Nonlinear Models for Repeated Measurement Data: An Overview and Update

Marie Davidian and David M. Giltinan

Nonlinear mixed effects models for data in the form of continuous, repeated measurements on each of a number of individuals, also known as hierarchical nonlinear models, are a popular platform for analysis when interest focuses on individual-specific characteristics. This framework first enjoyed widespread attention within the statistical research community in the late 1980s, and the 1990s saw vigorous development of new methodological and computational techniques for these models, the emergence of general-purpose software, and broad application of the models in numerous substantive fields. This article presents an overview of the formulation, interpretation, and implementation of nonlinear mixed effects models and surveys recent advances and applications.

Key Words: Hierarchical model; Inter-individual variation; Intra-individual variation; Nonlinear mixed effects model; Random effects; Serial correlation; Subject-specific.

1. INTRODUCTION

A common challenge in biological, agricultural, environmental, and medical applications is to make inference on features underlying profiles of continuous, repeated measurements from a sample of individuals from a population of interest. For example, in pharmacokinetic analysis (Sheiner and Ludden 1992), serial blood samples are collected from each of several subjects following doses of a drug and assayed for drug concentration, and the objective is to characterize pharmacological processes within the body that dictate the time-concentration relationship for individual subjects and the population of subjects. Similar objectives arise in a host of other applications; see Section 2.1.

The nonlinear mixed effects model, also referred to as the hierarchical nonlinear model, has gained broad acceptance as a suitable framework for such problems. Analyses based on this model are now routinely reported across a diverse spectrum of subject-matter literature, and software has become widely available. Extensions and modifications of the model to

Marie Davidian is Professor, Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695 (E-mail: davidian@stat.ncsu.edu). David Giltinan is Staff Scientist, Genentech, Inc., 1 DNA Way South San Francisco, CA 94080-4990 (E-mail: giltinan@gene.com).

handle new substantive challenges continue to emerge. Indeed, since the publication of Davidian and Giltinan (1995) and Vonesh and Chinchilli (1997), two monographs offering detailed accounts of the model and its practical application, much has taken place.

The objective of this article is to provide an updated look at the nonlinear mixed effects model, summarizing from a current perspective its formulation, interpretation, and implementation and surveying new developments. In Section 2, we describe the model and situations for which it is an appropriate framework. Section 3 offers an overview of popular techniques for implementation and associated software. Recent advances and extensions that build on the basic model are reviewed in Section 4. Presentation of a comprehensive bibliography is impossible, as, pleasantly, the literature has become vast. Accordingly, we note only a few early, seminal references and refer the reader to Davidian and Giltinan (1995) and Vonesh and Chinchilli (1997) for a fuller compilation prior to the mid-1990s. The remaining work cited represents what we hope is an informative sample from the extensive modern literature on the methodology and application of nonlinear mixed effects models that will serve as a starting point for readers interested in deeper study of this topic.

2. NONLINEAR MIXED EFFECTS MODEL

2.1 The Setting

To exemplify circumstances for which the nonlinear mixed effects model is an appropriate framework, we review challenges from several diverse applications.

Figure 1 goes here

Pharmacokinetics. Figure 1 shows data typical of an intensive pharmacokinetic study (Davidian and Giltinan 1995, sec. 5.5) carried out early in drug development to gain insight into within-subject pharmacokinetic processes of absorption, distribution, and elimination governing concentrations of drug achieved. Twelve subjects were given the same oral dose (mg/kg) of the anti-asthmatic agent theophylline, and blood samples drawn at several times following administration were assayed for theophylline concentration. As ordinarily observed in this context, the concentration profiles have a similar shape for all subjects; however, peak concentration achieved, rise, and decay vary substantially. These differences are believed to be attributable to inter-subject variation in the underlying pharmacokinetic processes, understanding of which is critical for developing dosing guidelines.

To characterize these processes formally, it is routine represent the body by a simple compartment model (e.g. Sheiner and Ludden 1991). For theophylline, a one-compartment model is standard, and solution of the corresponding differential equations yields

$$C(t) = \frac{Dk_a}{V(k_a - Cl/V)} \left\{ \exp(-k_a t) - \exp\left(-\frac{Cl}{V} t \right) \right\}, \tag{1}$$

where $C(t)$ is drug concentration at time t for a single subject following oral dose D at $t = 0$. Here, k_a is the fractional rate of absorption describing how drug is absorbed from the gut into the bloodstream; V is roughly the volume required to account for all drug in the body, reflecting the extent of drug distribution; and Cl is the clearance rate representing the volume of blood from which drug is eliminated per unit time. Thus, the parameters (k_a, V, Cl) in (1) summarize the pharmacokinetic processes for a given subject.

More precisely stated, the goal is to determine, based on the observed profiles, mean or median values of (k_a, V, Cl) and how they vary in the population of subjects in order to design repeated dosing regimens to maintain drug concentrations in a desired range. If inter-subject variation in (k_a, V, Cl) is large, it may be difficult to design an "all-purpose" regimen; however, if some of the variation is associated with subject characteristics such as age or weight, this might be used to develop strategies tailored for certain subgroups.

Figure 2 goes here

HIV Dynamics. With the advent of assays capable of quantifying the concentration of viral particles in the blood, monitoring of such "viral load" measurements is now a routine feature of care of HIV-infected individuals. Figure 2 shows viral load-time profiles for ten participants in AIDS Clinical Trial Group (ACTG) protocol 315 (Wu and Ding 1999) following initiation of potent antiretroviral therapy. Characterizing mechanisms of virus-immune system interaction that lead to such patterns of viral decay (and eventual rebound for many subjects) enhances understanding of the progression of HIV disease.

Considerable recent interest has focused on representing within-subject mechanisms by a system of differential equations whose parameters characterize rates of production, infection, and death of immune system cells and viral production and clearance (Wu and Ding 1999, sec. 2). Under assumptions discussed by these authors, typical models for $V(t)$, the concentration of virus particles at time t following treatment initiation, are of the form

$$V(t) = P_1 \exp(-\lambda_1 t) + P_2 \exp(-\lambda_2 t), \tag{2}$$

where $(P_1, \lambda_1, P_2, \lambda_2)$ describe patterns of viral production and decay. Understanding the "typical" values of these parameters, how they vary among subjects, and whether they are associated with measures of disease status at initiation of therapy such as baseline viral load (e.g., Notermans et al. 1998) may guide use of drugs in the anti-HIV arsenal. A complication is that viral loads below the lower limit of detection of the assay are not quantifiable and are usually coded as equal to the limit (100 copies/ml for ACTG 315).

Dairy Science. Inflammation of the mammary gland in dairy cattle has serious economic implications for the dairy industry (Rodriguez-Zas, Gianola, and Shook 2000). Milk somatic cell score (SCS) is a measure reflecting udder health during lactation, and elucidating processes underlying its evolution may aid disease management. Rodriguez-Zas et al. (2000) represent time-SCS profiles by nonlinear models whose parameters characterize these processes. Rekaya et al. (2001) describe longitudinal patterns of milk yield by nonlinear models involving quantities such as peak yield, time-to-peak-yield, and persistency (reflecting capacity to maintain milk production after peak yield). Understanding how these parameters vary among cows and are related to cow-specific attributes is valuable for breeding purposes.

Forestry. Forest growth and yield and the impact of silvicultural treatments such as herbicides and fertilizers are often evaluated on the basis of repeated measurements on permanent sample plots. Fang and Bailey (2001) and Hall and Bailey (2001) describe nonlinear models for these measures as a function of time that depend on meaningful parameters such as asymptotic growth or yield and rates of change. Objectives are to understand among-plot/tree variation in these parameters and determine whether some of this variation is associated with site preparation treatments, soil type, and tree density to monitor and predict changes in forest stands with different attributes, and to predict growth and yield for individual plots/trees to aid forest management decisions. Similarly, Gregoire and Schabenberger (1996ab) note that forest inventory practices rely on prediction of the volume of the bole (trunk) of a standing tree to assess the extent of merchantable product available in trees of a particular age using measurements on diameter-at-breast-height D and stem height H. Based on data on D, H, and repeated measurements of cumulative bole volume at three-foot intervals along the bole for several similarly-aged trees, they develop a nonlinear model to be used for prediction whose parameters governing asymptote, growth rate, and shape vary across trees depending on D and H; see also Davidian and Giltinan (1995, sec. 11.3).

Further Applications. The range of fields where nonlinear mixed models are used is vast. We briefly cite a few more applications in biological, agricultural, and environmental sciences.

In toxicokinetics (e.g., Gelman et al. 1996; Mezetti et al. 2003), physiologically-based pharmacokinetic models, complex compartment models including organs and tissue groups, describe concentrations in breath or blood of toxic substances as implicit solutions to a system of differential equations involving meaningful physiological parameters. Knowledge of the parameters aids understanding of mechanisms underlying toxic effects. McRoberts, Brooks, and Rogers (1998) represent size-age relationships for black bears via nonlinear models whose parameters describe features of the growth processes. Morrell et al. (1995), Law, Taylor, and Sandler (2002), and Pauler and Finkelstein (2002) characterize longitudinal prostate-specific antigen trajectories in men by subject-specific nonlinear models; the latter two papers relate underlying features of the profiles to cancer recurrence. Müller and Rosner (1997) describe patterns of white blood cell and granulocyte counts for cancer patients undergoing high-dose chemotherapy by nonlinear models whose parameters characterize important features of the profiles; knowledge of the relationship of these features to patient characteristics aids evaluation of toxicity and prediction of hematologic profiles for future patients. Mikulich et al. (2003) use nonlinear mixed models in analyses of data on circadian rhythms. Additional applications are found in the literature in fisheries science (Pilling, Kirkwood, and Walker 2002) and plant and soil sciences (Schabenberger and Pierce 2001).

Summary. These situations share several features: (i) repeated observations of a continuous response on each of several individuals (e.g., subjects, plots, cows) over time or other condition (e.g., intervals along a tree bole); (ii) variability in the relationship between response and time or other condition across individuals, and (iii) availability of a scientifically-relevant model characterizing individual behavior in terms of meaningful parameters that vary across individuals and dictate variation in patterns of time-response. Objectives are to understand the "typical" behavior of the phenomena represented by the parameters; the extent to which the parameters, and hence these phenomena, vary across individuals; and whether some of the variation is systematically associated with individual attributes. Individual-level prediction may also be of interest. As we now describe, the nonlinear mixed effects model is an appropriate framework within which to formalize these objectives.

2.2 The Model

Basic model. We consider a basic version of the model here and in Section 3; extensions are discussed in Section 4. Let y_{ij} denote the jth measurement of the response, e.g., drug concentration or milk yield, under condition t_{ij}, $j = 1, \ldots, n_i$, and possible additional conditions $\boldsymbol{u}_i$. E.g., for theophylline, t_{ij} is the time associated with the jth drug concentration for subject i following dose $\boldsymbol{u}_i = D_i$ at time zero. In many applications, t_{ij} is time and $\boldsymbol{u}_i$ is empty, and we use the word "time" generically below. We write for brevity $\boldsymbol{x}_{ij} = (t_{ij}, \boldsymbol{u}_i)$, but note dependence on t_{ij} where appropriate. Assume that there may be a vector of characteristics $\boldsymbol{a}_i$ for each individual that do not change with time, e.g., age, weight, or diameter-at-breast-height. Letting $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, it is ordinarily assumed that the triplets $(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{a}_i)$ are independent across i, reflecting the belief that individuals are "unrelated." The usual nonlinear mixed effects model may then be written as a two-stage hierarchy as follows:

Stage 1: Individual-Level Model.
$$y_{ij} = f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_i) + e_{ij}, \quad j = 1, \ldots, n_i. \tag{3}$$

In (3), f is a function governing within-individual behavior, such as (1)–(2), depending on a $(p \times 1)$ vector of parameters $\boldsymbol{\beta}_i$ specific to individual i. For example, in (1), $\boldsymbol{\beta}_i = (k_{ai}, V_i, Cl_i)^T = (\beta_{1i}, \beta_{2i}, \beta_{3i})^T$, where k_{ai}, V_i, and Cl_i are absorption rate, volume, and clearance for subject i. The intra-individual deviations $e_{ij} = y_{ij} - f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_i)$ are assumed to satisfy $E(e_{ij} | \boldsymbol{u}_i, \boldsymbol{\beta}_i) = 0$ for all j; we say more about other properties of the e_{ij} shortly.

Stage 2: Population Model.
$$\boldsymbol{\beta}_i = \boldsymbol{d}(\boldsymbol{a}_i, \boldsymbol{\beta}, \boldsymbol{b}_i), \quad i = 1, \ldots, m, \tag{4}$$

where $\boldsymbol{d}$ is a p-dimensional function depending on an $(r \times 1)$ vector of fixed parameters, or *fixed effects*, $\boldsymbol{\beta}$ and a $(k \times 1)$ vector of *random effects* $\boldsymbol{b}_i$ associated with individual i. Here, (4) characterizes how elements of $\boldsymbol{\beta}_i$ vary among individuals, due both to systematic association with individual attributes in $\boldsymbol{a}_i$ and to "unexplained" variation in the population of individuals, e.g., natural, biological variation, represented by $\boldsymbol{b}_i$. The distribution of the $\boldsymbol{b}_i$ conditional on $\boldsymbol{a}_i$ is usually taken not to depend on $\boldsymbol{a}_i$ (i.e., the $\boldsymbol{b}_i$ are independent of the $\boldsymbol{a}_i$), with $E(\boldsymbol{b}_i|\boldsymbol{a}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \text{var}(\boldsymbol{b}_i) = \boldsymbol{D}$. Here, $\boldsymbol{D}$ is an unstructured covariance matrix that is the same for all i, and $\boldsymbol{D}$ characterizes the magnitude of "unexplained" variation in the elements of $\boldsymbol{\beta}_i$ and associations among them; a standard such assumption is $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$. We discuss this assumption further momentarily.

