-1)}) \\ & \quad \times \dots \end{aligned}$$

$$\times f(x_{ij2}|x_{ij1}, \alpha_2)f(x_{ij1}|\alpha_1), \quad (38)$$

where  $x_{ijm}$  is the  $m$ th covariate for individual  $i$  at time  $j$ ,  $\alpha_k$  is a vector of indexing parameters for the  $k$ th conditional distribution,  $\alpha = (\alpha'_1, \dots, \alpha'_p)'$ , and the  $\alpha_k$ 's are distinct. Note that (38) only needs to be specified for those covariates that are missing. Second, identifiability of the model needs to be carefully considered. Third, efficient computational strategies are needed since this model can be computationally intensive in general.

### 5.1 Model Assessment and Sensitivity

Unfortunately, the parametric forms of the assumed missing data mechanism and the covariate model are not testable from the data. Many models need to be evaluated owing to the numerous possibilities for the missing data mechanism and/or the covariate distribution and for carrying out sensitivity analyses. In addition, issues related to bias, efficiency, and model fit need to be addressed. In the presence of missing data, Lipsitz et al (2001) examine and Fitzmaurice et al (2001) examine bias issues in longitudinal data. Chen, Ibrahim, Chen, Senchaudhuri (2008) examine bias and efficiency issues in regression models with missing responses and/or missing covariates. To address general issues regarding model fit and assessment in the presence of missing data, new methods are needed for defining residuals, diagnostic measures, assessing model fit, and assessing the influence of model perturbations for all types of models, such as GLMs, survival models, and models for longitudinal data. This is a currently a growing, active and open research area. AIC and BIC are common model assessment tools under the frequentist paradigm. In the presence of missing data, the definition of the AIC/BIC criterion is not clear. Ibrahim (Zhu) define AIC as  $AIC = -2Q(\hat{\gamma}|\hat{\gamma}) + 2d$ , where  $d$  is the total number of parameters in the models and  $Q(\hat{\gamma}|\hat{\gamma})$  is the  $Q$  function from the EM algorithm at convergence. Similarly, they define BIC as  $BIC = -2Q(\hat{\gamma}|\hat{\gamma}) + \log(N)d$ . Such measures can be used to assess fit in models for longitudinal data.

A more general framework for model assessment in complete data problems is given in Cook (1986), where he describes a method for assessing the local influence of minor perturbations of a statistical model. His method uses the geometric normal curvature to characterize the behavior of an influence graph based on a well-behaved likelihood function. In the context of the linear mixed model with complete data, Beckman, Nachtsheim and Cook (1987) use local influence to assess the effect of perturbing the error variances, the random-effects variances, and the response vector. Lesaffre and Verbeke (1998) show that the local influence approach is also useful for the detection of influential subjects in a longitudinal data analysis. Zhu and Lee (2001) apply Cook's approach to the conditional expectation of the complete data log-likelihood function in the EM algorithm instead of the more complicated observed data log-likelihood function. Their Q-displacement function,  $2[Q(\hat{\gamma}|\hat{\gamma}) - Q(\hat{\gamma}(\omega)|\hat{\gamma})]$ , will be explored as a method of assessing the local influence of perturbations of selection models with nonignorable missing data. Zhu, Ibrahim, Shi (2009) examine residuals, diagnostic measures and goodness of fit statistics for GLMs with missing covariate data. Shi, Zhu, Ibrahim (2009) examines local influence approaches for GLMs with missing covariate data, and Garcia, Ibrahim, Zhu (2009) investigates variable selection in GLMs with missing covariate data using penalized likelihood approaches. These procedures are currently being extended to longitudinal data.



## 6 Shared-parameter Models

Interest in methods for joint modelling of longitudinal and survival time data has developed considerably in recent years (see, e.g., Pawitan and Self, 1993; DeGruttola and Tu, 1994; Taylor, Cumberland, and Sy, 1994; Faucett and Thomas, 1996; Lavalley and DeGruttola, 1996; Hogan and Laird, 1997, 1998; Henderson, Diggle and Dobson, 2000; Xu and Zeger, 2001; Brown and Ibrahim, 2003a,b; Ibrahim, Chen, Sinha, 2004; Chen, Ibrahim, Shao, 2004; Brown, Ibrahim, DeGruttola, 2005; Chi and Ibrahim, 2006, 2007).

