José C. Pinheiro
Douglas M. Bates

Mixed-Effects Models in S and S-PLUS

Springer

José C. Pinheiro
Department of Biostatistics
Novartis Pharmaceuticals
One Health Plaza
East Hanover, NJ 07936-1080
USA
jose.pinheiro@pharma.novartis.com

Douglas M. Bates
Department of Statistics
University of Wisconsin
Madison, WI 53706-1685
USA
bates@stat.wisc.edu

Library of Congress Cataloging-in-Publication Data
Pinheiro, José C.
Mixed-effects models in S and S-PLUS/José C. Pinheiro, Douglas M. Bates
p. cm.—(Statistics and computing)
Includes bibliographical references and index.
ISBN 0-387-98957-9 (alk. paper)
I. Bates, Douglas M. II. Title. III. Series.
QA76.73.S15P56 2000
005.13'3—dc21 99-053566

Printed on acid-free paper.

Printed in the United States of America. (HAM/BP)

9 8 7

springer.com

7

Theory and Computational Methods for Nonlinear Mixed-Effects Models

This chapter presents the theory for the nonlinear mixed-effects model introduced in Chapter 6. A general formulation of NLME models is presented and illustrated with examples. Estimation methods for fitting NLME models, based on approximations to the likelihood function, are described and discussed. The computational methods used in the `nlme` function to fit NLME models are also described. An extended class of nonlinear regression models with heteroscedastic, correlated errors, but with no random effects, is presented.

The objective of this chapter is to give an overall description of the theoretical and computational aspects of NLME models so as to allow one to evaluate the strengths and limitations of such models in practice. It is not the purpose of this chapter to present a thorough theoretical description of NLME models. Such a comprehensive treatment of the theory of nonlinear mixed-effects models can be found, for example, in Davidian and Giltinan (1995) and in Vonesh and Chinchilli (1997).

Readers who are more interested in the applications of NLME models and the use of the functions and methods in the nlme library to fit such models can, without loss of continuity, skip this chapter and go straight to Chapter 8. If you decide to skip this chapter at a first reading, it is recommended that you return to it (especially §7.1) at a later time to get a good understanding of the NLME model formulation and its assumptions and limitations.

7.1 The NLME Model Formulation

Nonlinear mixed-effects models are mixed-effects models in which some, or all, of the fixed and random effects occur nonlinearly in the model function. They can be regarded either as an extension of linear mixed-effects models in which the conditional expectation of the response given the random effects is allowed to be a nonlinear function of the coefficients, or as an extension of nonlinear regression models for independent data (Bates and Watts, 1988) in which random effects are incorporated in the coefficients to allow then to vary by group, thus inducing correlation within the groups.

This section presents a general formulation for NLME models proposed by Lindstrom and Bates (1990). The NLME model for single-level grouped data, which includes repeated measures and longitudinal data, is presented in §7.1.1. The multilevel NLME model is described in §7.1.2.

7.1.1 *Single-Level of Grouping*

By far the most common application of NLME models is for repeated measures data—in particular, longitudinal data. The nonlinear mixed-effects model for repeated measures proposed by Lindstrom and Bates (1990) can be thought of as a hierarchical model. At one level the jth observation on the ith group is modeled as

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}) + \epsilon_{ij}, \qquad i = 1, \dots, M, \; j = 1, \dots, n_i, \tag{7.1}$$

where M is the number of groups, n_i is the number of observations on the ith group, f is a general, real-valued, differentiable function of a group-specific parameter vector $\boldsymbol{\phi}_{ij}$ and a covariate vector $\boldsymbol{v}_{ij}$, and ϵ_{ij} is a normally distributed within-group error term. The function f is nonlinear in at least one component of the group-specific parameter vector $\boldsymbol{\phi}_{ij}$, which is modeled as

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \tag{7.2}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of *fixed effects* and $\boldsymbol{b}_i$ is a q-dimensional *random effects* vector associated with the ith group (not varying with j) with variance–covariance matrix $\boldsymbol{\Psi}$. The matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are of appropriate dimensions and depend on the group and possibly on the values of some covariates at the jth observation. This model is a slight generalization of that described in Lindstrom and Bates (1990) in that $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ can depend on j. This generalization allows the incorporation of "time-varying" covariates in the fixed effects or the random effects for the model. It is assumed that observations corresponding to different groups are independent and that the within-group errors ϵ_{ij} are independently distributed as $\mathcal{N}(0, \sigma^2)$ and independent of the $\boldsymbol{b}_i$. The assumption of independence and homoscedasticity for the within-group errors can be relaxed, as shown in §7.4.