For instance, if $\boldsymbol{a}_i = (w_i, c_i)^T$, where for subject i w_i is weight (kg) and $c_i = 0$ if creatinine clearance is ≤ 50 ml/min, indicating impaired renal function, and $c_i = 1$ otherwise, then for a pharmacokinetic study under (1), an example of (4), with $\boldsymbol{b}_i = (b_{1i}, b_{2i}, b_{3i})^T$, is

$$k_{ai} = \exp(\beta_1 + b_{1i}), \quad V_i = \exp(\beta_2 + \beta_3 w_i + b_{2i}), \quad Cl_i = \exp(\beta_4 + \beta_5 w_i + \beta_6 c_i + \beta_7 w_i c_i + b_{3i}). \quad (5)$$

Model (5) enforces positivity of the pharmacokinetic parameters for each i. Moreover, if $\boldsymbol{b}_i$ is multivariate normal, k_{ai}, Cl_i, V_i are each lognormally distributed in the population, consistent with the widely-acknowledged phenomenon that these parameters have skewed population distributions. Here, the assumption that the distribution of $\boldsymbol{b}_i$ given $\boldsymbol{a}_i$ does not depend on $\boldsymbol{a}_i$ corresponds to the belief that variation in the parameters "unexplained" by the systematic relationships with w_i and c_i in (5) is the same regardless of weight or renal status, similar to standard assumptions in ordinary regression modeling. For example, if $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, then $\log Cl_i$ is normal with variance $\boldsymbol{D}_{33}$, and thus Cl_i is lognormal with coefficient of variation (CV) $\exp(\boldsymbol{D}_{33}) - 1$, neither of which depends on w_i, c_i. On the other hand, if this variation is thought to be different, the assumption may be relaxed by taking $\boldsymbol{b}_i|\boldsymbol{a}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{a}_i)\}$, where now the covariance matrix depends on $\boldsymbol{a}_i$. E.g., if the parameters are more variable among subjects with normal renal function, one may assume $\boldsymbol{D}(\boldsymbol{a}_i) = \boldsymbol{D}_0(1 - c_i) + \boldsymbol{D}_1 c_i$, so that the covariance matrix depends on $\boldsymbol{a}_i$ through c_i and equals $\boldsymbol{D}_0$ in the subpopulation of individuals with renal impairment and $\boldsymbol{D}_1$ in that of healthy subjects.

In (5), each element of $\boldsymbol{\beta}_i$ is taken to have an associated random effect, reflecting the belief that each component varies nonnegligibly in the population, even after systematic relationships with subject characteristics are taken into account. In some settings, "unexplained" variation in one component of $\boldsymbol{\beta}_i$ may be very small in magnitude relative to that in others. It is common to approximate this by taking this component to have no associated random effect; e.g., in (5), specify instead $V_i = \exp(\beta_2 + \beta_2 w_i)$, which attributes all variation in volumes across subjects to differences in weight. Usually, it is biologically implausible for there to be no "unexplained" variation in the features represented by the parameters, so one must recognize that such a specification is adopted mainly to achieve parsimony and numerical stability in fitting rather than to reflect belief in perfect biological consistency across individuals. Analyses in the literature to determine "whether elements of $\boldsymbol{\beta}_i$ are fixed or random effects" should be interpreted in this spirit.

A common special case of (4) is that of a linear relationship between $\boldsymbol{\beta}_i$ and fixed and random effects as in usual, empirical statistical linear modeling, i.e.,

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \tag{6}$$

where $\boldsymbol{A}_i$ is a design matrix depending on elements of $\boldsymbol{a}_i$, and $\boldsymbol{B}_i$ is a design matrix typically involving only zeros and ones allowing some elements of $\boldsymbol{\beta}_i$ to have no associated random effect. For example, consider the linear alternative to (5) given by

$$k_{ai} = \beta_1 + b_{1i}, \quad V_i = \beta_2 + \beta_3 w_i, \quad Cl_i = \beta_4 + \beta_5 w_i + \beta_6 c_i + \beta_7 w_i c_i + b_{3i} \tag{7}$$

with $\boldsymbol{b}_i = (b_{1i}, b_{3i})^T$, which may be represented as in (6) with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_7)^T$, and

$$\boldsymbol{A}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & w_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & w_i & c_i & w_i c_i \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix};$$

(7) takes V_i to vary negligibly relative to k_{ai} and Cl_i. Alternatively, if $V_i = \beta_2 + \beta_3 w_i + b_{2i}$, so including "unexplained" variation in this parameter above and beyond that explained by weight, $\boldsymbol{b}_i = (b_{1i}, b_{2i}, b_{3i})^T$ and $\boldsymbol{B}_i = \boldsymbol{I}_3$, where $\boldsymbol{I}_q$ is a q-dimensional identity matrix.

If $\boldsymbol{b}_i$ is multivariate normal, a linear specification as in (7) may be unrealistic for applications where population distributions are thought to be skewed, as in pharmacokinetics.

Rather than adopt a nonlinear population model as in (5), a common alternative tactic is to reparameterize the model f. For example, if (1) is represented in terms of parameters $(k_a^*, V^*, Cl^*)^T = (\log k_a, \log V, \log Cl)^T$, then $\boldsymbol{\beta}_i = (k_{ai}^*, V_i^*, Cl_i^*)^T$, and

$$k_{ai}^* = \beta_1 + b_{1i}, \quad V_i^* = \beta_2 + \beta_3 w_i + b_{2i}, \quad Cl_i^* = \beta_4 + \beta_5 w_i + \beta_6 c_i + \beta_7 w_i c_i + b_{3i} \tag{8}$$

is a linear population model in the same spirit as (5).

In summary, specification of the population model (4) is made in accordance with usual considerations for regression modeling and subject-matter knowledge. Taking $\boldsymbol{A}_i = \boldsymbol{I}_p$ in (6) with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ assumes $E(\beta_{\ell i}) = \beta_\ell$ for $\ell = 1, \ldots, p$, which may be done in the case where no individual covariates $\boldsymbol{a}_i$ are available or as a part of a model-building exercise; see Section 3.6. If the data arise according to a design involving fixed factors, e.g., a factorial experiment, $\boldsymbol{A}_i$ may be chosen to reflect this for each component of $\boldsymbol{\beta}_i$ in the usual way.

Within-individual variation. To complete the full nonlinear mixed model, a specification for variation within individuals is required. Considerations underlying this task are often not elucidated in the applied literature, and there is some apparent confusion regarding the notion of "within-individual correlation." Thus, we discuss this feature in some detail, focusing on phenomena taking place within a single individual i. Our discussion focuses on model (3), but the same considerations are relevant for linear modeling.

Figure 3 goes here

A conceptual perspective on within-individual variation discussed by Diggle et al. (2001, Ch. 5) and Verbeke and Molenberghs (2000, sec. 3.3) is depicted in Figure 3 in the context of the one-compartment model (1) for theophylline, where y is concentration and t is time following dose D. According to the individual model (3), $E(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = f(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$, so that f represents what happens "on average" for subject i, shown as the solid line in Figure 3. A finer interpretation of this may be appreciated as follows. Because f may not capture all within-individual physiological processes, or because response may exhibit "local fluctuations" (e.g., drug may not be "perfectly mixed" within the body as the compartment model assumes), when i is observed, the response profile *actually realized* follows the dotted line. In addition, as an assay is used to ascertain the value of the realized response at any time t, measurement error may be introduced, so that the measured values y, represented by the solid symbols, do not exactly equal the realized values. More precisely, then, $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ is the average over all possible realizations of actual concentration trajectory and measurement errors that could arise if subject i is observed. The implication is that $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ may be viewed as the "inherent tendency" for i's responses to evolve over time, and $\boldsymbol{\beta}_i$ is an "inherent characteristic" of i dictating this tendency. Hence, a fundamental principle underlying nonlinear mixed effects modeling is that such "inherent" properties of individuals, rather than actual response realizations, are of central scientific interest.

Thus, the y_{ij} observed by the analyst are each the sum of one realized profile and one set of measurement errors at intermittent time points t_{ij}, formalized by writing (3) as

$$y_{ij} = f(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i) + e_{R,ij} + e_{M,ij}, \tag{9}$$

where e_{ij} has been partitioned into deviations from $f(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ due to the particular realization observed, $e_{R,ij}$, and possible measurement error, $e_{M,ij}$, at each t_{ij}. In (9), the "actual" realized response at t_{ij}, if it could be observed without error, is thus $f(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i) + e_{R,ij}$. We may think of (9) as following from a within-subject stochastic process of the form

$$y_i(t, \boldsymbol{u}_i) = f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i) + e_{R,i}(t, \boldsymbol{u}_i) + e_{M,i}(t, \boldsymbol{u}_i) \tag{10}$$

with $E\{e_{R,i}(t, \boldsymbol{u}_i)|\boldsymbol{u}_i, \boldsymbol{\beta}_i\} = E\{e_{M,i}(t, \boldsymbol{u}_i)|\boldsymbol{u}_i, \boldsymbol{\beta}_i\} = 0$, where $e_{R,i}(t_{ij}, \boldsymbol{u}_i) = e_{R,ij}$, $e_{M,i}(t_{ij}, \boldsymbol{u}_i) = e_{M,ij}$, and hence $E(e_{R,ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = E(e_{M,ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = 0$. The process $e_{M,i}(t, \boldsymbol{u}_i)$ is similar to the "nugget" effect in spatial models. Thus, characterizing within-individual variation corresponds to characterizing autocorrelation and variance functions (conditional on $\boldsymbol{u}_i, \boldsymbol{\beta}_i$) describing the pattern of correlation and variation of realizations of $e_{R,i}(t, \boldsymbol{u}_i)$ and $e_{M,i}(t, \boldsymbol{u}_i)$ and how they combine to produce an overall pattern of intra-individual variation.

First consider $e_{R,i}(t, \boldsymbol{u}_i)$. From Figure 3, two realizations (dotted line) at times close together tend to occur "on the same side" of the "inherent" trajectory, implying that $e_{R,ij}$ and $e_{R,ij'}$ for $t_{ij}, t_{ij'}$ "close" would tend to be positive or negative together. On the other hand, realizations far apart bear little relation to one another. Thus, realizations over time are likely to be positively correlated within an individual, with correlation "damping out" as observations are more separated in time. This suggests models such as the exponential correlation function $\text{corr}\{e_{R,i}(t, \boldsymbol{u}_i), e_{R,i}(s, \boldsymbol{u}_i)|\boldsymbol{u}_i, \boldsymbol{\beta}_i\} = \exp(-\rho|t - s|)$; other popular models are described, for example, by Diggle et al. (2001, Ch. 5). If "fluctuations" in the realization process are believed to be of comparable magnitude over time, then $\text{var}\{e_{R,i}(t,\boldsymbol{u}_i)|\boldsymbol{u}_i,\boldsymbol{\beta}_i\} = \sigma_R^2$ for all t. Alternatively, the variance of the realization process need not be constant over time; e.g., if "actual" responses at any time within an individual are thought to have a skewed distribution with constant CV σ_R, then $\text{var}\{e_{R,i}(t,\boldsymbol{u}_i)|\boldsymbol{u}_i,\boldsymbol{\beta}_i\} = \sigma_R^2 f^2(t,\boldsymbol{u}_i,\boldsymbol{\beta}_i)$. Letting $\boldsymbol{e}_{R,i} = (e_{R,i1},\ldots,e_{R,in_i})^T$, under specific variance and autocorrelation functions, define $\boldsymbol{T}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\delta})$ to be the $(n_i \times n_i)$ diagonal matrix with diagonal elements $\text{var}(e_{R,ij}|\boldsymbol{u}_i,\boldsymbol{\beta}_i)$ depending on parameters $\boldsymbol{\delta}$, say, and $\boldsymbol{\Gamma}_i(\boldsymbol{\rho})$ $(n_i \times n_i)$ with (j,j') elements $\text{corr}(e_{R,ij},e_{R,ij'}|\boldsymbol{u}_i,\boldsymbol{\beta}_i)$ depending on parameters $\boldsymbol{\rho}$, then

$$\text{var}(\boldsymbol{e}_{R,i}|\boldsymbol{u}_i,\boldsymbol{\beta}_i) = \boldsymbol{T}_i^{1/2}(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\delta})\boldsymbol{\Gamma}_i(\boldsymbol{\rho})\boldsymbol{T}_i^{1/2}(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\delta}), \quad (n_i \times n_i), \tag{11}$$

is the covariance matrix for $\boldsymbol{e}_{R,i}$. E.g., with constant CV σ_R and exponential correlation, (11) has (j,j') element $\sigma_R^2 f(t_{ij},\boldsymbol{u}_i,\boldsymbol{\beta}_i)f(t_{ij'},\boldsymbol{u}_i,\boldsymbol{\beta}_i)\exp(-\rho|t_{ij}-t_{ij'}|)$, and $\boldsymbol{\delta} = \sigma_R^2$, $\boldsymbol{\rho} = \rho$.

The process $e_{M,i}(t,\boldsymbol{u}_i)$ characterizes error in measuring "actual" realized responses on i. Often, the error committed by a measuring device at one time is unrelated to that at another; e.g., samples are assayed separately. Thus, it is usually reasonable to assume (conditional) independence of instances of the measurement process, so that $\text{corr}\{e_{M,i}(t,\boldsymbol{u}_i),e_{M,i}(s,\boldsymbol{u}_i)|\boldsymbol{u}_i,\boldsymbol{\beta}_i\} = 0$ for all $t > s$. The form of $\text{var}\{e_{M,i}(t,\boldsymbol{u}_i)|\boldsymbol{u}_i,\boldsymbol{\beta}_i\}$ reflects the nature of error variation; e.g., $\text{var}\{e_{M,i}(t,\boldsymbol{u}_i)|\boldsymbol{u}_i,\boldsymbol{\beta}_i\} = \sigma_M^2$ implies this is constant regardless of the magnitude of the response. Defining $\boldsymbol{e}_{M,i} = (e_{M,i1},\ldots,e_{M,in_i})^T$, the covariance matrix of $\boldsymbol{e}_{M,i}$ would thus ordinarily be diagonal with diagonal elements $\text{var}(e_{M,ij}|\boldsymbol{u}_i,\boldsymbol{\beta}_i)$; i.e.,

$$\text{var}(\boldsymbol{e}_{M,i}|\boldsymbol{u}_i,\boldsymbol{\beta}_i) = \boldsymbol{\Lambda}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\theta}), \quad (n_i \times n_i), \tag{12}$$

a diagonal matrix depending on a parameter $\boldsymbol{\theta}$. (12) may also depend on $\boldsymbol{\beta}_i$; an example is given below. For constant measurement error variance, $\boldsymbol{\Lambda}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\theta}) = \sigma_M^2\boldsymbol{I}_{n_i}$ and $\boldsymbol{\theta} = \sigma_M^2$.