Broadly speaking, there are three main reasons to consider such models. First, a time-to-event outcome may be measured alongside a longitudinal covariate. Such a joint model then allows, in a natural way, for incorporation of measurement error present in the longitudinal covariate into the model. Second, a number of researchers have used joint modeling methods to exploit longitudinal markers as surrogates for survival. Tsiatis, DeGruttola and Wulfsohn (1995), for instance, propose a model for the relationship of survival to longitudinal data measured with error and, using Prentice's (1989) criteria, examine whether CD4 counts may serve as a useful surrogate marker for survival in patients with AIDS. Xu and Zeger (2001) investigate the issue of evaluating multiple surrogate endpoints and discuss a joint latent model for a time to clinical event and for repeated measures over time on multiple biomarkers that are potential surrogates. In addition, they propose two complementary measures to assess the relative benefit of using multiple surrogates as opposed to a single one. Another aspect of the problem, discussed by Henderson, Diggle and Dobson (2000), Brown and Ibrahim (2003a), Brown and Ibrahim (2003b), Ibrahim, Chen, Sinha (2004), Chen, Ibrahim, Sinha (2004(@)), Brown, Ibrahim, DeGruttola (2005), Chi and Ibrahim (2006), and Chi and Ibrahim (2007), is the identification of longitudinal markers for survival. These authors focus on the use of longitudinal marker trajectories to investigate the association between a longitudinal marker and survival. Renard et al (2002) used a joint model to explore the usefulness of prostate-specific antigen as a marker for prostate cancer.

Third, and most relevant for us here, such joint models can be used when incomplete longitudinal data are collected. Whenever data are incomplete, one should *a priori* consider the joint distribution of responses and missing data process. In this sense, selection models and pattern-mixture models are merely convenient ways to decompose this joint distribution. In a number of applications, it may be attractive to write this joint distribution in terms of latent variables, latent classes, or random effects. This leads to so-called shared-parameter models. In principle, one can augment the full-data distribution with random effects

$$f(y_i, r_i, b_i | X_i, W_i, Z_i, \theta, \psi, \xi), \quad (39)$$

and then still consider the selection-model factorization

$$\begin{aligned} & f(y_i, r_i, b_i | X_i, W_i, Z_i, \theta, \psi) \\ &= f(y_i | X_i, b_i, \theta) f(r_i | y_i, b_i, W_i, \psi) f(b_i | Z_i, \xi) \end{aligned} \quad (40)$$

and the pattern-mixture model factorization

$$\begin{aligned} & f(y_i, r_i, b_i | X_i, W_i, Z_i, \theta, \psi, \xi) \\ &= f(y_i | r_i, b_i, X_i, \theta) f(r_i | b_i, W_i, \psi) f(b_i | Z_i, \xi). \end{aligned} \quad (41)$$

Here,  $Z_i$  and  $\xi$  are covariates and parameters, respectively, describing the random-effects distribution. Little (1995) refers to such decompositions as random-coefficient selection and pattern-mixture models, respectively.

Important early references to such models are Wu and Carroll (1988) and Wu and Bailey (1988, 1989). Wu and Carroll (1988) proposed this kind of model for what they termed informative right censoring. For a continuous response, Wu and Carroll suggested using a conventional Gaussian random-coefficient model combined with an appropriate model for time to dropout, such as proportional hazards, logistic or probit regression. The combination of probit and Gaussian response allows explicit solution of the integral and was used in their application.

In a slightly different approach to modeling dropout time as a continuous variable in the latent variable setting, Schluchter (1992) and DeGruttola and Tu (1994) proposed joint multivariate Gaussian distributions for the latent variable(s) of the response process and a variable representing time to dropout. The correlation between these variables induces dependence between dropout and response. To permit more realistic distributions for dropout time, Schluchter proposed that dropout time itself should be some monotone transformation of the corresponding Gaussian variable. The use of a joint Gaussian representation does simplify computational problems associated with the likelihood. There are clear links here with the Tobit model and this is made explicit by Cowles, Carlin and Connett (1996) who use a number of correlated latent variables to represent various aspects of an individual's behaviour, such as compliance and attendance at scheduled visits. Models of this type handle non-monotone missingness quite conveniently. There are many ways in which such models can be extended and generalized.