Because f can be any nonlinear function of $\boldsymbol{\phi}_{ij}$, the representation of the group-specific coefficients $\boldsymbol{\phi}_{ij}$ could be chosen so that $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are always simple incidence matrices. However, it is desirable to encapsulate as much modeling of the $\boldsymbol{\phi}_{ij}$ as possible in this second stage, as this simplifies the calculation of the derivatives of the model function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, used in the optimization algorithm. In a call to `nlme` the arguments `fixed` and `random` are used to specify the $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ matrices, respectively.

We can write (7.1) and (7.2) in matrix form as

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{f}_i(\boldsymbol{\phi}_i, \boldsymbol{v}_i) + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\phi}_i &= \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \end{aligned} \tag{7.3}$$

for $i = 1, \dots, M$, where

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \; \boldsymbol{\phi}_i = \begin{bmatrix} \boldsymbol{\phi}_{i1} \\ \vdots \\ \boldsymbol{\phi}_{in_i} \end{bmatrix}, \; \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}, \; \boldsymbol{f}_i(\boldsymbol{\phi}_i, \boldsymbol{v}_i) = \begin{bmatrix} f(\boldsymbol{\phi}_{i1}, \boldsymbol{v}_{i1}) \\ \vdots \\ f(\boldsymbol{\phi}_{in_i}, \boldsymbol{v}_{in_i}) \end{bmatrix},$$

$$\boldsymbol{v}_i = \begin{bmatrix} \boldsymbol{v}_{i1} \\ \vdots \\ \boldsymbol{v}_{in_i} \end{bmatrix}, \quad \boldsymbol{A}_i = \begin{bmatrix} \boldsymbol{A}_{i1} \\ \vdots \\ \boldsymbol{A}_{in_i} \end{bmatrix}, \quad \boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{B}_{i1} \\ \vdots \\ \boldsymbol{B}_{in_i} \end{bmatrix}. \tag{7.4}$$

We use the examples of Chapter 6 to illustrate the general NLME model formulation.

Indomethacin Kinetics

The final model obtained in §6.2 for the indomethacin data, represented by the object `fm4Indom.nlme`, expresses the concentration measurement y_{ij} for the ith subject at time t_j as

$$y_{ij} = \phi_{1i} \exp\left[-\exp\left(\phi'_{2i}\right) t_j\right] + \phi_{3i} \exp\left[-\exp\left(\phi'_{4i}\right) t_j\right] + \epsilon_{ij},$$

$$\underbrace{\begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix}}_{\boldsymbol{\phi}_{ij}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{A}_{ij}} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{B}_{ij}} \underbrace{\begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix}}_{\boldsymbol{b}_i},$$

$$\boldsymbol{b}_i \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \psi_{11} & \psi_{12} & 0 \\ \psi_{12} & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}\right), \quad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right).$$

In this case, the individual coefficients $\boldsymbol{\phi}_{ij}$ and the design matrices $\boldsymbol{A}_{ij} = \boldsymbol{I}$ and $\boldsymbol{B}_{ij}$ do not vary with time. The $\boldsymbol{\Psi}$ matrix for the random effects is block-diagonal.

Growth of Soybean Plants

The fitted object `fm4Soy.nlme` represents the final model for the soybean data obtained in §6.3. We use plot `1990P8`, for which `Year = 1990` and `Variety = P`, to illustrate the general NLME model representation, expressing the average leaf weight y_{ij} for plot i at t_{ij} days after planting as

$$y_{ij} = \frac{\phi_{1i}}{1 + \exp\left[-\left(t_{ij} - \phi_{2i}\right)/\phi_{3i}\right]} + \epsilon_{ij},$$

$$\underbrace{\begin{bmatrix}\phi_{1i}\\ \phi_{2i}\\ \phi_{3i}\end{bmatrix}}_{\boldsymbol{\phi}_{ij}} = \underbrace{\begin{bmatrix}1&0&1&1&0&1&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&1&1&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&1&0&1\end{bmatrix}}_{\boldsymbol{A}_{ij}} \underbrace{\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5\\ \beta_6\\ \beta_7\\ \beta_8\\ \beta_9\\ \beta_{10}\\ \beta_{11}\\ \beta_{12}\\ \beta_{13}\end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix}1\\ 0\\ 0\end{bmatrix}}_{\boldsymbol{B}_{ij}} \underbrace{\begin{bmatrix}b_{1i}\end{bmatrix}}_{b_i},$$

$$b_i \sim \mathcal{N}\left(0, \psi\right), \quad \epsilon_{ij}|\boldsymbol{\phi}_i \sim \mathcal{N}\left(0, \sigma^2\left[\mathrm{E}\left(y_{ij}|\boldsymbol{\phi}_i\right)\right]^{\theta}\right).$$