A common assumption (e.g., Diggle et al. 2001, Ch. 5) is that the realization and measurement error processes in (10) are conditionally independent, which implies

$$\text{var}(\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{\beta}_i) = \text{var}(\boldsymbol{e}_{R,i}|\boldsymbol{u}_i,\boldsymbol{\beta}_i) + \text{var}(\boldsymbol{e}_{M,i}|\boldsymbol{u}_i,\boldsymbol{\beta}_i). \tag{13}$$

If such independence were not thought to hold, $\text{var}(\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{\beta}_i)$ would also involve a conditional covariance term. Thus, combining the foregoing considerations and adopting (13) as is customary, a general representation of the components of within-subject variation is

$$\text{var}(\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{\beta}_i) = \boldsymbol{T}_i^{1/2}(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\delta})\boldsymbol{\Gamma}_i(\boldsymbol{\rho})\boldsymbol{T}_i^{1/2}(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\delta}) + \boldsymbol{\Lambda}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\theta}) = \boldsymbol{R}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\xi}), \qquad (14)$$

where $\boldsymbol{\xi} = (\boldsymbol{\delta}^T, \boldsymbol{\rho}^T, \boldsymbol{\theta}^T)^T$ fully describes the overall pattern of within-individual variation.

The representation (14) provides a framework for thinking about sources that contribute to the overall pattern of within-individual variation. It is common in practice to adopt models that are simplifications of (14). For example, measurement error may be nonexistent; if response is age of a plant or tree and recorded exactly, $e_{M,i}(t,\boldsymbol{u}_i)$ may be eliminated from the model altogether, so that (14) only involves $\boldsymbol{T}_i$ and $\boldsymbol{\Gamma}_i$. In many instances, although ideally $e_{R,i}(t,\boldsymbol{u}_i)$ may have non-zero autocorrelation function, the observation times t_{ij} are far enough apart relative to the "locality" of the process that correlation has "damped out" sufficiently between any t_{ij} and $t_{ij'}$ as to be virtually negligible. Thus, for the particular times involved, $\boldsymbol{\Gamma}_i(\boldsymbol{\rho}) = \boldsymbol{I}_{n_i}$ in (11) may be a reasonable approximation.

In still other contexts, measurement error may be taken to be the primary source of variation about f. Karlsson, Beal, and Sheiner (1995) note that this is a common approximation in pharmacokinetics. Although this is generally done without comment, a rationale for this approximation may be deduced from the standpoint of (14). A well-known property of assays used to quantify drug concentration is that errors in measurement are larger in magnitude the larger the (true) concentration being measured. Thus, as the "actual" concentration in a sample at time t_{ij} is $f(t_{ij},\boldsymbol{u}_i,\boldsymbol{\beta}_i)+e_{R,ij}$, measurement errors $e_{M,ij}$ at t_{ij} would be thought to vary as a function this value. This would violate the usual assumption (13) that $e_{M,ij}$ is independent of $e_{R,ij}$. However, if the magnitude of "fluctuations" represented by the $e_{R,ij}$ is "small" relative to errors in measurement, variation in measurement errors at t_{ij} may be thought mainly to be a function of $f(t_{ij},\boldsymbol{u}_i,\boldsymbol{\beta}_i)$. In fact, if $\text{var}(e_{R,ij}|\boldsymbol{u}_i,\boldsymbol{\beta}_i) << \text{var}(e_{M,ij}|\boldsymbol{u}_i,\boldsymbol{\beta}_i)$, one might regard the first term in (14) as negligible. This leads to a standard model in this application, where $\boldsymbol{R}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\xi}) = \boldsymbol{\Lambda}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\theta})$, with $\boldsymbol{\Lambda}_i(\boldsymbol{u}_i,\boldsymbol{\beta}_i,\boldsymbol{\theta})$ a diagonal matrix with diagonal elements $\sigma^2 f^{2\theta}(t_{ij},\boldsymbol{u}_i,\boldsymbol{\beta}_i)$ for some $\boldsymbol{\theta} = (\sigma^2,\theta)^T$; often $\theta = 1$. Karlsson et al. (1995) argue that this approximation, where effectively $e_{R,i}(t,\boldsymbol{u}_i)$ is entirely disregarded, may be optimistic, as serial correlation may be apparent. To demonstrate how considering (14) may lead to a more realistic model, suppose that, although although $\mathrm{var}(e_{R,ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i)$ may be small relative to $\mathrm{var}(e_{M,ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i)$, so that $\boldsymbol{\Lambda}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})$ is reasonably modeled as above, $e_{R,i}(t, \boldsymbol{u}_i)$ is not disregarded. As drug concentrations are positive, it may be appropriate to assume that realized concentrations have a skewed distribution for each t and take $\boldsymbol{T}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\delta})$ to have diagonal elements $\sigma_R^2 f^2(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$. If $\theta = 1$ and $\boldsymbol{\Gamma}_i(\boldsymbol{\rho})$ is a correlation matrix, one is led to $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$ with diagonal elements $(\sigma_R^2 + \sigma_M^2) f^2(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ and (j, j') off-diagonal elements $\sigma_R^2 f(t_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}_i) f(t_{ij'}, \boldsymbol{u}_i, \boldsymbol{\beta}_i) \boldsymbol{\Gamma}_{i,jj'}(\boldsymbol{\rho})$, similar to models studied by these authors.

In summary, in specifying a model for $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$ in (14), the analyst must be guided by subject-matter and practical considerations. As discussed below, accurate characterization of $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$ may be less critical when among-individual variation is dominant.

In this formulation of $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$, the parameters $\boldsymbol{\xi}$ are common to all individuals. Just as $\boldsymbol{\beta}_i$ vary across individuals, in some contexts, parameters of the processes in (10) may be thought to be individual-specific. For instance, if the same measuring device is used for all subjects, $\mathrm{var}\{e_{M,i}(t, \boldsymbol{u}_i)|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = \sigma_M^2$ for all i is plausible; however, "fluctuations" may differ in magnitude for different subjects, suggesting, e.g., $\mathrm{var}\{e_{R,i}(t, \boldsymbol{u}_i)|\boldsymbol{u}_i, \boldsymbol{\beta}_i\} = \sigma_{R,i}^2$, and it is natural to think of $\sigma_{R,i}^2$ as depending on covariates and random effects, as in (4). Such specifications are possible although more difficult (see Zeng and Davidian 1997 for an example), but have not been widely used. If inter-individual variation in these parameters is not large, postulating a common parameter may be a reasonable approximation.

The foregoing discussion tacitly assumes that $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ is a correct specification of "inherent" individual behavior, with actual realizations fluctuating about $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$. Thus, deviations arising from $e_{R,i}(t, \boldsymbol{u}_i)$ in (10) are regarded as part of within-individual variation and hence not of direct interest. An alternative view is that a component of these deviations is due to misspecification of $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$. E.g., in pharmacokinetics, compartment models used to characterize the body are admittedly simplistic, so systematic deviations from $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ due to their failure to capture the true "inherent" trajectory are possible. If, for example, other "local" variation is negligible, the process $e_{R,i}(t, \boldsymbol{u}_i)$ in (10) in fact reflects entirely such misspecification, and the true, "inherent trajectory" would be $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i) + e_{R,i}(t, \boldsymbol{u}_i)$, so that $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ alone is a biased representation of the true trajectory. Here, $e_{R,i}(t, \boldsymbol{u}_i)$ should be regarded as part of the "signal" rather than as "noise." More generally, under misspecification, part of $e_{R,i}(t, \boldsymbol{u}_i)$ is due to such systematic deviations and part is due to "fluctuations," but without knowledge of the exact nature of misspecification, distinguishing bias from within-individual variation makes within-individual covariance modeling a nearly impossible challenge. In the particular context of nonlinear mixed effects models, there are also obvious implications for the meaning and relevance of $\boldsymbol{\beta}_i$ under a misspecified model. We do not pursue this further; however, it is vital to recognize that published applications of nonlinear mixed effects models are almost always predicated on correctness of $f(t, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$.

Summary. We are now in a position to summarize the basic nonlinear mixed effects model. Let $\boldsymbol{f}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i) = \{f(\boldsymbol{x}_{i1}, \boldsymbol{\beta}_i), \ldots, f(\boldsymbol{x}_{in_i}, \boldsymbol{\beta}_i)\}^T$, and let $\boldsymbol{z}_i = (\boldsymbol{u}_i^T, \boldsymbol{a}_i^T)^T$ summarize all covariate information on subject i. Then, we may write the model in (3) and (4) succinctly as

Stage 1: Individual-Level Model.

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i), \quad \text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i) = \boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi}) = \boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi}). \quad (15)$$

Stage 2: Population Model. $$\boldsymbol{\beta}_i = \boldsymbol{d}(\boldsymbol{a}_i, \boldsymbol{\beta}, \boldsymbol{b}_i), \quad \boldsymbol{b}_i \sim (\boldsymbol{0}, \boldsymbol{D}). \quad (16)$$

In (15), dependence of $\boldsymbol{f}_i$ and $\boldsymbol{R}_i$ on the covariates $\boldsymbol{a}_i$ and fixed and random effects through $\boldsymbol{\beta}_i$ is emphasized. This model represents individual behavior conditional on $\boldsymbol{\beta}_i$ and hence on $\boldsymbol{b}_i$, the random component in (16). In (16), we assume that the distribution of $\boldsymbol{b}_i|\boldsymbol{a}_i$ does not depend on $\boldsymbol{a}_i$, so that all $\boldsymbol{b}_i$ have common distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D}$. We adopt this assumption in the sequel, as it is routine in the literature, but the methods we discuss may be extended if it is relaxed in the manner described in Section 2.2.

"Within-individual correlation." The nonlinear mixed model (15)–(16) implies a model for the marginal mean and covariance matrix of $\boldsymbol{y}_i$ given all covariates $\boldsymbol{z}_i$; i.e, averaged across the population. Letting $F_b(\boldsymbol{b}_i)$ denote the cumulative distribution function of $\boldsymbol{b}_i$, we have

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i) = \int \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\, dF_b(\boldsymbol{b}_i), \quad \text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i) = E\{\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi})|\boldsymbol{z}_i\} + \text{var}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)|\boldsymbol{z}_i\}, \quad (17)$$

where expectation and variance are with respect to the distribution of $\boldsymbol{b}_i$. In (17), $E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ characterizes the "typical" response profile among individuals with covariates $\boldsymbol{z}_i$. In the literature, $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$ is often referred to as the "within-subject covariance matrix;" however, this is misleading. In particular, $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$ involves two terms: $E\{\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi})|\boldsymbol{z}_i\}$, which averages realization and measurement variation that occur *within individuals* across individuals having covariates $\boldsymbol{z}_i$; and $\text{var}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)|\boldsymbol{z}_i\}$, which describes how "inherent trajectories" vary *among individuals* sharing the same $\boldsymbol{z}_i$. Note $E\{\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi})|\boldsymbol{z}_i\}$ is a diagonal matrix only if $\boldsymbol{\Gamma}_i(\boldsymbol{\rho})$ in (14), reflecting correlation due to within-individual realizations, is an identity matrix. However, $\text{var}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)|\boldsymbol{z}_i\}$ has non-zero off-diagonal elements in general due to common dependence of all elements of $\boldsymbol{f}_i$ on $\boldsymbol{b}_i$. Thus, correlation at the *marginal* level is always expected due to variation *among* individuals, while there is correlation from *within*-individual sources only if serial associations among intra-individual realizations are nonnegligible. In general, then, *both* terms contribute to the overall pattern of correlation among responses on the same individual represented in $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$.

Thus, the terms "within-individual covariance" and "within-individual correlation" are better reserved to refer to phenomena associated with the realization process $e_{R,i}(t, \boldsymbol{u}_i)$. We prefer "aggregate correlation" to denote the overall population-averaged pattern of correlation arising from both sources. It is important to recognize that within-individual variance and correlation are relevant even if scope of inference is limited to a given individual only.

As noted by Diggle et al. (2001, Ch. 5) and Verbeke and Molenberghs (2000, sec. 3.3), in many applications, the effect of within-individual serial correlation reflected in the first term of $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$ is dominated by that from among-individual variation in $\text{var}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)|\boldsymbol{z}_i\}$. This explains why many published applications of nonlinear mixed models adopt simple, diagonal models for $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$ that emphasize measurement error. Davidian and Giltinan (1995, secs. 5.2.4 and 11.3), suggest that, here, how one models within-individual correlation, or, in fact, whether one improperly disregards it, may have inconsequential effects on inference. It is the responsibility of the data analyst to evaluate critically the rationale for and consequences of adopting a simplified model in a particular application.

2.3 Inferential Objectives

We now state more precisely routine objectives of analyses based on the nonlinear mixed effects model, discussed at the end of Section 2.1. Implementation is discussed in Section 3.

Understanding the "typical" values of the parameters in f, how they vary across individuals in the population, and whether some of this variation is associated with individual characteristics may be addressed through inference on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{D}$. The components of $\boldsymbol{\beta}$ describe both the "typical values" and the strength of systematic relationships between elements of $\boldsymbol{\beta}_i$ and individual covariates $\boldsymbol{a}_i$. Often, the goal is to deduce an appropriate specification $\boldsymbol{d}$ in (16); i.e., as in ordinary regression modeling, identify a parsimonious functional form involving the elements of $\boldsymbol{a}_i$ for which there is evidence of associations. In most of the applications in Section 2.1, knowledge of which individual characteristics in $\boldsymbol{a}_i$ are "important" in this way has significant practical implications. For example, in pharmacokinetics, understanding whether and to what extent weight, smoking behavior, renal status, etc. are associated with drug clearance may dictate whether and how these factors must be considered in dosing. Thus, an analysis may involve postulating and comparing several such models to arrive at a final specification. Once a final model is selected, inference on $\boldsymbol{D}$ corresponding to the included random effects provides information on the variation among subjects not explained by the available covariates. If such variation is relatively large, it may be difficult to make statements that are generalizable even to particular subgroups with certain covariate configurations. In HIV dynamics, for example, for patients with baseline CD4 count, viral load, and prior treatment history in a specified range, if λ_2 in (2) characterizing long-term viral decay varies considerably, the difficulty of establishing broad treatment recommendations based only on these attributes will be highlighted, indicating the need for further study of the population to identify additional, important attributes.