An important simplification arises when  $Y_i$  and  $R_i$  are assumed independent, given the random effects. We then obtain shared-parameter decomposition:

$$\begin{aligned} f(y_i, r_i, b_i | X_i, W_i, Z_i, \theta, \psi, \xi) \\ = f(y_i | X_i, b_i, \theta) f(r_i | W_i, b_i, \psi) f(b_i | Z_i, \xi). \end{aligned} \quad (42)$$

This route was followed by Follman and Wu (1995). Note that, when  $b_i$  is assumed to be discrete, a latent-class or mixture model follows. Rizopoulos, Verbeke and Molenberghs (2008) study the impact of random-effects misspecification in a shared parameter model. Beunckens et al (2008) combine continuous random effects with latent classes, leading to the simultaneous use of mixture and mixed-effects models ideas. It is very natural to handle random-coefficient models, and in particular shared-parameter models, in a Bayesian framework. Examples in the missing data setting are provided by Best et al (1996) and Carpenter, Pocock and Lamm (2002).

## 7 Bayesian Methods

Daniels and Hogan (2008) provide a comprehensive survey of Bayesian methods for longitudinal models with missing data. We refer the reader to their book and the many references therein. Here, we only provide a brief discussion of implementational and methodologic issues for the Bayesian paradigm in the presence of missing data. Fully Bayesian methods require specifying priors all of the parameters as well as specifying distributions for the missing covariates and/or missing data mechanisms, along with the sampling distribution of the response variable. We note here that Bayesian methods for any missing data problem, are in principal, quite straightforward to implement compared to the no missing data situation. This is due to the fact that all one needs to do in the Bayesian paradigm is to add additional steps to the Gibbs sampler, for example, to sample from the full conditional distributions of the missing data. Such steps can be easily incorporated into an existing Gibbs

sampler for a no missing data problem, and will be generally easier to implement than the MCEM algorithm discussed earlier. These fully Bayesian procedures can be easily implemented in WinBUGS or PROC MCMC in SAS for all types of models including models for longitudinal data.

However, new issues arise in fitting Bayesian models with missing data that do not arise in the frequentist development. First, one has to ensure that the posterior distribution is proper when using improper priors, as it is very easy for the posterior to be improper especially in nonignorable missing data settings. These issues as well as other modeling and elicitation issues are discussed in Ibrahim et al (2001) (Chapter 8), [Ibrahim et al (2002), Chen et al (2002), Chen, Ibrahim, Shao (2004), Chen, Ibrahim, Shao (2006), and Ibrahim Chen, Kim (2008)]. Second, even when using proper priors, if the model is weakly identifiable, which is often the case in many nonignorable missing data problems, the inferences may be quite sensitive to the choices of the hyperparameters, and one needs clever strategies for specifying informative priors that do not dominate the likelihood. Such strategies are outlined in Huang, Chen, Ibrahim (2005) for GLMs that can be easily extended to models for longitudinal data. Thirdly, it is conceivable that fully Bayesian methods may be more computationally intensive than their frequentist counterparts and Markov chain Monte Carlo convergence may not be easily achieved.

## 8 Concluding Remarks

Problems associated with incompletely gathered data, especially in longitudinal and clinical studies, have received considerable attention in recent times (Verbeke and Molenberghs, 2000; Fitzmaurice et al, 2004; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007; Daniels and Hogan, 2008; Fitzmaurice, Davidian, Verbeke, Molenberghs, 2008).

To efficiently describe these issues, a formal taxonomy, as laid out in this paper, is called for. We have placed emphasis on: (1) missing data patterns (monotone, non-monotone); (2) missing data mechanisms (MCAR, MAR, MNAR); (3) modeling frameworks (selection, pattern-mixture, and shared-parameter models); (4) inferential paradigms (likelihood, Bayesian, frequentist); (5) ignorability; and (6) outcomes types (continuous/linear, non-continuous/generalized linear). Finally, some attention has been devoted to sensitivity analysis frameworks.