The correspondence between the fixed effects, $\boldsymbol{\beta}$, and the coefficient names used in `fm4Soy.nlme` is: $\beta_1 =$ `Asym.(Intercept)`, $\beta_2 =$ `Asym.Year1989`, $\beta_3 =$ `Asym.Year1990`, $\beta_4 =$ `Asym.Variety`, $\beta_5 =$ `Asym.Year1989Variety`, $\beta_6 =$ `Asym.Year1990Variety`, $\beta_7 =$ `xmid.(Intercept)`, $\beta_8 =$ `xmid.Year1989`, $\beta_9 =$ `xmid.Year1990`, $\beta_{10} =$ `xmid.Variety`, $\beta_{11} =$ `scal.(Intercept)`, $\beta_{12} =$ `scal.Year1989`, and $\beta_{13} =$ `scal.Year1990`. The design matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ do not vary with j in this example and, as a result, neither do the coefficients $\boldsymbol{\phi}_{ij}$. The use of variance functions for the within-group errors is discussed in §7.4, when we present extensions to the basic NLME model.

Clinical Study of Phenobarbital Kinetics

The final model for the phenobarbital data, represented in §6.4 by the object `fm3Pheno.nlme`, includes only the infant's birth weight w_i as a covariate for the fixed effects. The phenobarbital concentration y_{ij} for individual i measured at time t_{ij}, following intravenous injections of dose D_{id} at times t_{id}, is expressed as

$$y_{ij} = \sum_{d:t_{id}<t_{ij}} \frac{D_{id}}{\exp\left(lV_i\right)} \exp\left[-\exp\left(lCl_i - lV_i\right)\left(t_{ij} - t_{id}\right)\right] + \epsilon_{ij},$$

$$\underbrace{\begin{bmatrix}lCl_i\\ lV_i\end{bmatrix}}_{\boldsymbol{\phi}_{ij}} = \underbrace{\begin{bmatrix}1&w_i&0&0\\ 0&0&1&w_i\end{bmatrix}}_{\boldsymbol{A}_{ij}} \underbrace{\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix}1&0\\ 0&1\end{bmatrix}}_{\boldsymbol{B}_{ij}} \underbrace{\begin{bmatrix}b_{1i}\\ b_{2i}\end{bmatrix}}_{\boldsymbol{b}_i},$$

$$\boldsymbol{b}_i \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix}\psi_{11}&0\\ 0&\psi_{22}\end{bmatrix}\right), \qquad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right).$$

The correspondence between the fixed effects, $\boldsymbol{\beta}$, and the coefficient names in the `fm3Pheno.nlme` object is: $\beta_1 =$ `lCl.(Intercept)`, $\beta_2 =$ `lCl.Wt`, $\beta_3 =$ `lV.(Intercept)`, and $\beta_4 =$ `lV.Wt`. A diagonal $\boldsymbol{\Psi}$ matrix is used to represent the independence between the random effects.

7.1.2 Multilevel NLME Models

The single-level NLME model (7.1) can be extended to data grouped according to multiple, nested factors by modifying the model for the random effects in (7.2). For example, the multilevel version of the Lindstrom and Bates (1990) model for two levels of nesting is written as a two-stage model in which the first stage expresses the response y_{ijk} for the kth observation on the jth second-level group of the ith first-level group as

$$\begin{aligned} y_{ijk} &= f\left(\boldsymbol{\phi}_{ijk}, \boldsymbol{v}_{ijk}\right) + \epsilon_{ijk},\\ &\qquad i = 1, \ldots, M, \quad j = 1, \ldots, M_i, \quad k = 1, \ldots, n_{ij}, \end{aligned} \tag{7.5}$$

where M is the number of first-level groups, M_i is the number of second-level groups within the ith first-level group, n_{ij} is the number of observations on the jth second-level group of the ith first-level group, and ϵ_{ijk} is a normally distributed within-group error term. As in the single-level model, f is a general, real-valued, differentiable function of a group-specific parameter vector $\boldsymbol{\phi}_{ijk}$ and a covariate vector $\boldsymbol{v}_{ijk}$. It is nonlinear in at least one component of $\boldsymbol{\phi}_{ij}$. The second stage of the model expresses $\boldsymbol{\phi}_{ij}$ as

$$\begin{aligned} \boldsymbol{\phi}_{ijk} &= \boldsymbol{A}_{ijk}\boldsymbol{\beta} + \boldsymbol{B}_{i,jk}\boldsymbol{b}_i + \boldsymbol{B}_{ijk}\boldsymbol{b}_{ij},\\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_2). \end{aligned} \tag{7.6}$$

As in the single-level model (7.2), $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects, with design matrix $\boldsymbol{A}_{ijk}$, which may incorporate time-varying covariates. The first-level random effects $\boldsymbol{b}_i$ are independently distributed

q_1-dimensional vectors with variance–covariance matrix $\boldsymbol{\Psi}_1$. The second-level random effects $\boldsymbol{b}_{ij}$ are q_2-dimensional independently distributed vectors with variance–covariance matrix $\boldsymbol{\Psi}_2$, assumed to be independent of the first-level random effects. The random effects design matrices $\boldsymbol{B}_{i,jk}$ and $\boldsymbol{B}_{ijk}$ depend on first- and second-level groups and possibly on the values of some covariates at the kth observation. The within-group errors ϵ_{ijk} are independently distributed as $\mathcal{N}(0, \sigma^2)$ and are independent of the random effects. The assumption of independence and homoscedasticity for the within-group errors can be relaxed, as shown in §7.4.