In many applications, an additional goal is to characterize behavior for specific individuals, so-called "individual-level prediction." In the context of (15)–(16), this involves inference on $\boldsymbol{\beta}_i$ or functions such as $f(t_0, \boldsymbol{u}_i, \boldsymbol{\beta}_i)$ at a particular time t_0. For instance, in pharmacokinetics, there is great interest in the potential for "individualized" dosing regimens based on subject i's own pharmacokinetic processes, characterized by $\boldsymbol{\beta}_i$. Simulated concentration profiles based on $\boldsymbol{\beta}_i$ under different regimens may inform strategies for i that maintain desired levels. Of course, given sufficient data on i, inference on $\boldsymbol{\beta}_i$ may in principle be implemented via standard nonlinear model-fitting techniques using i's data only. However, sufficient data may not be available, particularly for a new patient. The nonlinear mixed model provides a framework that allows "borrowing" of information from similar subjects; see Section 3.6. Even if n_i is large enough to facilitate estimation of $\boldsymbol{\beta}_i$, as i is drawn from a population of subjects, intuition suggests that it may be advantageous to exploit the fact that i may have similar pharmacokinetic behavior to subjects with similar covariates.

2.4 "Subject-Specific" or "Population-Averaged?"

The nonlinear mixed effects model (15)–(16) is a *subject-specific* (SS) model in what is now standard terminology. As discussed by Davidian and Giltinan (1995, sec. 4.4), the distinction between SS and *population averaged* (PA, or marginal) models may not be important for linear mixed effects models, but it is critical under nonlinearity, as we now exhibit. A PA model assumes that interest focuses on parameters that describe, in our notation, the marginal distribution of $\boldsymbol{y}_i$ given covariates $\boldsymbol{z}_i$. From the discussion following (17), if $E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ were modeled *directly* as a function of $\boldsymbol{z}_i$ and a parameter $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ would represent the parameter corresponding to the "typical (average) response profile" among individuals with covariates $\boldsymbol{z}_i$. This is to be contrasted with the meaning of $\boldsymbol{\beta}$ in (16) as the "typical value" of individual-specific parameters $\boldsymbol{\beta}_i$ in the population.

Consider first linear such models. A linear SS model with second stage $\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i$ as in (6), for design matrix $\boldsymbol{A}_i$ depending on $\boldsymbol{a}_i$, and first stage $E(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = \boldsymbol{U}_i\boldsymbol{\beta}_i$, where $\boldsymbol{U}_i$ is a design matrix depending on the t_{ij} and $\boldsymbol{u}_i$, leads to the linear mixed effects model $E(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ for $\boldsymbol{X}_i = \boldsymbol{U}_i\boldsymbol{A}_i$ and $\boldsymbol{Z}_i = \boldsymbol{U}_i\boldsymbol{B}_i$, where $\boldsymbol{X}_i$ thus depends on $\boldsymbol{z}_i$. From (17), this model implies that

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i) = \int (\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)\, dF_b(\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta},$$

as $E(\boldsymbol{b}_i) = \boldsymbol{0}$. Thus, $\boldsymbol{\beta}$ in a linear SS model fully characterizes *both* the "typical value" of $\boldsymbol{\beta}_i$ and the "typical response profile," so that either interpretation is valid. Here, then, postulating the linear SS model is equivalent to postulating a PA model of the form $E(\boldsymbol{y}_i|\boldsymbol{z}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ directly in that both approaches yield the same representation of the marginal mean and hence allow the same interpretation of $\boldsymbol{\beta}$. Consequently, the distinction between SS and PA approaches has not generally been of concern in the literature on linear modeling.

For nonlinear models, however, this is no longer the case. For definiteness, suppose that $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, and consider a SS model of the form in (15) and (16) for some function f nonlinear in $\boldsymbol{\beta}_i$ and hence in $\boldsymbol{b}_i$. Then, from (17), the implied marginal mean is

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i) = \int \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) p(\boldsymbol{b}_i; \boldsymbol{D}) d\boldsymbol{b}_i, \tag{18}$$

where $p(\boldsymbol{b}_i; \boldsymbol{D})$ is the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ density. For nonlinear f such as (1), this integral is clearly intractable, and $E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ is a complicated expression, one that is not even available in a closed form and evidently depends on *both* $\boldsymbol{\beta}$ and $\boldsymbol{D}$ in general. Consequently, if we start with a nonlinear SS model, the implied PA marginal mean model involves both the "typical value" of $\boldsymbol{\beta}_i$ ($\boldsymbol{\beta}$) *and* $\boldsymbol{D}$. Accordingly, $\boldsymbol{\beta}$ does not fully characterize the "typical response profile" and thus cannot enjoy both interpretations. Conversely, if we were to take a PA approach and model the marginal mean directly as a function of $\boldsymbol{z}_i$ and a parameter $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ would indeed have the interpretation of describing the "typical response profile." But it seems unlikely that it could also have the interpretation as the "typical value" of individual-specific parameters $\boldsymbol{\beta}_i$ in a SS model; indeed, identifying a corresponding SS model for which the integral in (18) turns out to be exactly the same function of $\boldsymbol{z}_i$ and $\boldsymbol{\beta}$ in (16) and does not depend on $\boldsymbol{D}$ seems an impossible challenge. Thus, for nonlinear models, the interpretation of $\boldsymbol{\beta}$ in SS and PA models cannot be the same in general; see Heagerty (1999) for related discussion. The implication is that the modeling approach must be carefully considered to ensure that the interpretation of $\boldsymbol{\beta}$ coincides with the questions of scientific interest.

In applications like those in Section 2.1, the SS approach and its interpretation are clearly more relevant, as a model to describe individual behavior like (1)–(2) is central to the scientific objectives. The PA approach of modeling $E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ directly, where averaging over the population has already taken place, does not facilitate incorporation of an individual-level model. Moreover, using such a model for population-level behavior is inappropriate, particularly when it is derived from theoretical considerations. E.g., representing the average of time-concentration profiles across subjects by the one-compartment model (1), although perhaps giving an acceptable empirical characterization of the average, does not enjoy meaningful subject-matter interpretation. Even when the "typical concentration profile" $E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ is of interest, Sheiner (2003, personal communication) argues that adopting a SS approach and averaging the subject-level model across the population, as in (17), is preferable, as this exploits the scientific assumptions about individual processes embedded in the model.

General statistical modeling of longitudinal data is often purely empirical in that there is no "scientific" model. Rather, linear or logistic functions are used to approximate relationships between continuous or discrete responses and covariates. The need to take into account (aggregate) correlation among elements of $\boldsymbol{y}_i$ is well-recognized, and both SS and PA models are used. In SS generalized linear mixed effects models, for which there is a large, parallel literature (Breslow and Clayton 1993; Diggle et al. 2001, Ch. 9) within-individual correlation is assumed negligible, and random effects represent (among-individual) correlation and generally do not correspond to "inherent" physical or mechanistic features as in "theoretical" nonlinear mixed models. In PA models, aggregate correlation is modeled directly. Here, for nonlinear such empirical models like the logistic, the above discussion implies that the choice between PA and SS approaches is also critical; the analyst must ensure that the interpretation of the parameters matches the subject-matter objectives (interest in the "typical response profile" versus the "typical" value of individual characteristics).

From a historical perspective, pharmacokineticists were among the first to develop in nonlinear mixed effects modeling in detail; see Sheiner, Rosenberg, and Marathe (1977).

3. IMPLEMENTATION AND INFERENCE

A number of inferential methods for the nonlinear mixed effects model are now in common use. We provide a brief overview, and refer the reader to the cited references for details.

3.1 The Likelihood

As in any statistical model, a natural starting point for inference is maximum likelihood. This is a starting point here because the analytical intractability of likelihood inference has motivated many approaches based on approximations; see Sections 3.2 and 3.3. Likelihood is also a fundamental component of Bayesian inference, discussed in Section 3.5.

The individual model (15) along with an assumption on the distribution of $\boldsymbol{y}_i$ given $(\boldsymbol{z}_i, \boldsymbol{b}_i)$

yields a conditional density $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$, say; the ubiquitous choice is the normal. Under the popular (although not always relevant) assumption that $\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi})$ is diagonal, the density may be written as the product of m contributions $p(y_{ij}|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$. Under this condition, the lognormal has also been used. At Stage 2, (16), adopting independence of $\boldsymbol{b}_i$ and $\boldsymbol{a}_i$, one assumes a k-variate density $p(\boldsymbol{b}_i; \boldsymbol{D})$ for $\boldsymbol{b}_i$. As with other mixed models, normality is standard. With these specifications, the joint density of $(\boldsymbol{y}_i, \boldsymbol{b}_i)$ given $\boldsymbol{z}_i$ is

$$p(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{z}_i, ; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) = p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})p(\boldsymbol{b}_i; \boldsymbol{D}). \tag{19}$$

(If the distribution of $\boldsymbol{b}_i$ given $\boldsymbol{a}_i$ is taken to depend on $\boldsymbol{a}_i$, one would substitute for $p(\boldsymbol{b}_i; \boldsymbol{D})$ in (19) the assumed k-variate density $p(\boldsymbol{b}_i|\boldsymbol{a}_i)$, which may depend on different parameters for different values of $\boldsymbol{a}_i$.) A likelihood for $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ may be based on the joint density of the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ given $\boldsymbol{z}_i$,

$$\prod_{i=1}^{m} \int p(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{z}_i, ; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})\, d\boldsymbol{b}_i = \prod_{i=1}^{m} \int p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i \tag{20}$$

by independence across i. Nonlinearity means that the m k-dimensional integrations in (20) generally cannot be done in a closed form; thus, iterative algorithms to maximize (20) in $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ require a way to handle these integrals. Although numerical techniques for evaluation of an integral are available, these can be computationally expensive when performed at each internal iteration of the algorithm. Hence, many approaches to fitting (15)–(16) are instead predicated on analytical approximations.

3.2 Methods Based on Individual Estimates

If the n_i are sufficiently large, an intuitive approach is to "summarize" the responses $\boldsymbol{y}_i$ for each i through individual-specific estimates $\widehat{\boldsymbol{\beta}}_i$ and then use these as the basis for inference on $\boldsymbol{\beta}$ and $\boldsymbol{D}$. In particular, viewing the conditional moments in (15) as functions of $\boldsymbol{\beta}_i$, i.e., $E(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_i)$, $\text{var}(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\beta}_i) = \boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$, fit the model specified by these moments for each individual. If $\boldsymbol{R}_i(\boldsymbol{u}_i, \boldsymbol{\beta}_i, \boldsymbol{\xi})$ is a matrix of the form $\sigma^2 \boldsymbol{I}_{n_i}$, as in ordinary regression where the y_{ij} are assumed (conditionally) independent, standard nonlinear OLS may be used. More generally, methods incorporating estimation of within-individual variance and correlation parameters $\boldsymbol{\xi}$ are needed so that intra-individual variation is taken into appropriate account, as described by Davidian and Giltinan (1993; 1995, sec. 5.2).

Usual large-sample theory implies that the individual estimators $\widehat{\boldsymbol{\beta}}_i$ are asymptotically normal. Each individual is treated separately, so the theory may be viewed as applying conditionally on $\boldsymbol{\beta}_i$ for each i, yielding $\widehat{\boldsymbol{\beta}}_i|\boldsymbol{u}_i, \boldsymbol{\beta}_i \stackrel{\cdot}{\sim} \mathcal{N}(\boldsymbol{\beta}_i, \boldsymbol{C}_i)$. Because of the nonlinearity of f in $\boldsymbol{\beta}_i$, $\boldsymbol{C}_i$ depends on $\boldsymbol{\beta}_i$ in general, so $\boldsymbol{C}_i$ is replaced by $\widehat{\boldsymbol{C}}_i$ in practice, where $\widehat{\boldsymbol{\beta}}_i$ is substituted. To see how this is exploited for inference on $\boldsymbol{\beta}$ and $\boldsymbol{D}$, consider the linear second-stage model (6); the same developments apply to any general $\boldsymbol{d}$ in (16) (Davidian and Giltinan 1995, sec. 5.3.4). The asymptotic result may be expressed alternatively as

$$\widehat{\boldsymbol{\beta}}_i \approx \boldsymbol{\beta}_i + \boldsymbol{e}_i^* = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i + \boldsymbol{e}_i^*, \quad \boldsymbol{e}_i^*|\boldsymbol{z}_i \stackrel{\cdot}{\sim} \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{C}}_i), \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \tag{21}$$

where $\widehat{\boldsymbol{C}}_i$ is treated as known for each i, so that the $\boldsymbol{e}_i^*$ do not depend on $\boldsymbol{b}_i$. This has the form of a linear mixed effects model with known "error" covariance matrix $\widehat{\boldsymbol{C}}_i$, which suggests using standard techniques for fitting such models to estimate $\boldsymbol{\beta}$ and $\boldsymbol{D}$. Steimer et al. (1984) propose use of the EM algorithm; the algorithm is given by Davidian and Giltinan (1995, sec. 5.3.2) in the case $k = p$ and $\boldsymbol{B}_i = \boldsymbol{I}_p$. Alternatively, it is possible to use linear mixed model software such as SAS `proc mixed` (Littell et al. 1996) or the Splus/R function `lme()` (Pinheiro and Bates 2000) to fit (21). Because $\widehat{\boldsymbol{C}}_i$ is known and different for each i, and the software default is to assume $\text{var}(\boldsymbol{e}_i^*|\boldsymbol{z}_i) = \sigma_e^2\boldsymbol{I}_p$, say, this is simplified by "preprocessing" as follows. If $\widehat{\boldsymbol{C}}_i^{-1/2}$ is the Cholesky decomposition of $\widehat{\boldsymbol{C}}_i^{-1}$; i.e., an upper triangular matrix satisfying $\widehat{\boldsymbol{C}}_i^{-1/2\,T}\widehat{\boldsymbol{C}}_i^{-1/2} = \widehat{\boldsymbol{C}}_i^{-1}$, then $\widehat{\boldsymbol{C}}_i^{-1/2}\widehat{\boldsymbol{C}}_i\widehat{\boldsymbol{C}}_i^{-1/2\,T} = \boldsymbol{I}_p$, so that $\widehat{\boldsymbol{C}}_i^{-1/2}\boldsymbol{e}_i^*$ has identity covariance matrix. Premultiplying (21) by $\widehat{\boldsymbol{C}}_i^{-1/2}$ yields the "new" linear mixed model

$$\widehat{\boldsymbol{C}}_i^{-1/2}\widehat{\boldsymbol{\beta}}_i \approx (\widehat{\boldsymbol{C}}_i^{-1/2}\boldsymbol{A}_i)\boldsymbol{\beta} + (\widehat{\boldsymbol{C}}_i^{-1/2}\boldsymbol{B}_i)\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p). \tag{22}$$

Fitting (22) to the "data" $\widehat{\boldsymbol{C}}_i^{-1/2}\widehat{\boldsymbol{\beta}}_i$ with "design matrices" $\widehat{\boldsymbol{C}}_i^{-1/2}\boldsymbol{A}_i$ and $\widehat{\boldsymbol{C}}_i^{-1/2}\boldsymbol{B}_i$ and the constraint $\sigma_e^2 = 1$ is then straightforward. For either this or the EM algorithm, valid approximate standard errors are obtained by treating (21) as exact.