Thanks to advances in terms of both available methodology and efficient implementations thereof, not in the least in generally available statistical software tools, such as SAS, SPSS, SPlus, and R, quite advanced analyses are within reach and there no longer is a need to focus on such simplistic methods as complete-case analysis or last observation carried forward, to name but a few. At the same time, all methods, no matter how sophisticated, rest to some extent on unverifiable assumptions, owing to the simple fact that the missing data are . . . unobserved. Therefore, rather than placing belief in a single such model, it should be supplemented with appropriate forms of sensitivity analysis.

## References

- Beckman RJ, Nachtsheim CJ, Cook RD (1987) Diagnostics for mixed-model analysis of variance. *Technometrics* 29:413–426
- Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG (1996) Bayesian analysis of realistically complex models. *J Roy Statist Soc Ser A* 159:323–342

- Beunckens C, Molenberghs G, Verbeke G, Mallinckrodt C (2008) A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics* 64(1):96–105
- Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. *J Amer Statist Assoc* 88:9–25
- Brown, ER, Ibrahim, JG (2003a) A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* 59:221–228
- Brown, ER, Ibrahim, JG (2003b) Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* 59:686–693
- Brown, ER, Ibrahim, JG, DeGruttola, V (2005) A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics* 61:64–73
- Carpenter J, Pocock S, Lamm CJ (2002) Coping with missing data in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine* 21:1043–1066
- Chen, M-H, Ibrahim, JG, Lipsitz, SR (2002) Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis* 8:117–146.
- Chen, M-H, Ibrahim, JG Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* 57:43–52
- Chen, M-H, Ibrahim, JG, Shao, Q-M (2004) Propriety of the posterior distribution and existence of the maximum likelihood estimator for regression models with covariates missing at random. *J Amer Statist Assoc* 99:421–438
- Chen, M-H, Ibrahim, JG, Shao, Q-M (2006) Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* 93:791–807
- Chen, M-H, Ibrahim, JG, Shao, Q-M (2009) Model identifiability for the Cox regression model with applications to missing covariates. *J Mult Anal*, in press
- Chen, M-H, Ibrahim, JG, Sinha D (2004) A new joint model for longitudinal and survival data with a cure fraction. *J Mult Anal* 91:18–34
- Chen, Q, Ibrahim, JG (2006) Missing covariate and response data in regression models. *Biometrics* 62:177–184
- Chen, Q, Zeng, D, Ibrahim, JG (2007) Sieve maximum likelihood estimation for regression models with covariates missing at random. *J Amer Statist Assoc* 102:1309–1317
- Chen, Q, Ibrahim, JG, Chen, M-H, Senchaudhuri, P (2008) Theory and inference for regression models with missing responses and covariates. *J Mult Anal* 99:1302–1331
- Chi, Y, Ibrahim, JG (2006) Joint models for multivariate longitudinal and survival data. *Biometrics* 62:432–445
- Chi, Y, Ibrahim, JG (2007) A new class of joint models for longitudinal and survival data accommodating zero and non-zero cure fractions: a case study of an international breast cancer study group trial. *Statistica Sinica* 17:445–462
- Cook RD (1986) Assessment of local influence. *J Roy Statist Soc Ser B* 48:133–169
- Cowles MK, Carlin BP, Connert JE (1996) Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *J Amer Statist Assoc* 91:86–98
- Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward, MG (2009) Shared-parameter models and missingness at random. Submitted for publication
- Daniels MJ, Hogan JW (2008) *Missing Data in Longitudinal Studies*. Chapman and Hall, London
- DeGruttola V, Tu XM (1994) Modelling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics* 50:1003–1014
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc Ser B* 39:1–38
- Diggle, PJ, Heagerty, P, Liang, K-Y, Zeger, SL (2002) *Analysis of longitudinal data*. Oxford University Press, Oxford
- Diggle P, Kenward MG (1994) Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* 43:49–93
- Ekhholm A, Skinner C (1998) The Muscatine children's obesity data reanalysed using pattern mixture models. *Applied Statistics* 47:251–263
- Faucett CL, Thomas DC (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 15:1663–1685
- Fitzmaurice, GM, Davidian, M, Verbeke, G, Molenberghs, M (2008) *Longitudinal Data Analysis*. Chapman and Hall, London
- Fitzmaurice GM, Laird NM (2000) Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* 1:141–156
- Fitzmaurice GM, Laird NM, Ware JH (2004) *Applied Longitudinal Analysis*. John Wiley and Sons, New York
- Fitmaurice, GM, Lipsitz, SR, Molenberghs, G, Ibrahim, JG (2001) Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics* 57:15–21
- Fitmaurice, GM, Lipsitz, SR, Ibrahim, JG, Gelber, R, Lipshultz, S (2006) Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics* 7:469–485
- Follman D, Wu M (1995) An approximate generalized linear model with random effects for informative missing data. *Biometrics* 51:151–168
- Garcia, RI, Ibrahim, JG, Zhu, H (2009) Variable selection for regression models with missing data. *Statistica Sinica*, in press.
- Gilks WR, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 41:337–348
- Henderson R, Diggle P, Dobson A (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1:465–480