We can express (7.5) and (7.6) in matrix form as

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{f}_{ij}\left(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}\right) + \boldsymbol{\epsilon}_{ij}, \\ \boldsymbol{\phi}_{ij} &= \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{i,j}\boldsymbol{b}_i + \boldsymbol{B}_{ij}\boldsymbol{b}_{ij}, \end{aligned} \tag{7.7}$$

for $i = 1, \ldots, M$, $j = 1, \ldots, M_i$, where

$$\boldsymbol{y}_{ij} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijn_{ij}} \end{bmatrix}, \quad \boldsymbol{\phi}_{ij} = \begin{bmatrix} \boldsymbol{\phi}_{ij1} \\ \vdots \\ \boldsymbol{\phi}_{ijn_{ij}} \end{bmatrix}, \quad \boldsymbol{\epsilon}_{ij} = \begin{bmatrix} \epsilon_{ij1} \\ \vdots \\ \epsilon_{ijn_{ij}} \end{bmatrix},$$

$$\boldsymbol{f}_{ij}\left(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}\right) = \begin{bmatrix} f(\boldsymbol{\phi}_{ij1}, \boldsymbol{v}_{ij1}) \\ \vdots \\ f(\boldsymbol{\phi}_{ijn_{ij}}, \boldsymbol{v}_{ijn_{ij}}) \end{bmatrix}, \quad \boldsymbol{v}_{ij} = \begin{bmatrix} \boldsymbol{v}_{ij1} \\ \vdots \\ \boldsymbol{v}_{ijn_{ij}} \end{bmatrix},$$

$$\boldsymbol{A}_{ij} = \begin{bmatrix} \boldsymbol{A}_{ij1} \\ \vdots \\ \boldsymbol{A}_{ijn_{ij}} \end{bmatrix}, \quad \boldsymbol{B}_{i,j} = \begin{bmatrix} \boldsymbol{B}_{i,j1} \\ \vdots \\ \boldsymbol{B}_{i,jn_{ij}} \end{bmatrix}, \quad \boldsymbol{B}_{ij} = \begin{bmatrix} \boldsymbol{B}_{ij1} \\ \vdots \\ \boldsymbol{B}_{ijn_{ij}} \end{bmatrix}.$$

Extensions of the NLME model to more than two levels of nesting are straightforward. For example, with three levels of nesting the second-stage model for the group-specific coefficients is

$$\begin{aligned} \boldsymbol{\phi}_{ijkl} &= \boldsymbol{A}_{ijkl}\boldsymbol{\beta} + \boldsymbol{B}_{i,jkl}\boldsymbol{b}_i + \boldsymbol{B}_{ij,kl}\boldsymbol{b}_{ij} + \boldsymbol{B}_{ijkl}\boldsymbol{b}_{ijk}, \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_2), \quad \boldsymbol{b}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_3). \end{aligned}$$

7.1.3 Other NLME Models

The first developments of nonlinear mixed-effects models appear in Sheiner and Beal (1980). Their models and estimation methods are incorporated in the NONMEM program (Beal and Sheiner, 1980), which is widely used in pharmacokinetics. They introduced a model similar to (7.1) and developed a maximum likelihood estimation method based on a first-order Taylor expansion of the model function around $\boldsymbol{0}$, the expected value of the random effects vector $\boldsymbol{b}$.

A nonparametric maximum likelihood method for nonlinear mixed-effects models was proposed by Mallet, Mentre, Steimer and Lokiek (1988). They use a model similar to (7.1), but make no assumptions about the distribution of the random effects. The conditional distribution of the response given the random effects is assumed to be known. The objective of the estimation procedure is to get the probability distribution of the group-specific coefficients $(\boldsymbol{\phi}_{ij})$ that maximizes the likelihood of the data. Mallet (1986) showed that the maximum likelihood solution is a discrete distribution with the number of discontinuity points less than or equal to the number of groups in the sample. Inference for this model is based on the maximum likelihood distribution from which summary statistics (e.g., means and variance–covariance matrices) and plots are obtained.

Davidian and Gallant (1992) introduced a smooth, nonparametric maximum likelihood estimation method for nonlinear mixed effects. Their model is similar to (7.1), but with a more general definition for the group-specific coefficients, $\boldsymbol{\phi}_{ij} = g(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{v}_{ij})$, where g is a general, possibly nonlinear function. As in Mallet et al. (1988), Davidian and Gallant assume that the conditional distribution of the response vector given the random effects is known (up to the parameters that define it), but the distribution of the random effects is free to vary within a class of smooth densities defined in Gallant and Nychka (1987).