A practical drawback of these methods is that no general-purpose software is available. Splus/R programs implementing both approaches that require some intervention by the user are available at `http://www.stat.ncsu.edu/~st762_info/`.

If the $\widehat{\boldsymbol{\beta}}_i$ are viewed roughly as conditional "sufficient statistics" for the $\boldsymbol{\beta}_i$, this approach may be thought of as approximating (20) with a change of variables to $\boldsymbol{\beta}_i$ by replacing

$p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$ by $p(\widehat{\boldsymbol{\beta}}_i|\boldsymbol{u}_i, \boldsymbol{\beta}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$. In point of fact, as the asymptotic result is not predicated on normality of $\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\beta}_i$, and because the estimating equations for $\boldsymbol{\beta}, \boldsymbol{D}$ solved by normal-based linear mixed model software lead to consistent estimators even if normality does not hold, this approach does not require normality to achieve valid inference. However, the n_i must be large enough for the asymptotic approximation to be justified.

3.3 Methods Based on Approximation of the Likelihood

This last point is critical when the n_i are not large. E.g., in population pharmacokinetics, sparse, haphazard drug concentrations over intervals of repeated dosing are collected on a large number of subjects along with numerous covariates $\boldsymbol{a}_i$. Although this provides rich information for building population models $\boldsymbol{d}$, there are insufficient data to fit the pharmacokinetic model f to any one subject (nor to assess its suitability). Implementation of (20) in principle imposes no requirements on the magnitude of the n_i. Thus, an attractive tactic is instead to approximate (20) in a way that avoids intractable integration. In particular, for each i, an approximation to $p(\boldsymbol{y}_i|\boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i$ is obtained.

First order methods. An approach usually attributed to Beal and Sheiner (1982) is motivated by letting $\boldsymbol{R}_i^{1/2}$ be the Cholesky decomposition of $\boldsymbol{R}_i$ and writing (15)–(16) as

$$\boldsymbol{y}_i = \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) + \boldsymbol{R}_i^{1/2}(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\xi})\boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i|\boldsymbol{z}_i, \boldsymbol{b}_i \sim (\boldsymbol{0}, \boldsymbol{I}_{n_i}). \tag{23}$$

As nonlinearity in $\boldsymbol{b}_i$ causes the difficulty for integration in (20), it is natural to consider a linear approximation. A Taylor series of (23) about $\boldsymbol{b}_i = \boldsymbol{0}$ to linear terms, disregarding the term involving $\boldsymbol{b}_i\boldsymbol{\epsilon}_i$ as "small" and letting $\boldsymbol{Z}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}^*) = \partial/\partial\boldsymbol{b}_i\{\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}|_{b_i=b^*}$ leads to

$$\boldsymbol{y}_i \approx \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0}) + \boldsymbol{Z}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{b}_i + \boldsymbol{R}_i^{1/2}(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0}, \boldsymbol{\xi})\boldsymbol{\epsilon}_i, \tag{24}$$

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i) \approx \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0}), \quad \mathrm{var}(\boldsymbol{y}_i|\boldsymbol{z}_i) \approx \boldsymbol{Z}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0}) + \boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0}, \boldsymbol{\xi}). \tag{25}$$

When $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$ in (20) is a normal density, (24) amounts to approximating it by another normal density whose mean and covariance matrix are linear in and free of $\boldsymbol{b}_i$, respectively. If $p(\boldsymbol{b}_i; \boldsymbol{D})$ is also normal, the integral is analytically calculable analogous to a linear mixed model and yields a n_i-variate normal density $p(\boldsymbol{y}_i|\boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})$ for each i with mean and covariance matrix (25). This suggests the proposal of Beal and Sheiner (1982) to estimate $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ by jointly maximizing $\prod_{i=1}^{m} p(\boldsymbol{y}_i|\boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})$, which is equivalent to maximum likelihood under the assumption the marginal distribution $\boldsymbol{y}_i|\boldsymbol{z}_i$ is normal with moments (25). The advantage is that this "approximate likelihood" is available in a closed form. Standard errors are obtained from the information matrix assuming the approximation is exact.

This approach is known as the `fo` (first-order) method in the package `nonmem` (Boeckmann, Sheiner, and Beal 1992, `http://www.globomax.net/products/nonmem.cfm`) favored by pharmacokineticists. It is also available in SAS `proc nlmixed` (SAS Institute 1999) via the `method=firo` option. As (25) defines an approximate marginal mean and covariance matrix for $\boldsymbol{y}_i$ given $\boldsymbol{z}_i$, an alternative approach is to estimate $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ by solving generalized estimating equations (GEEs) (Diggle et al. 2001, Ch. 8). Here, the mean is not of the "generalized linear model" type and the covariance matrix is not a "working" model but rather an approximation to the true structure dictated by the nonlinear mixed model; however, the GEE approach is broadly applicable to any marginal moment model. Implementation is available in the SAS `nlinmix` macro (Littell et al. 1996, Ch. 12) with the `expand=zero` option. The SAS `nlinmix` macro and `proc nlmixed` are distinct pieces of software. `nlinmix` implements only the first order approximate marginal moment model (25) using the GEE approach and a related first order conditional method, discussed momentarily. In contrast, in addition to implementing (25) via the normal likelihood method above, `nlmixed` offers several, different alternative approaches to "exact" likelihood inference, described shortly.

Vonesh and Chinchilli (1997, Chs. 8,9) and Davidian and Giltinan (1995, sec. 6.2.4) describe the connections between GEEs and the approximate methods discussed in this section. Briefly, because the approximation to $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$ in (25) depends on $\boldsymbol{\beta}$, maximizing the approximate normal likelihood corresponds to what has been called a "GEE2" method, which takes full account of this dependence. The ordinary GEE approach noted above is an example of "GEE1;" here, the dependence of $\text{var}(\boldsymbol{y}_i|\boldsymbol{z}_i)$ on $\boldsymbol{\beta}$ only plays a role in the formation of "weighting matrices" for each data vector $\boldsymbol{y}_i$.

Vonesh and Carter (1992) and Davidian and Giltinan (1995, sec. 6.2.3) discuss further, related methods. An obvious drawback of all first order methods is that the approximation may be poor, as they essentially replace $E(\boldsymbol{y}_i|\boldsymbol{z}_i) = \int f(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i$ by $f(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{0})$. This suggests that more refined approximation would be desirable.

First order conditional methods. As $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})$ and $p(\boldsymbol{b}_i; \boldsymbol{D})$ are ordinarily normal densities, a natural way to approximate integrals like those in (20) is to exploit Laplace's method, a standard technique to approximate an integral of the form $\int e^{-\ell(\boldsymbol{b})}\, d\boldsymbol{b}$ that follows from a Taylor series expansion of $-\ell(\boldsymbol{b})$ about the value $\widehat{\boldsymbol{b}}$, say, maximizing $\ell(\boldsymbol{b})$. Wolfinger and Lin (1997) provide references for this method and a sketch of steps involved in using this approach to approximate (20) in a special case of (15)–(16); see also Wolfinger (1993) and Vonesh (1996). The derivations of these authors assume that the within-individual covariance matrix $\boldsymbol{R}_i$ does not depend on $\boldsymbol{\beta}_i$ and hence $\boldsymbol{b}_i$, which we write as $\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\xi})$. The result is that $p(\boldsymbol{y}_i|\boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})$ may be approximated by a normal density with

$$E(\boldsymbol{y}_i|\boldsymbol{z}_i) \approx \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i) - \boldsymbol{Z}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\widehat{\boldsymbol{b}}_i, \;\; \mathrm{var}(\boldsymbol{y}_i|\boldsymbol{z}_i) \approx \boldsymbol{Z}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i) + \boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\xi}), \quad (26)$$

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\boldsymbol{R}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\xi})\{\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\}, \quad (27)$$

where $\boldsymbol{Z}_i$ is defined as before, and $\widehat{\boldsymbol{b}}_i$ maximizes $\ell(\boldsymbol{b}_i) = \{\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}^T \boldsymbol{R}_i^{-1}(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\xi})\{\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\} + \boldsymbol{b}_i^T\boldsymbol{D}\boldsymbol{b}_i$ in $\boldsymbol{b}_i$. In fact, $\widehat{\boldsymbol{b}}_i$ maximizes in $\boldsymbol{b}_i$ the posterior density for $\boldsymbol{b}_i$

$$p(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) = \frac{p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\xi})p(\boldsymbol{b}_i; \boldsymbol{D})}{p(\boldsymbol{y}_i|\boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})}. \quad (28)$$

Lindstrom and Bates (1990) instead derive (26) by a Taylor series of (23) about $\boldsymbol{b}_i = \widehat{\boldsymbol{b}}_i$.

Equations (26)–(27) suggest an iterative scheme whose essential steps are (i) given current estimates $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{D}}$ and $\widehat{\boldsymbol{b}}_i$, say, update $\widehat{\boldsymbol{b}}_i$ by substituting these in the right hand side of (27); and (ii) holding $\widehat{\boldsymbol{b}}_i$ fixed, update estimation of $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ based on the moments in (26). Software is available implementing variations on this theme. The Splus/R function `nlme()` (Pinheiro and Bates 2000) and the SAS macro `nlinmix` with the `expand=eblup` option carry out step (ii) by a "GEE1" method. The `nonmem` package with the `foce` option instead uses a "GEE2" approach. Additional software packages geared to pharmacokinetic analysis also implement both this and the "first order" approach; e.g., `winnonmix` (`http://www.pharsight.com/products/winnonmix`) and `nlmem` (Galecki 1998). In all cases, standard errors are obtained assuming the approximation is exactly correct.

In principle, the Laplace approximation is valid only if n_i is large. However, Ko and Davidian (2000) note that it should hold if the magnitude of intra-individual variation is small relative to that among individuals, which is the case in many applications, even if n_i are small. When $\boldsymbol{R}_i$ depends on $\boldsymbol{\beta}_i$, the above argument no longer holds, as noted by Vonesh (1996), but Ko and Davidian (2000) argue that is still valid approximately for "small" intra-individual variation. Davidian and Giltinan (1995, sec. 6.3) present a two-step algorithm incorporating dependence of $\boldsymbol{R}_i$ on $\boldsymbol{\beta}_i$. The software packages above all handle this more general case (e.g., Pinheiro and Bates 2000, sec. 5.2). In fact, the derivation of (26) involves an additional approximation in which a "negligible" term is ignored (e.g., Wolfinger and Lin 1997, p. 472; Pinheiro and Bates 2000, p. 317). The `nonmem laplacian` method includes this term and invokes Laplace's method "as-is" in the case $\boldsymbol{R}_i$ depends on $\boldsymbol{b}_i$.

It is well-documented by numerous authors that these "first order conditional" approximations work extremely well in general, even when n_i are not large or the assumptions of normality that dictate the form of (28) on which $\widehat{\boldsymbol{b}}_i$ is based are violated (e.g., Hartford and Davidian 2000). These features and the availability of supported software have made this approach probably the most popular way to implement nonlinear mixed models in practice.

Remarks. The methods in this section may be implemented for any n_i. Although they involve closed-form expressions for $p(\boldsymbol{y}_i|\boldsymbol{z}_i;\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{D})$ and moments (25) and (26), maximization or solution of likelihoods or estimating equations can still be computationally challenging, and selection of suitable starting values for the algorithms is essential. Results from first order methods may also be used as starting values for a more refined "conditional" fit. A common practical strategy is to first fit a simplified version of the model and use the results to suggest starting values for the intended analysis. For instance, one might take $\boldsymbol{D}$ to be a diagonal matrix, which can often speed convergence of the algorithms; this implies the elements of $\boldsymbol{\beta}_i$ are uncorrelated in the population and hence the phenomena they represent are unrelated, which is usually highly unrealistic. The analyst must bear in mind that failure to achieve convergence in general is not valid justification for adopting a model specification that is at odds with features dictated by the science.

3.4 Methods Based on the "Exact" Likelihood

The foregoing methods invoke analytical approximations to the likelihood (20) or first two moments of $p(\boldsymbol{y}_i|\boldsymbol{z}_i;\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{D})$. Alternatively, advances in computational power have made routine implementation of "exact" likelihood inference feasible for practical use, where "exact" refers to techniques where (20) is maximized directly using deterministic or stochastic approximation to handle the integral. In contrast to an analytical approximation as in Section 3.3, whose accuracy depends on the sample size n_i, these approaches can be made as accurate as desired at the expense of greater computational intensity.

When $p(\boldsymbol{b}_i;\boldsymbol{D})$ is a normal density, numerical approximation of the integrals in (20) may be achieved by Gauss-Hermite quadrature. This is a standard deterministic method of approximating an integral by a weighted average of the integrand evaluated at suitably chosen points over a grid, where accuracy increases with the number of grid points. As the integrals over $\boldsymbol{b}_i$ in (20) are k-dimensional, Pinheiro and Bates (1995, sec. 2.4) and Davidian and Gallant (1993) demonstrate how to transform them into a series of one-dimensional integrals, which simplifies computation. As a grid is required in each dimension, the number of evaluations grows quickly with k, increasing the computational burden of maximizing the likelihood with the integrals so represented. Pinheiro and Bates (1995, 2000, Ch. 7) propose an approach they refer to as adaptive Gaussian quadrature; here, the $\boldsymbol{b}_i$ grid is centered around $\widehat{\boldsymbol{b}}_i$ maximizing (28) and scaled in a way that leads to a great reduction in the number of grid points required to achieve suitable accuracy. Use of one grid point in each dimension reduces to a Laplace approximation as in Section 3.3. We refer the reader to these references for details of these methods. Gauss-Hermite and adaptive Gaussian quadrature are implemented in SAS `proc nlmixed` (SAS Institute 1999); the latter is the default method for approximating the integrals, and the former is obtained via `method=gauss noad`.