- Herring, AH, Ibrahim, JG (2001) Likelihood-based methods for missing covariates in the Cox proportional hazards model. *J Amer Statist Assoc* 96:292–302
- Herring, AH, Ibrahim, JG (2002) Maximum likelihood estimation in random effects cure rate models with nonignorably missing covariates. *Biostatistics* 3:387–405
- Herring, AH, Ibrahim, JG, Lipsitz, SR (2002) Frailty models with missing covariates. *Biometrics* 58:98–109
- Herring, AH, Ibrahim, JG, Lipsitz, SR (2004) Nonignorably missing covariate data in survival analysis: a case study of an international breast cancer study group trial. *Applied Statistics* 53:293–310
- Hogan JW, Laird NM (1997) Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 16:239–257
- Hogan JW, Laird NM (1998) Increasing efficiency from censored survival data using random effects from longitudinal covariates. *Statistical Methods in Medical Research* 7:28–48
- Huang, L, Chen, M-H, Ibrahim, JG (2005) Bayesian analysis for generalized linear models with nonignorably missing covariates. *Bioometrics* 61:767–780
- Ibrahim JG (1990) Incomplete data in generalized linear models. *J Amer Statist Assoc* 85:765–769
- Ibrahim, JG, Chen, M-H, Kim, S (2008) Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Analysis* 14:496–520
- Ibrahim, JG, Chen, MH, Lipsitz, SR (1999) Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* 55:591–596
- Ibrahim JG, Chen M-H, Lipsitz SR (2001) Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 88:551–564
- Ibrahim JG, Chen M-H, Lipsitz SR (2002) Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics* 30:55–78
- Ibrahim, JG, Chen, M-H, Sinha, D (2004) Bayesian methods for joint modeling of longitudinal and survival data with applicants to cancer vaccine trials. *Statistica Sinica* 14:863–883
- Ibrahim, JG, Chen, M-H, Lipsitz, SR, Herring, AH (2005) Missing data methods in generalized linear models: a comparative review. *J Amer Statist Assoc* 100:332–346.
- Ibrahim, JG, Lipsitz, SR, Chen, M-H (1999) Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J Roy Statist Soc Ser B* 61:173–190.
- Ibrahim, JG, Zhu, H, Tang, N (2008) Model selection criteria for missing data problems using the EM algorithm. *J Amer Statist Assoc* 103:1648–1658
- Jennrich RI, Schluchter MD (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42:805–820
- Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* 38:963–974
- Lavalley MP, DeGruttola V (1996) Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine* 15:2289–2305
- Lesaffre E, Verbeke G (1998) Local influence in linear mixed models. *Biometrics* 54:570–582
- Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22
- Lipsitz, SR, Ibrahim, JG, Fitzmaurice, GM (1999) Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* 55:214–223
- Lipsitz, SR, Ibrahim, JG, Molenberghs, G (2000) Using a Box-Cox transformation in the analysis of longitudinal data with incomplete responses. *Applied Statistics* 49:287–296
- Lipsitz, SR, Parzen, M, Molenberghs, G, Ibrahim, JG (2001) Testing for bias in weighted estimating equations. *Biostatistics* 2:295–307
- Lipsitz, SR, Fitzmaurice, GM, Ibrahim, JG, Gelber, R, Lipschutz, S (2002) Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* 58:621–630
- Little RJA (1993) Pattern-mixture models for multivariate incomplete data. *J Amer Statist Assoc* 88:125–134
- Little RJA (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika* 81:471–483
- Little RJA (1995) Modeling the drop-out mechanism in repeated-measures studies. *J Amer Statist Assoc* 90:1113–1121
- Little RJA, Rubin DB (2002) *Statistical Analysis With Missing Data*. John Wiley, New York
- Little RJA, Wang Y (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 52:98–111
- Louis T (1982) Finding the observed information matrix when using the EM algorithm. *J Roy Statist Soc Ser B* 44:226–233
- Meilijson I (1989) A fast improvement to the EM algorithm on its own terms. *J Roy Statist Soc Ser B* 51: 127–138
- Molenberghs G, Kenward MG (2007) *Missing Data in Clinical Studies*. John Wiley and Sons, Chichester
- Molenberghs G, Verbeke G (2005) *Models for Discrete Longitudinal Data*. Springer, New York
- Molenberghs G, Kenward MG, Lesaffre E (1997) The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* 84:33–4
- Pawitan Y, Self S (1993) Modeling disease marker processes in AIDS. *J Amer Statist Assoc* 88:719–726
- Prentice RL (1989) Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* 8:431–440
- Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M (2002) Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biom J* 44:921–935