A Bayesian approach using hierarchical models for nonlinear mixed effects is described in Bennett and Wakefield (1993) and Wakefield (1996). The first stage model is similar to (7.1) and the distributions of both the random effects and the errors ϵ_{ij} are assumed known up to population parameters. Prior distributions for the population parameters must be provided then Markov-chain Monte Carlo methods, such as the Gibbs sampler (Geman and Geman, 1984) or the Metropolis algorithm (Hastings, 1970), are used to approximate the posterior density of the random effects.

Vonesh and Carter (1992) developed a mixed-effects model that is nonlinear in the fixed effects, but linear in the random effects. Their model is

$$\boldsymbol{y}_i = \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{v}_i) + \boldsymbol{Z}_i(\boldsymbol{\beta})\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where, as before, $\boldsymbol{\beta}$, $\boldsymbol{b}_i$, and $\boldsymbol{\epsilon}_i$ denote, respectively, the fixed effects, the random effects, and the within-group error term, $\boldsymbol{v}_i$ is a matrix of covariates, and $\boldsymbol{Z}_i$ is a full-rank matrix of known functions of the fixed effects $\boldsymbol{\beta}$. It is further assumed that $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and the two vectors are independent. In some sense, Vonesh and Carter incorporate in the model the approximations suggested by Sheiner and Beal (1980) and Lindstrom and Bates (1990). Their approach concentrates more on inferences about the fixed effects, and less on the variance–covariance components of the random effects.

7.2 Estimation and Inference in NLME Models

Different methods have been proposed to estimate the parameters in the NLME model described in §7.1.1 and §7.1.2. In this book we restrict ourselves to methods based on the likelihood function. Descriptions and comparisons of other estimation methods proposed for NLME models can be found, for example, in Ramos and Pantula (1995), Davidian and Giltinan (1995), and Vonesh and Chinchilli (1997).

7.2.1 Likelihood Estimation

Because the random effects are unobserved quantities, maximum likelihood estimation in mixed-effects models is based on the marginal density of the responses $\boldsymbol{y}$, which, for a model with Q levels of nesting, is calculated as

$$p\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\Psi}_1,\dots,\boldsymbol{\Psi}_Q\right)=\int p\left(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2\right)\,p\left(\boldsymbol{b}|\boldsymbol{\Psi}_1,\dots,\boldsymbol{\Psi}_Q\right)\,d\boldsymbol{b},\qquad(7.8)$$

where $p\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\Psi}_1,\dots,\boldsymbol{\Psi}_Q\right)$ is the marginal density of $\boldsymbol{y}$, $p(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2)$ is the conditional density of $\boldsymbol{y}$ given the random effects $\boldsymbol{b}$, and the marginal distribution of $\boldsymbol{b}$ is $p\left(\boldsymbol{b}|\boldsymbol{\Psi}_1,\dots,\boldsymbol{\Psi}_Q\right)$. For the NLME model (7.1), expressing the random effects variance–covariance matrix in terms of the precision factor $\boldsymbol{\Delta}$, so that $\boldsymbol{\Psi}^{-1}=\sigma^{-2}\boldsymbol{\Delta}^T\boldsymbol{\Delta}$ as described in §2.1.1, provides the marginal density of $\boldsymbol{y}$ as

$$p\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\Delta}\right)=\\ \frac{|\boldsymbol{\Delta}|^M}{\left(2\pi\sigma^2\right)^{(N+Mq)/2}}\prod_{i=1}^{M}\int\exp\left\{\frac{\|\boldsymbol{y}_i-\boldsymbol{f}_i\left(\boldsymbol{\beta},\boldsymbol{b}_i\right)\|^2+\|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2}{-2\sigma^2}\right\}d\boldsymbol{b}_i,\qquad(7.9)$$

where $\boldsymbol{f}_i\left(\boldsymbol{\beta},\boldsymbol{b}_i\right)=\boldsymbol{f}_i\left[\boldsymbol{\phi}_i\left(\boldsymbol{\beta},\boldsymbol{b}_i\right),\boldsymbol{v}_i\right]$.

Because the model function f can be nonlinear in the random effects, the integral in (7.8) generally does not have a closed-form expression. To make the numerical optimization of the likelihood function a tractable problem, different approximations to (7.8) have been proposed. Some of these methods consist of taking a first-order Taylor expansion of the model function f around the expected value of the random effects (Sheiner and Beal, 1980; Vonesh and Carter, 1992), or around the conditional (on $\boldsymbol{\Delta}$) modes of the random effects (Lindstrom and Bates, 1990). Gaussian quadrature rules have also been used (Davidian and Gallant, 1992).