Although the assumption of normal $\boldsymbol{b}_i$ is commonplace, as in most mixed effects modeling, it may be an unrealistic representation of true "unexplained" population variation. For example, the population may be more prone to individuals with "unusual" parameter values than indicated by the normal. Alternatively, the apparent distribution of the $\boldsymbol{\beta}_i$, even after accounting for systematic relationships, may appear multimodal due to failure to take into account an important covariate. These considerations have led several authors (e.g., Mallet 1986; Davidian and Gallant 1993) to place no or mild assumptions on the distribution of the random effects. The latter authors assume only that the density of $\boldsymbol{b}_i$ is in a "smooth" class that includes the normal but also skewed and multimodal densities. The density represented in (20) by a truncated series expansion, where the degree of truncation controls the flexibility of the representation, and the density is estimated with other model parameters by maximizing (20); see Davidian and Giltinan (1995, sec. 7.3). This approach is implemented in the Fortran program `nlmix` (Davidian and Gallant 1992a), which uses Gauss-Hermite quadrature to do the integrals and requires the user to write problem-specific code. Other authors impose no assumptions and work with the $\boldsymbol{\beta}_i$ directly, estimating their distribution nonparametrically when maximizing the likelihood. Mallet (1986) shows that the resulting estimate is discrete, so that integrations in the likelihood are straightforward. With covariates $\boldsymbol{a}_i$, Mentré and Mallet 1994) consider nonparametric estimation of the joint distribution of $(\boldsymbol{\beta}_i, \boldsymbol{a}_i)$; see Davidian and Giltinan (1995, sec. 7.2). Software for pharmacokinetic analysis implementing this type of approach via an EM algorithm (Schumitzky 1991) is available at `http://www.usc.edu/hsc/lab_apk/software/uscpack.html`. Methods similar in spirit in a Bayesian framework (Section 3.5) have been proposed by Müller and Rosner (1997). Advantages of methods that relax distributional assumptions are potential insight on the structure of the population provided by the estimated density or distribution and more realistic inference on individuals (Section 3.6). Davidian and Gallant (1992) demonstrate how this can be advantageous in the context of selection of covariates for inclusion in $\boldsymbol{d}$.

Other approaches to "exact" likelihood are possible; e.g., Walker (1996) presents an EM algorithm to maximize (20), where the "E-step" is carried out using Monte Carlo integration.

3.5 Methods Based on a Bayesian Formulation

The hierarchical structure of the nonlinear mixed effects model makes it a natural candidate for Bayesian inference. Historically, a key impediment to implementation of Bayesian analyses in complex statistical models was the intractability of the numerous integrations required. However, vigorous development of Markov chain Monte Carlo (MCMC) techniques to facilitate such integration in the early 1990s and new advances in computing power have made such Bayesian analysis feasible. The nonlinear mixed model served as one of the first examples of this capability (e.g., Rosner and Müller 1994; Wakefield et al. 1994). We provide only a brief review of the salient features of Bayesian inference for (15)–(16); see Davidian and Giltinan (1995, Ch. 8) for an introduction in this specific context and Carlin and Louis (2000) for comprehensive general coverage of modern Bayesian analysis.

From the Bayesian perspective, $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ and $\boldsymbol{b}_i$, $i = 1, \ldots, m$, are all regarded as random vectors on an equal footing. Placing the model within a Bayesian framework requires specification of distributions for (15) and (16) and adoption of a third "hyperprior" stage

Stage 3: Hyperprior. $$(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) \sim p(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}). \tag{29}$$

The hyperprior distribution is usually chosen to reflect weak prior knowledge, and typically $p(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) = p(\boldsymbol{\beta})p(\boldsymbol{\xi})p(\boldsymbol{D})$. Given such a full model (15), (16), and (29), Bayesian analysis proceeds by identifying the posterior distributions induced; i.e., the marginal distributions of $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ and the $\boldsymbol{b}_i$ or $\boldsymbol{\beta}_i$ given the observed data, upon which inference is based. Here, because the $\boldsymbol{b}_i$ are treated as "parameters," they are ordinarily not "integrated out" as in the foregoing frequentist approaches. Rather, writing the observed data as $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T$, and defining $\boldsymbol{b}$ and $\boldsymbol{z}$ similarly, the joint posterior density of $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}, \boldsymbol{b})$ is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}, \boldsymbol{b} | \boldsymbol{y}, \boldsymbol{z}) = \frac{\prod_{i=1}^m p(\boldsymbol{y}_i, \boldsymbol{b}_i | \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}) p(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})}{p(\boldsymbol{y} | \boldsymbol{z})}, \tag{30}$$

where $p(\boldsymbol{y}_i, \boldsymbol{b}_i | \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D})$ is given in (19), and the denominator follows from integration of the numerator with respect to $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}, \boldsymbol{b}$. The marginals are then obtained by integration of (30); e.g., the posterior for $\boldsymbol{\beta}$ is $p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{z})$, and an "estimate" of $\boldsymbol{\beta}$ is the mean or mode, with uncertainty measured by spread of $p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{z})$. The integration involved is a daunting analytical task (see Davidian and Giltinan 1995, p. 220).

MCMC techniques yield simulated samples from the relevant posterior distributions, from which any desired feature, such as the mode, may then be approximated. Because of the nonlinearity of f (and possibly $\boldsymbol{d}$) in $\boldsymbol{b}_i$, generation of such simulations is more complex than in simpler linear hierarchical models and must be tailored to the specific problem in many instances (e.g. Wakefield 1996; Carlin and Louis 2000, sec. 7.3). This complicates implementation via all-purpose software for Bayesian analysis such as `WinBUGS` (`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`). For pharmacokinetic analysis, where certain compartment models are standard, a `WinBUGS` interface, `PKBugs` (`http://www.med.ic.ac.uk/divisions/60/pkbugs_web/home.html`), is available.

It is beyond our scope to provide a full account of Bayesian inference and MCMC implementation. The work cited above and Wakefield (1996), Müller and Rosner (1997), and Rekaya et al. (2001) are only a few examples of detailed demonstrations in the context of specific applications. With weak hyperprior specifications, inferences obtained by Bayesian and frequentist methods agree in most instances; e.g., results of Davidian and Gallant (1992) and Wakefield (1996) for a pharmacokinetic application are remarkably consistent.

A feature of the Bayesian framework that is particularly attractive when a "scientific" model is the focus is that it provides a natural mechanism for incorporating known constraints on values of model parameters and other subject-matter knowledge through the specification of suitable proper prior distributions. Gelman et al. (1996) demonstrate this capability in the context of toxicokinetic modeling.

3.6 Individual Inference

As discussed in Section 2.3, elucidation of individual characteristics may be of interest. Whether from a frequentist or Bayesian standpoint, the nonlinear mixed model assumes that individuals are drawn from a population and thus share common features. The resulting phenomenon of "borrowing strength" across individuals to inform inference on a randomly-chosen such individual is often exhibited for the normal linear mixed effects model (f linear in $\boldsymbol{b}_i$) by showing that $E(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i)$ can be written as linear combination of population- and individual-level quantities (Davidian and Giltinan 1995, sec. 3.3; Carlin and Louis 2000, sec. 3.3). The spirit of this result carries over to general models and suggests using the posterior distribution of the $\boldsymbol{b}_i$ or $\boldsymbol{\beta}_i$ for this purpose.

In particular, such posterior distributions are a by-product of MCMC implementation of Bayesian inference for the nonlinear mixed model, so are immediately available. Alternatively, from a frequentist standpoint, an analogous approach is to base inference on $\boldsymbol{b}_i$ on the mode or mean of the posterior distribution (28), where now $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{D}$ are regarded as fixed. As these parameters are unknown, it is natural to substitute estimates for them in (28). This leads to what is known as empirical Bayes inference (e.g., Carlin and Louis 2000, Ch. 3). Accordingly, $\widehat{\boldsymbol{b}}_i$ in (27) are often referred to as empirical Bayes "estimates." For a general second stage model (16), such "estimates" for $\boldsymbol{\beta}_i$ are then obtained as $\widehat{\boldsymbol{\beta}}_i = \boldsymbol{d}(\boldsymbol{a}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}_i)$.

In both frequentist and Bayesian implementations, "estimates" of the $\boldsymbol{b}_i$ are often exploited in an *ad hoc* fashion to assist with identification of an appropriate second-stage model $\boldsymbol{d}$. Specifically, a common tactic is to fit an initial model in which no covariates $\boldsymbol{a}_i$ are included, such as $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$; obtain Bayes or empirical Bayes "estimates" $\widehat{\boldsymbol{b}}_i$; and plot the components of $\widehat{\boldsymbol{b}}_i$ against each element of $\boldsymbol{a}_i$. Apparent systematic patterns are taken to reflect the need to include that element of $\boldsymbol{a}_i$ in $\boldsymbol{d}$ and also suggest a possible functional form for this dependence. Davidian and Gallant (1992) and Wakefield (1996) demonstrate this approach in a specific application. Mandema, Verotta, and Sheiner (1992) use generalized additive models to aid interpretation. Of course, such graphical techniques may be supplemented by standard model selection techniques; e.g., likelihood ratio tests to distinguish among nested such models or inspection of information criteria.

3.7 Summary

With a plethora of methods available for implementation, the analyst has a range of options for nonlinear mixed model analysis. With sparse individual data (n_i "small"), the choice is limited to the approaches in Sections 3.3, 3.4, and 3.5. First order conditional methods yield reliable inferences for both rich (n_i "large") and sparse data situations; this is in contrast to their performance when applied to generalized linear mixed models for binary data, where they can lead to unacceptable biases. Here, we can recommend them for most practical applications. "Exact" likelihood and Bayesian methods in require more sophistication and commitment on the part of the user. If rich intra-individual data are available for all i, in our experience, methods based on individual estimates (Section 3.6), are attractive, both on the basis of performance and the ease with which they are explained to non-statisticians.

Demidenko (1997) has shown that these methods are equivalent asymptotically to first-order conditional methods when m and n_i are large; see also Vonesh (2002).

Many of the issues that arise for linear mixed models carry over, at least approximately, to the nonlinear case. For small samples, estimation of $\boldsymbol{D}$ and $\boldsymbol{\xi}$ may be poor, leading to concern over the impact on estimation of $\boldsymbol{\beta}$. Lindstrom and Bates (1990) and Pinheiro and Bates (2000, sec. 7.2.1) propose approximate "restricted maximum likelihood" estimation of these parameters in the context of first order conditional methods. Moreover, the reliability of standard errors for estimators for $\boldsymbol{\beta}$ may be poor in small samples in part due to failure of the approximate formulæ to take adequate account of uncertainty in estimating $\boldsymbol{D}$ and $\boldsymbol{\xi}$. The issue of testing whether all elements of $\boldsymbol{\beta}_i$ should include associated random effects, discussed in Section 2.2, involves the same considerations as those arising for inference on variance components in linear mixed models; e.g., a null hypothesis that a diagonal element of $\boldsymbol{D}$, representing the variance of the corresponding random effect, is equal to zero, is on the boundary of the allowable parameter space for variances. Under these conditions, it is well-known for the linear mixed model that the usual test statistics do not have standard null sampling distributions. E.g., the approximate distribution of the likelihood ratio test statistic is a mixture of chi-squares; see, for example, Verbeke and Molenberghs (2000, sec. 6.3.4). In the nonlinear case, the same issues apply to "exact" or approximate likelihood (under the first order or first order conditional approaches) inference.

4. EXTENSIONS AND RECENT DEVELOPMENTS

The end of the twentieth century and beginning of the twenty-first saw an explosion of research on nonlinear mixed models. We cannot hope to do this vast literature justice, so note only selected highlights. The cited references should be consulted for details.

New approximations and computation. Vonesh et al. (2002) propose a higher-order conditional approximation to the likelihood than that obtained by the Laplace approximation and show that this method can achieve gains in efficiency over first order conditional methods and approach the performance of "exact" maximum likelihood. These authors also establish large-m/large-n_i theoretical properties. Raudenbush et al. (2000) discuss use of a sixth-order

Laplace-type approximation and Clarkson and Zhan (2002) apply so-called spherical-radial integration methods (Monahan and Genz 1997) for deterministic and stochastic approximation of an integral based on a transformation of variables, both in the context of generalized linear mixed models. These methods could also be applied to the likelihood (20).

Multilevel models and inter-occasion variation. In some settings, individuals may be observed longitudinally over more than one "occasion." One example is in pharmacokinetics, where concentration measurements may be taken over several distinct time intervals following different doses. In each interval, covariates such as enzyme levels, weight, or measures of renal function may also be obtained and are thus "time-dependent" in the sense that their values change across intervals for each subject. If pharmacokinetic parameters such as clearance and volume of distribution are associated with such covariates, then it is natural to expect their values to change with changing covariate values. If there are q dosing intervals I_h, say, $h = 1, \ldots, q$, then this may be represented by modifying (16) to allow the individual parameters to change; i.e., write $\boldsymbol{\beta}_{ij} = \boldsymbol{d}(\boldsymbol{a}_{ih}, \boldsymbol{\beta}, \boldsymbol{b}_i)$ to denote the value of the parameters at t_{ij} when $t_{ij} \in I_h$, where $\boldsymbol{a}_{ih}$ gives the values of the covariates during I_h. This assumes that such "inter-occasion" variation is entirely attributable to changes in covariates.

Similarly, Karlsson and Sheiner (1993) note that pharmacokinetic behavior may vary naturally over time. Thus, even without changing covariates, if subjects are observed over several dosing intervals, parameters values may fluctuate. This is accommodated by a second-stage model with nested random effects for individual and interval-within-individual. Again letting $\boldsymbol{\beta}_{ij}$ denote the value of the parameters when $t_{ij} \in I_h$, modify (16) to $\boldsymbol{\beta}_{ij} = \boldsymbol{d}(\boldsymbol{a}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ih})$, where now $\boldsymbol{b}_i$ and $\boldsymbol{b}_{ih}$ are independent with means zero and covariance matrices $\boldsymbol{D}$ and $\boldsymbol{G}$, say. See Pinheiro and Bates (2000 sec. 7.1.2) for a discussion of such multilevel models. Levels of nested random effects are also natural in other settings. Hall and Bailey (2001) and Hall and Clutter (2003) discuss studies in forestry where longitudinal measures of yield or growth may be measured on each tree within a plot. Similarly, Rekaya et al. (2001) consider milk yield data where each cow is observed longitudinally during it first three lactations.