- Rizopoulos D, Verbeke G, Molenberghs G (2008) Shared parameter models under random-effects misspecification. *Biometrika* 94:63–74
- Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Amer Statist Assoc* 90(429):106–121
- Rotnitzky A, Robins JM, Scharfstein DO (1998) Semiparametric regression for repeated outcomes with nonignorable non-response. *J Amer Statist Assoc* 93:1321–1339
- Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592
- Rubin DB (1987) Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley and Sons, New York
- Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Amer Statist Assoc* 94:1096–1120
- Schluchter MD (1992) Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 11:1861–1870
- Shi, X, Zhu, H, Ibrahim, JG (2009) Local influence for generalized linear models with missing covariates. *Biometrics*, in press
- Stubbendick, AL, Ibrahim, JG (2003) Maximum likelihood methods for nonignorable responses and covariates in random effects models. *Biometrics* 59:1140–1150
- Stubbendick, AL, Ibrahim, JG (2006) Likelihood-based inference with nonignorably missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica* 16:1143–1167
- Taylor JMG, Cumberland WG, Sy JP (1994) A stochastic model for analysis of longitudinal AIDS data. *J Amer Statist Assoc* 89:727–736
- Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D (2002) Strategies to fit pattern-mixture models. *Biostatistics* 3:245–265
- Troxel AB, Harrington DP, Lipsitz SR (1998) Analysis of longitudinal data with nonignorable nonmonotone missing values. *Applied Statistics* 47:425–438
- Troxel AB, Lipsitz SR, Harrington DP (1998) Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* 85:661–672
- Tsiatis AA, DeGruttola V, Wulfsohn MS (1995) Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J Amer Statist Assoc* 90:27–37
- Verbeke G, Molenberghs, G (2000) *Linear Mixed Models for Longitudinal Data*. Springer, New York
- Wedderburn RWM (1974) Quasi-likelihood methods, generalised linear models, and the Gauss-Newton method. *Biometrika* 61:439–447
- Wei GC, Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Amer Statist Assoc* 85:699–704
- Wolfinger R, O'Connell M (1993) Generalized linear models: a pseudo-likelihood approach. *J Stat Comput Simul* 48:233–243
- Woolson RF, Clarke WR (1984) Analysis of categorical incomplete longitudinal data. *J Roy Statist Soc Ser A* 147:87–99
- Wu MC, Bailey KR (1988) Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* 7:337–346
- Wu MC, Bailey KR (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 45:939–955
- Wu MC, Carroll RJ (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44:175–188
- Xu J, Zeger SL (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* 50:375–387
- Zeger SL, Liang K-Y (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121–130
- Zhu, H, Shi, Ibrahim, JG, Shi, X (2009) Diagnostic measures for generalized linear models with missing covariates. *Scandinavian Journal of Statistics*, in press
- Zhu H-T, Lee S-Y (2001) Local influence for incomplete-data models. *J Roy Statist Soc Ser B* 63:111–126