We describe three different methods for approximating the likelihood function in the NLME model. The first, proposed by Lindstrom and Bates (1990), approximates (7.8) by the likelihood of a linear mixed-effects model. We call this the LME approximation. It is the basis of the estimation algorithm currently implemented in the `nlme` function. The second method uses a Laplacian approximation to the likelihood function, and the last method uses an adaptive Gaussian quadrature rule to improve the Laplacian approximation. The LME, Laplacian, and adaptive Gaussian approximations have increasing degrees of accuracy, at the cost of increasing computational complexity. The three approximations to the NLME likelihood are discussed and compared in Pinheiro and Bates (1995).

Lindstrom and Bates Algorithm

The estimation algorithm described by Lindstrom and Bates (1990) alternates between two steps, a penalized nonlinear least squares (PNLS) step, and a linear mixed effects (LME) step, as described below. We initially consider the alternating algorithm for the single-level NLME model (7.1).

In the PNLS step, the current estimate of $\boldsymbol{\Delta}$ (the precision factor) is held fixed, and the conditional modes of the random effects $\boldsymbol{b}_i$ and the conditional estimates of the fixed effects $\boldsymbol{\beta}$ are obtained by minimizing a penalized nonlinear least squares objective function

$$\sum_{i=1}^{M}\left[\|\boldsymbol{y}_i-\boldsymbol{f}_i\left(\boldsymbol{\beta},\boldsymbol{b}_i\right)\|^2+\|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2\right].\qquad(7.10)$$

The LME step updates the estimate of $\boldsymbol{\Delta}$ based on a first-order Taylor expansion of the model function f around the current estimates of $\boldsymbol{\beta}$ and the conditional modes of the random effects $\boldsymbol{b}_i$, which we will denote by $\widehat{\boldsymbol{\beta}}^{(w)}$ and $\widehat{\boldsymbol{b}}_i^{(w)}$, respectively. Letting

$$\widehat{\boldsymbol{X}}_i^{(w)}=\left.\frac{\partial\boldsymbol{f}_i}{\partial\boldsymbol{\beta}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)},\widehat{\boldsymbol{b}}_i^{(w)}},\qquad\widehat{\boldsymbol{Z}}_i^{(w)}=\left.\frac{\partial\boldsymbol{f}_i}{\partial\boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)},\widehat{\boldsymbol{b}}_i^{(w)}},\\ \widehat{\boldsymbol{w}}_i^{(w)}=\boldsymbol{y}_i-\boldsymbol{f}_i(\widehat{\boldsymbol{\beta}}^{(w)},\widehat{\boldsymbol{b}}_i^{(w)})+\widehat{\boldsymbol{X}}_i^{(w)}\widehat{\boldsymbol{\beta}}^{(w)}+\widehat{\boldsymbol{Z}}_i^{(w)}\widehat{\boldsymbol{b}}_i^{(w)},\qquad(7.11)$$

the approximate log-likelihood function used to estimate $\boldsymbol{\Delta}$ is

$$\ell_{\text{LME}}\left(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Delta}\mid\boldsymbol{y}\right)=-\frac{N}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\boldsymbol{\Sigma}_i(\boldsymbol{\Delta})\right|\right.\\ \left.+\sigma^{-2}\left[\widehat{\boldsymbol{w}}_i^{(w)}-\widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta})\left[\widehat{\boldsymbol{w}}_i^{(w)}-\widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]\right\},\qquad(7.12)$$

where $\boldsymbol{\Sigma}_i(\boldsymbol{\Delta})=\boldsymbol{I}+\widehat{\boldsymbol{Z}}_i^{(w)}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^{-T}\widehat{\boldsymbol{Z}}_i^{(w)^T}$. This log-likelihood is identical to that of a linear mixed-effects model in which the response vector is given by $\widehat{\boldsymbol{w}}^{(w)}$ and the fixed- and random-effects design matrices are given by $\widehat{\boldsymbol{X}}^{(w)}$ and $\widehat{\boldsymbol{Z}}^{(w)}$, respectively. Using the results in §2.2, one can express the optimal values of $\boldsymbol{\beta}$ and σ^2 as functions of $\boldsymbol{\Delta}$ and work with the profiled log-likelihood of $\boldsymbol{\Delta}$, greatly simplifying the optimization problem.