Multivariate response. Often, more than one response measurement may be taken longitudinally on each individual. A key example is again from pharmacology, where both drug concentrations and measures of some physiological response are collected over time on the same individual. The goal is to develop a joint pharmacokinetic/pharmacodynamic model, where a pharmacodynamic model for the relationship between concentration and response is postulated in terms of subject-specific parameters and linked to a pharmacokinetic model for the time-concentration relationship. This yields a version of (15) where responses of each type are "stacked" and depend on random effects corresponding to parameters in each model, which are in turn taken to be correlated in (16). Examples are given by Davidian and Giltinan (1995, sec. 9.5) and Bennett and Wakefield (2001).

Motivated by studies of timber growth and yield in forestry, where multiple such measures are collected, Hall and Clutter (2003) extend the basic model and first order conditional fitting methods to handle both multivariate response and multiple levels of nested effects.

Mismeasured/missing covariates and censored response. As in any statistical modeling context, missing, mismeasured, and censored data may arise. Wang and Davidian (1996) study the implications for inference when the observation times for each individual are recorded incorrectly. Ko and Davidian (2000) develop first order conditional methods applicable when components of $\boldsymbol{a}_i$ are measured with error. An approach to take appropriate account of censored responses due to a lower limit of detection as in the HIV dynamics setting in Section 2.1 is proposed by Wu (2002). Wu also extends the model to handle mismeasured and missing covariates $\boldsymbol{a}_i$. Wu and Wu (2002b) propose a multiple imputation approach to accommodate missing covariates $\boldsymbol{a}_i$.

Semiparametric models. Ke and Wang (2001) propose a generalization of the nonlinear mixed effects model where the model f is allowed to depend on a completely unspecified function of time and elements of $\boldsymbol{\beta}_i$. The authors suggest that the model provides flexibility for accommodating possible model misspecification and may be used as a diagnostic tool for assessing the form of time dependence in a fully parametric nonlinear mixed model. Li et al. (2002) study related methods in the context of pharmacokinetic analysis. Lindstrom (1995) develops methods for nonparametric modeling of longitudinal profiles that involve random effects and may be fitted using standard nonlinear mixed model techniques.

Other topics. Methods for model selection and determination are studied by Vonesh, Chinchilli, and Pu (1996), Dey, Chen, and Chang (1997), and Wu and Wu (2002a). Young, Zerbe, and Hay (1997) propose confidence intervals for ratios of components of the fixed effects $\boldsymbol{\beta}$. Oberg and Davidian (2000) propose methods for estimating a parametric transformation of the response under which the Stage 1 conditional density is normal with constant variance. Concordet and Nunez (2000) and Chu et al. (2001) discuss interesting applications in veterinary science involving calibration and prediction problems. Methods for combining data from several studies where each is represented by a nonlinear mixed model are discussed by Wakefield and Rahman (2000) and Lopes, Müller, and Rosner (2003). Yeap and Davidian (2001) propose "robust" methods for accommodating "outlying" responses within individual or "outlying" individuals. Lai and Shih (2003) develop alternative methods to those of Mentré and Mallet (1994) cited in Section 3.4 that do not require consideration of the joint distribution of $\boldsymbol{\beta}_i$ and $\boldsymbol{a}_i$. For a model of the form (16), the distribution of the $\boldsymbol{b}_i$ is left completely unspecified and is estimated nonparametrically; these authors also derive large-sample properties of the approach.

Pharmaceutical applications. A topic that has generated great recent interest in the pharmaceutical industry is so-called "clinical trial simulation." Here, a hypothetical population is simulated to which the analyst may apply different drug regimens according to different designs to evaluate potential study outcomes. Nonlinear mixed models are at the heart of this enterprise; subjects are simulated from mixed models for pharmacokinetic and pharmacodynamic behavior that incorporate variation due to covariates and "unexplained" sources thought to be present. See, for example, `http://www.pharsight.com/products/trial_simulator`. More generally, nonlinear mixed effects modeling techniques have been advocated for population pharmacokinetic analysis in a guidance issued by the U.S. Food and Drug Administration (`http://www.fda.gov/cder/guidance/1852fnl.pdf`).

5. DISCUSSION

This review of nonlinear mixed effects modeling is of necessity incomplete, as it is beyond the limits of a single article to document fully the extensive literature. We have chosen to focus much of our attention on revisiting the considerations underlying the basic model from an updated standpoint, and we hope that this will offer readers familiar with the topic additional insight and provide those new to the model a foundation for appreciating its rationale and utility. We have not presented an analysis of a specific application; the references cited in Section 2.1 and Clayton et al. (2003) and Yeap et al. (2003) in this issue offer detailed demonstrations of the formulation, implementation, and interpretation of nonlinear mixed models in practice. We look forward to continuing methodological developments for and new applications of this rich class of models in the statistical and subject-matter literature.

ACKNOWLEDGMENTS

This research was partially supported by NIH grants R01-CA085848 and R37-AI031789.

REFERENCES

Beal, S.L., and Sheiner, L.B. (1982), "Estimating Population Pharmacokinetics," *CRC Critical Reviews in Biomedical Engineering*, 8, 195–222.

Bennett, J., and Wakefield, J. (2001), "Errors-in-Variables in Joint Population Pharmacokinetic/Pharmacodynamic Modeling," *Biometrics*, 57, 803–812.

Boeckmann, A.J., Sheiner, L.B., and Beal S.L. (1992), *NONMEM User's Guide, Part V, Introductory Guide*, San Francisco: University of California.

Breslow, N.E., and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

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

Chu, K.K., Wang, N.Y., Stanley, S., and Cohen, N.D. (2001), "Statistical Evaluation of the Regulatory Guidelines for Use of Furosemide in Race Horses," *Biometrics*, 57, 294–301.

Clarkson, D.B., and Zhan, Y.H. (2002), "Using Spherical-Radial Quadrature to Fit Generalized Linear Mixed Effects Models," *Journal of Computational and Graphical Statistics*, 11, 639–659.

Clayton, C.A., T.B. Starr, R.L. Sielken, Jr., R.L. Williams, P.G. Pontal, A.J. Tobia. (2003), 'Using a Non-linear Mixed Effects Model to Characterize Cholinesterase Activity in Rats Exposed to Aldicarb," *Journal of Agricultural, Biological, and Environmental Statistics*, 8, ????-????.

Concordet, D., and Nunez, O.G. (2000), "Calibration for Nonlinear Mixed Effects Models: An Application to the Withdrawal Time Prediction," *Biometrics*, 56, 1040–1046.

Davidian, M., and Gallant, A. R. (1992a). Nlmix: A program for maximum likelihood estimation of the nonlinear mixed effects model with a smooth random effects density. Department of Statistics, North Carolina State University.

Davidian, M., and Gallant, A.R. (1992b), "Smooth Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, With Application to Quinidine," *Journal of Pharmacokinetics and Biopharmaceutics*, 20, 529–556.

Davidian, M., and Gallant, A.R. (1993), "The Nonlinear Mixed Effects Model With a Smooth Random Effects Density," *Biometrika*, 80, 475–488.

Davidian, M., and Giltinan, D.M. (1993), "Some simple methods for estimating intra-individual variability in nonlinear mixed effects models," *Biometrics*, 49, 59–73.

Davidian, M., and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data*, New York: Chapman and Hall.

Demidenko, E. (1997), "Asymptotic Properties of Nonlinear Mixed Effects Models," in *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, eds., T.G. Gregoire, D.R., Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren, and R.D. Wolfinger, New York: Springer.

Dey, D.K., Chen, M.H., and Chang, H. (1997), "Bayesian Approach for Nonlinear Random Effects Models," *Biometrics*, 53, 1239–1252.

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2001), *Analysis of Longitudinal Data, Second Edition*, Oxford: Oxford University Press.

Hall, D.B., and Bailey, R.L. (2001), "Modeling and Prediction of Forest Growth Variables Based on Multilevel Nonlinear Mixed Models," *Forest Science*, 47, 311–321.

Hall, D.B., and Clutter, M. (2003), "Multivariate multilevel nonlinear mixed effects models for timber yield predictions." *Biometrics*, in press.

Hartford, A., and Davidian, M. (2000), "Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models," *Computational Statistics and Data Analysis*, 34, 139–164.

Heagerty, P. (1999), "Marginally Specified Logistic-Normal Models for Longitudinal Binary Data," *Biometrics*, 55, 688–698.

Fang, Z. and Bailey, R.L. (2001), "Nonlinear Mixed Effects Modeling for Slash Pine Dominant Height Growth Following Intensive Silvicultural Treatments," *Forest Science*, 47, 287–300.

Galecki, A.T. (1998), "NLMEM: A New SAS/IML Macro for Hierarchical Nonlinear Models," *Computer Methods and Programs in Biomedicine*, 55, 207–216.

Gelman, A., Bois, F., and Jiang, L.M. (1996), "Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions,"*Journal of the American Statistical Association*, 91, 1400–1412.

Gregoire, T.G., and Schabenberger, O. (1996a), "Nonlinear Mixed-Effects Modeling of Cumulative Bole Volume With Spatially-Correlated Within-Tree Data," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 107–119.

Gregoire, T.G., and Schabenberger, O. (1996b), "A Non-Linear Mixed-Effects Model to Predict Cumulative Bole Volume of Standing Trees," *Journal of Applied Statistics*, 23, 257–271.

Karlsson, M.O., Beal, S.L., and Sheiner, L.B. (1995), "Three New Residual Error Models for Population PK/PD Analyses," *Journal of Pharmacokinetics and Biopharmaceutics*, 23, 651–672.

Karlsson, M.O., and Sheiner, L.B. (1993), "The Importance of Modeling Inter-Occasion Variability in Population Pharmacokinetic Analyses," *Journal of Pharmacokinetics and Biopharmaceutics*, 21, 735–750.

Ke, C. and Wang, Y. (2001), "Semiparametric Nonlinear Mixed Models and Their Applications," *Journal of the American Statistical Association*, 96, 1272–1298.

Ko, H.J., and Davidian, M. (2000), "Correcting for Measurement Error in Individual-Level Covariates in Nonlinear Mixed Effects Models," *Biometrics*, 56, 368–375.

Lai, T.L., and Shih, M.-C. (2003), "Nonparametric Estimation in Nonlinear Mixed Effects Models," *Biometrika*, 90, 1–13.

Law, N.J., Taylor, J.M.G., and Sandler, H. (2002), “The Joint Modeling of a Longitudinal Disease Progression Marker and the Failure Time Process in the Presence of a Cure,” *Biostatistics*, 3, 547–563.

Li, L., Brown, M.B., Lee, K.H., and Gupta, S. (2002), “Estimation and Inference for a Spline-Enhanced Population Pharmacokinetic Model,” *Biometrics*, 58, 601–611.

Lindstrom, M.J. (1995), “Self-Modeling With Random Shift and Scale Parameters and a Free-Knot Spline Shape Function,” *Statistics in Medicine*, 14, 2009–2021.

Lindstrom, M.J., and Bates, D.M. (1990), “Nonlinear Mixed Effects Models for Repeated Measures Data,” *Biometrics*, 46, 673–687.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996), *SAS System for Mixed Models*, Cary NC: SAS Institute Inc.

Lopes, H.F., Müller, P., and Rosner, G.L. (2003), “Bayesian meta-analysis for longitudinal data models using multivariate mixture priors, *Biometrics*, 59, 66–75.

Mallet, A. (1986), “A Maximum Likelihood Estimation Method for Random Coefficient Regression Models,” *Biometrika*, 73, 645–656.

Mandema, J.W., Verotta, D., and Sheiner, L.B., (1992), “Building Population Pharmacokinetic/Pharmacodynamic Models,” *Journal of Pharmacokinetics and Biopharmaceutics*, 20, 511–529.

McRoberts, R.E., Brooks, R.T., and Rogers, L.L. (1998), “Using Nonlinear Mixed Effects Models to Estimate Size-Age Relationships for Black Bears,” *Canadian Journal of Zoology*, 76, 1098–1106.

Mentré, F., and Mallet, A. (1994), “Handling Covariates in Population Pharmacokinetics,” *International Journal of Biomedical Computing*, 36, 25–33.

Mezzetti, M., Ibrahim, J,.G., Bois, F.Y., Ryan, L.M., Ngo, L., and Smith, T.J. (2003), “A Bayesian Compartmental Model for the Evaluation of 1,3-Butadiene Metabolism,” *Applied Statistics*, 52, 291–305.

Mikulich, S.K., Zerbe, G.O., Jones, R.H., and Crowley, T.J. (2003), “Comparing Linear and Nonlinear Mixed Model Approaches to Cosinor Analysis,” *Statistics in Medicine*, 22, 3195–3211.

Monahan, J., and Genz, A. (1997), “Spherical-Radial Integration Rules for Bayesian Computation,” *Journal of the American Statistical Association*, 92, 664–674.

Morrell, C.H., Pearson, J.D., Carter, H.B., and Brant, L.J. (1995), “Estimating Unknown Transition Times Using a Piecewise Nonlinear Mixed-Effects Model in Men With Prostate Cancer,” *Journal of the American Statistical Association*, 90, 45–53.

Müller, P., and Rosner, G.L. (1997), “A Bayesian Population Model With Hierarchical Mixture Priors Applied to Blood Count Data,” *Journal of the American Statistical Association*, 92, 1279–1292.

Notermans, D.W., Goudsmit, J., Danner, S.A., de Wolf, F., Perelson, A.S., and Mittler, J. (1998). Rate of HIV-1 decline following antiretroviral therapy is related to viral load at baseline and drug regimen. *AIDS* **12**, 1483–1490.

Oberg, A., and Davidian, M. (2000), “Estimating Data Transformations in Nonlinear Mixed Effects Models,” *Biometrics*, 56, 65–72.

Pauler, D. and Finkelstein, D. (2002), “Predicting Time to Prostate Cancer Recurrence Based on Joint Models for Non-linear Longitudinal Biomarkers and Event Time,” *Statistics in Medicine*, 21, 3897–3911.

Pilling, G.M., Kirkwood, G.P., and Walker, S.G. (2002), “An Improved Method for Estimating Individual Growth Variability in Fish, and the Correlation Between von Bertalanffy Growth Parameters,” *Canadian Journal of Fisheries and Aquatic Sciences*, 59, 424–432.

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

Pinheiro, J.C., and Bates, D.M. (2000), *Mixed-Effects Models in S and Splus*, New York: Springer.