Lindstrom and Bates (1990) also proposed a restricted maximum likelihood estimation method for $\boldsymbol{\Delta}$, which consists of replacing the log-likelihood in the LME step of the alternating algorithm by the log-restricted-likelihood

$$\ell_{\text{LME}}^{R}\left(\sigma^2, \boldsymbol{\Delta} \mid \boldsymbol{y}\right) =$$
$$\ell_{\text{LME}}\left(\widehat{\boldsymbol{\beta}}\left(\boldsymbol{\Delta}\right), \sigma^2, \boldsymbol{\Delta} \mid \boldsymbol{y}\right) - \frac{1}{2}\log\left|\sum_{i=1}^{M} \sigma^{-2}\widehat{\boldsymbol{X}}_i^{(w)^T} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta})\widehat{\boldsymbol{X}}_i^{(w)}\right|. \quad (7.13)$$

Note that, because $\widehat{\boldsymbol{X}}_i^{(w)}$ depends on both $\widehat{\boldsymbol{\beta}}^{(w)}$ and $\widehat{\boldsymbol{b}}_i^{(w)}$, changes in either the fixed effects model or the random effects model imply changes in the penalty factor for the log-restricted-likelihood (7.13). Therefore, log-restricted-likelihoods from NLME models with different fixed or random effects models are not comparable.

The algorithm alternates between the PNLS and LME steps until a convergence criterion is met. Such alternating algorithms tend to be more efficient when the estimates of the variance–covariance components ($\boldsymbol{\Delta}$ and σ^2) are not highly correlated with the estimates of the fixed effects ($\boldsymbol{\beta}$). Pinheiro (1994) has shown that, in the linear mixed-effects model, the maximum likelihood estimates of $\boldsymbol{\Delta}$ and σ^2 are asymptotically independent of the maximum likelihood estimates of $\boldsymbol{\beta}$. These results have not yet been extended to the nonlinear mixed-effects model (7.1).

Lindstrom and Bates (1990) only use the LME step to update the estimate of $\boldsymbol{\Delta}$. However, the LME step also produces updated estimates of $\boldsymbol{\beta}$ and the conditional modes of $\boldsymbol{b}_i$. Thus, one can iterate LME steps by re-evaluating (7.11) and (7.12) (or (7.13) for the log-restricted-likelihood) at the updated estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, as described in Wolfinger (1993). Because the updated estimates correspond to the values obtained in the first iteration of a Gauss–Newton algorithm for the PNLS step, iterated LME steps will converge to the same values as the alternating algorithm, though possibly not as quickly.

Wolfinger (1993) also shows that, when a flat prior is assumed for $\boldsymbol{\beta}$, the LME approximation to the log-restricted-likelihood (7.13) is equivalent to a Laplacian approximation (Tierney and Kadane, 1986) to the integral (7.9).

The alternating algorithm and the LME approximation to the NLME log-likelihood can be extended to multilevel models. For example, for an NLME model with two levels of nesting, the PNLS step consists of minimizing the penalized nonlinear least-squares function

$$\sum_{i=1}^{M}\left\{\sum_{j=1}^{M_i}\left[\|\boldsymbol{y}_{ij} - \boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij})\|^2 + \|\boldsymbol{\Delta}_2\boldsymbol{b}_{ij}\|^2\right] + \|\boldsymbol{\Delta}_1\boldsymbol{b}_i\|^2\right\} \quad (7.14)$$

to obtain estimates for the fixed effects $\boldsymbol{\beta}$ and the conditional (on $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$) modes of the random effects $\boldsymbol{b}_i$ and $\boldsymbol{b}_{ij}$.

Letting

$$\widehat{\boldsymbol{X}}_{ij}^{(w)} = \left.\frac{\partial \boldsymbol{f}_{ij}}{\partial \boldsymbol{\beta}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}, \widehat{\boldsymbol{b}}_{ij}^{(w)}}, \quad \widehat{\boldsymbol{Z}}_{i,j}^{(w)} = \left.\frac{\partial \boldsymbol{f}_{ij}}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}, \widehat{\boldsymbol{b}}_{ij}^{(w)}},$$
$$\widehat{\boldsymbol{Z}}_{ij}^{(w)} = \left.\frac{\partial \boldsymbol{f}_{ij}}{\partial \boldsymbol{b}_{ij}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}, \widehat{\boldsymbol{b}}_{ij}^{(w)}},$$
$$\widehat{\boldsymbol{w}}_{ij}^{(w)} = \boldsymbol{y}_{ij} - \boldsymbol{f}_{ij}(\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}, \widehat{\boldsymbol{b}}_{ij}^{(w)}) + \widehat{\boldsymbol{X}}_{ij}^{(w)}\widehat{\boldsymbol{\beta}}^{(w)} + \widehat{\boldsymbol{Z}}_{i,j}^{(w)}\widehat{\boldsymbol{b}}_i^{(w)} + \widehat{\boldsymbol{Z}}_{ij}^{(w)}\widehat{\boldsymbol{b}}_{ij}^{(w)},$$
$$\widehat{\boldsymbol{X}}_i^{(w)} = \begin{bmatrix} \widehat{\boldsymbol{X}}_{i1}^{(w)} \\ \vdots \\ \widehat{\boldsymbol{X}}_{iM_i}^{(w)} \end{bmatrix}, \quad \widehat{\boldsymbol{Z}}_i^{(w)} = \begin{bmatrix} \widehat{\boldsymbol{Z}}_{i,1}^{(w)} \\ \vdots \\ \widehat{\boldsymbol{Z}}_{i,M_i}^{(w)} \end{bmatrix}, \quad \widehat{\boldsymbol{w}}_i^{(w)} = \begin{bmatrix} \widehat{\boldsymbol{w}}_{i1}^{(w)} \\ \vdots \\ \widehat{\boldsymbol{w}}_{iM_i}^{(w)} \end{bmatrix}, \quad (7.15)$$