SAS Institute (1999), *PROC NLMIXED, SAS OnlineDoc, Version 8*, Cary, NC: SAS Institute Inc.

Schumitzky, A. (1991), "Nonparametric EM Algorithms for Estimating Prior Distributions," *Applied Mathematics and Computation*, 45, 143–157

Steimer, J.L., Mallet, A., Golmard, J.L., and Boisvieux, J.F. (1984), "Alternative Approaches to Estimation of Population Pharmacokinetic Parameters: Comparison with the Nonlinear Mixed Effect Model," *Drug Metabolism Reviews*, 15, 265–292.

Raudenbush, S.W., Yang, M.L., and Yosef, M. (2000), "Maximum Likelihood for Generalized Linear Models With Nested Random Effects Via High-Order, Multivariate Laplace Approximation," *Journal of Computational and Graphical Statistics*, 9, 141–157.

Rekaya, R., Weigel, K.A., and Gianola, D. (2001), "Hierarchical Nonlinear Model for Persistency of Milk Yield in the First Three Lactations of Holsteins," *Lifestock Production Science*, 68, 181–187.

Rodriguez-Zas, S.L., Gianola, D., and Shook, G.E. (2000), "Evaluation of models for somatic cell score lactation patterns in Holsteins," *Lifestock Production Science*, 67, 19–30.

Rosner, G.L, and Müller, P. (1994), "Pharmacokinetic/Pharmacodynamic Analysis of Hematologic Profiles," *Journal of of Pharmacokinetics and Biopharmaceutics*, 22, 499–524.

Schabenberger, O., and Pierce, F.J. (2002), *Contemporary Statistical Models for the Plant and Soil Sciences*, New York: CRC Press.

Sheiner, L.B., and Ludden, T.M. (1992), "Population pharmacokinetics/pharmacodynamics," *Annual Review of Pharmacological Toxicology*, 32, 185–209.

Sheiner, L.B., Rosenberg, B., and Marathe, V.V. (1977), "Estimation of Population Characteristics of Population Pharmacokinetic Parameters From Routine Clinical Data," *Journal of Pharmacokinetics and Biopharmaceutics*, 8, 635–651.

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

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

Vonesh, E.F. (1992), "Mixed-Effects Nonlinear Regression for Unbalanced Repeated Measures," *Biometrics*, 48, 1–17.

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

Vonesh, E.F., Chinchilli, V.M., and Pu, K.W. (1996), "Goodness-Of-Fit in Generalized Nonlinear Mixed-Effects Models," *Biometrics*, 52, 572–587.

Vonesh, E.G., Wang, H., Nie, L., and Majumdar, D. (2002), "Conditional Second-Order Generalized Estimating Equations for Generalized Linear and Nonlinear Mixed-Effects Models," *Journal of the American Statistical Association*, 97, 271–283.

Wakefield, J. (1996), "The Bayesian Analysis of Population Pharmacokinetic Models," *Journal of the American Statistical Association*, 91, 62–75.

Wakefield, J. and Rahman, N. (2000), "The Combination of Population Pharmacokinetic Studies," *Biometrics*, 56, 263–270.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., and Gelfand, A.E. (1994), "Bayesian Analysis of Linear and Nonlinear Population Models by Using the Gibbs Sampler," *Applied Statistics*, 43, 201–221.

Walker, S.G. (1996), “An EM algorithm for Nonlinear Random Effects Models,” *Biometrics*, 52. 934–944.

Wang, N., and Davidian, M. (1996), “A Note on Covariate Measurement Error in Nonlinear Mixed Effects Models,” *Biometrika*, 83, 801–812.

Wolfinger, R. (1993), “Laplace’s Approximation for Nonlinear Mixed Models,” *Biometrika*, 80, 791–795.

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

Wu, H.L., and Ding, A.A. (1999), “Population HIV-1 Dynamics in vivo: Applicable Models and Inferential Tools for Virological Data From AIDS Clinical Trials,” *Biometrics*, 55, 410–418.

Wu, H.L., and Wu, L. (2002a), “Identification of Significant Host Factors for HIV Dynamics Modelled by Non-Linear Mixed-Effects Models,” *Statistics in Medicine*, 21, 753–771.

Wu, L. (2002), “A Joint Model for Nonlinear Mixed-Effects Models With Censoring and Covariates Measured With Error, With Application to AIDS Studies,” *Journal of the American Statistical Association*, 97, 955–964.

Wu, L., and Wu, H.L. (2002b), “Missing Time-Dependent Covariates in Human Immunodeficiency Virus Dynamic Models,” *Applied Statistics*, 51, 2002.

Yeap, B.Y., Catalano, P.J., Ryan, L.M., and Davidian, M. (2003), ‘Robust Two-Stage Approach to Repeated Measurements Analysis of Chronic Ozone Exposure in Rats,” *Journal of Agricultural, Biological, and Environmental Statistics*, 8, ????-????.

Yeap, B.Y., and Davidian, M. (2001), “Robust Two-Stage Estimation in Hierarchical Nonlinear Models,” *Biometrics*, 57, 266–272.

Young, D.A., Zerbe, G.O., and Hay, W.W. (1997), “Fieller’s Theorem, Scheff’e Simultaneous Confidence Intervals, and Ratios of Parameters of Linear and Nonlinear Mixed-Effects Models,” *Biometrics*, 53, 838–347.

Zeng, Q., and Davidian, M. (1997), “Testing Homgeneity of Intra-run Variance Parameters in Immunoassay,” *Statistics in Medicine*, 16, 1765–1776.

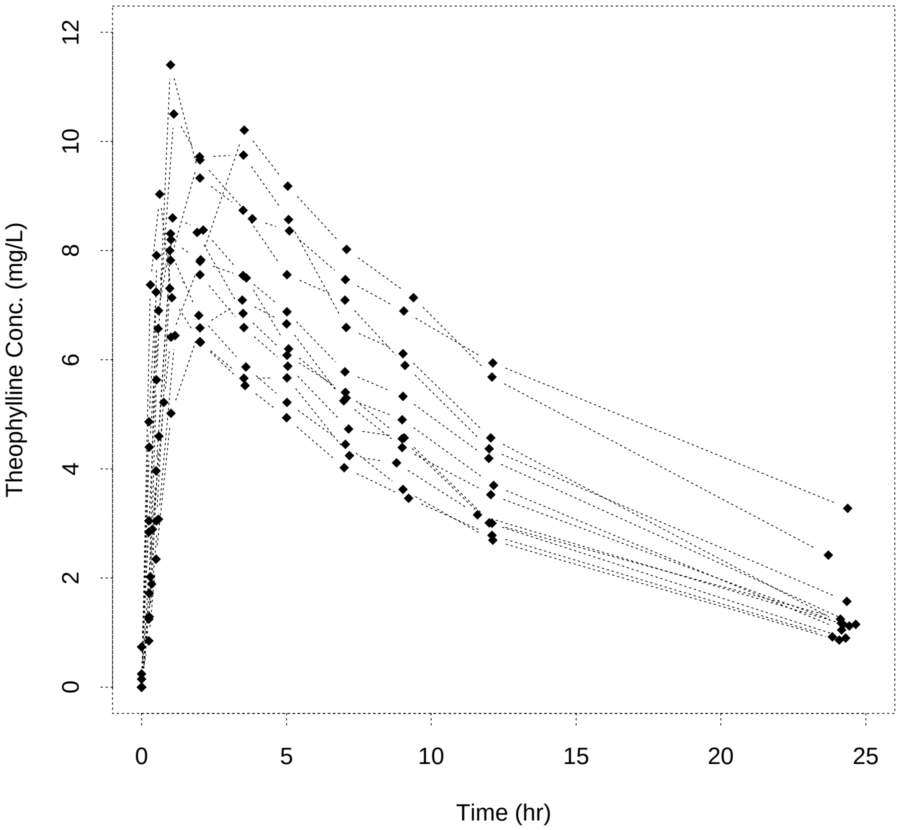

Figure 1. Theophylline concentrations for 12 subjects following a single oral dose.

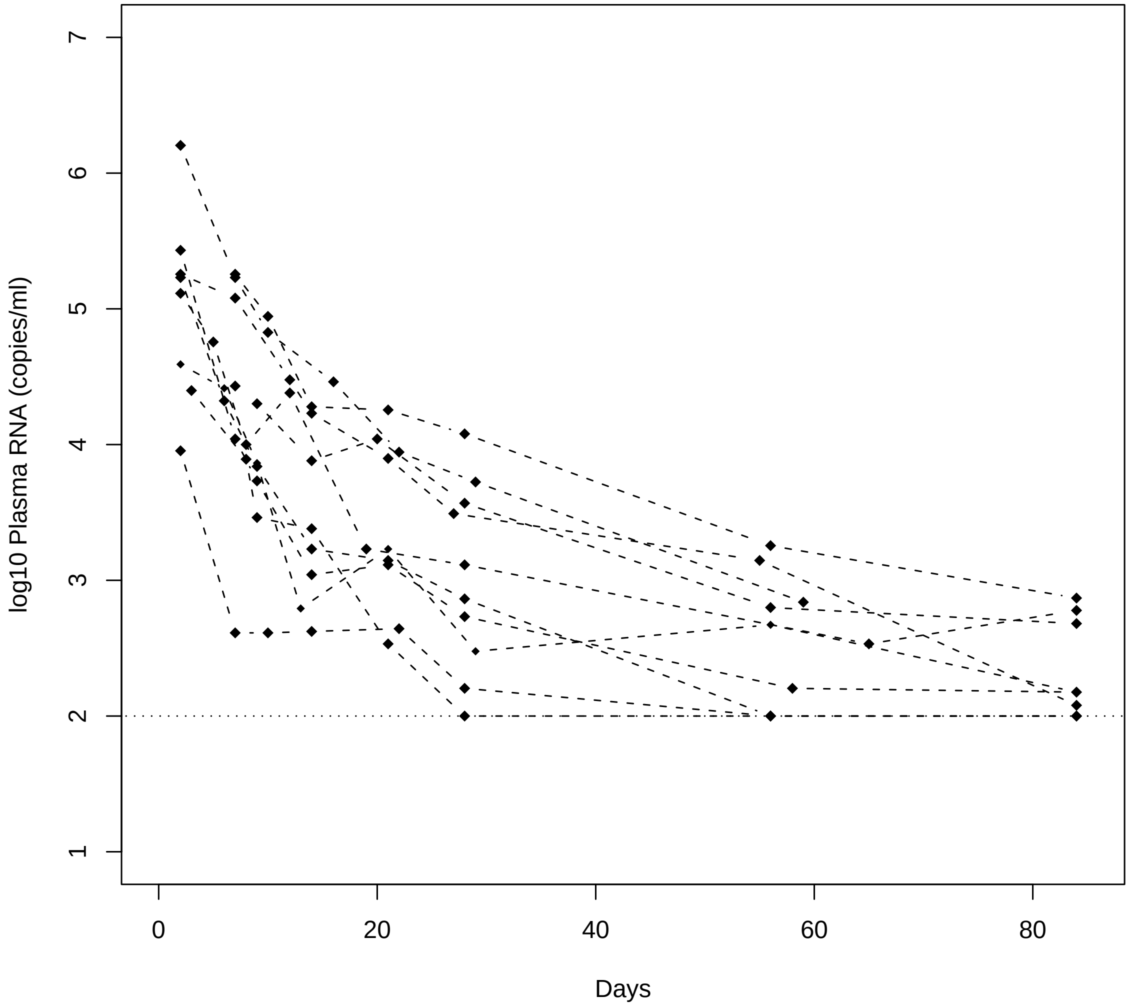

Figure 2. Viral load profiles for 10 subjects from the ACTG 315 study. The lower limit of detection of 100 copies/ml is denoted by the dotted line.

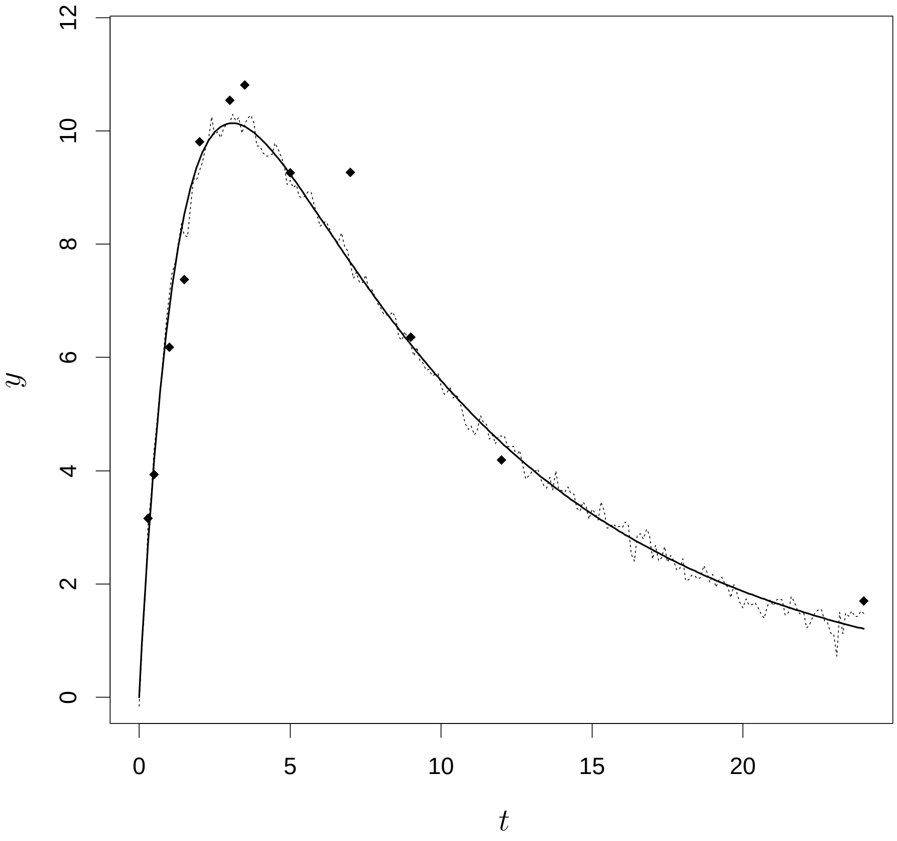

Figure 3. Intra-individual sources of variation. The solid line is the "inherent" trajectory, the dotted line is the "realization" of the response process that actually takes place, and the solid diamonds are measurements of the "realization" at particular time points that are subject to error.