the approximate log-likelihood function used to estimate $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ in the two-level NLME models is

$$\ell_{\text{LME}}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \mid \boldsymbol{y}\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\boldsymbol{\Sigma}_i(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)\right|\right.$$
$$\left. + \sigma^{-2}\left[\widehat{\boldsymbol{w}}_i^{(w)} - \widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)\left[\widehat{\boldsymbol{w}}_i^{(w)} - \widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]\right\},$$

where $\boldsymbol{\Sigma}_i(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) = \boldsymbol{I} + \widehat{\boldsymbol{Z}}_i^{(w)}\boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_1^{-T}\widehat{\boldsymbol{Z}}_i^{(w)^T} + \bigoplus_{j=1}^{M_i}\widehat{\boldsymbol{Z}}_{ij}^{(w)}\boldsymbol{\Delta}_2^{-1}\boldsymbol{\Delta}_2^{-T}\widehat{\boldsymbol{Z}}_{ij}^{(w)^T}$ and $\oplus$ denotes the direct sum operator. The corresponding log-restricted-likelihood is

$$\ell_{\text{LME}}^{R}\left(\sigma^2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \mid \boldsymbol{y}\right) = \ell_{\text{LME}}\left(\widehat{\boldsymbol{\beta}}\left(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2\right), \sigma^2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \mid \boldsymbol{y}\right)$$
$$- \frac{1}{2}\log\left|\sum_{i=1}^{M}\sigma^{-2}\widehat{\boldsymbol{X}}_i^{(w)^T}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)\widehat{\boldsymbol{X}}_i^{(w)}\right|.$$

This formulation can be extended to multilevel NLME models with an arbitrary number of levels.

The alternating algorithm is the only estimation algorithm used in the `nlme` function. It is implemented for maximum likelihood and restricted maximum likelihood estimation in single and multilevel NLME models.

Laplacian Approximation

Laplacian approximations are used frequently in Bayesian inference to estimate marginal posterior densities and predictive distributions (Tierney and Kadane, 1986; Leonard, Hsu and Tsui, 1989). These techniques can also be used for approximating the likelihood function in NLME models.

We consider initially the single-level NLME models. The integral that we want to estimate to obtain the marginal distribution of $\boldsymbol{y}_i$ in (7.9) can be written as

$$p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}) = \int \left(2\pi\sigma^2\right)^{-(n_i+q)/2} |\boldsymbol{\Delta}| \exp\left[-g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i)/2\sigma^2\right] d\boldsymbol{b}_i,$$

where $g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i) = \|\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2$, the sum of which is the objective function for the PNLS step of the alternating algorithm defined in (7.10). Let

$$\begin{aligned}
\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{b}}_i\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right) &= \arg\min_{\boldsymbol{b}_i} g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i), \\
g'\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i\right) &= \frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}, \qquad (7.16)\\
g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i\right) &= \frac{\partial^2 g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T},
\end{aligned}$$

and consider a second-order Taylor expansion of g around $\widehat{\boldsymbol{b}}_i$

$$g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i\right) \simeq g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right) + \frac{1}{2}\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right]^T g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right)\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right]. \qquad (7.17)$$

(The linear term in the expansion vanishes because $g'(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i) = \boldsymbol{0}$.) The Laplacian approximation is defined as

$$\begin{aligned}
&p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}\right) \\
&\quad \simeq \left(2\pi\sigma^2\right)^{-\frac{N}{2}} |\boldsymbol{\Delta}|^M \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{M} g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)\right] \\
&\quad \times \prod_{i=1}^{M} \int \left(2\pi\sigma^2\right)^{\frac{q}{2}} \exp\left\{-\tfrac{1}{2\sigma^2}\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right]^T g''(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right]\right\} d\boldsymbol{b}_i \\
&= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} |\boldsymbol{\Delta}|^M \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{M} g(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)\right] \prod_{i=1}^{M}\left|g''(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)\right|^{-\frac{1}{2}}.
\end{aligned}$$

The Hessian

$$\begin{aligned}
g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right) = &-\left.\frac{\partial^2 \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i}\left[\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\right] \\
&+ \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\right|_{\widehat{\boldsymbol{b}}_i} \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i} + \boldsymbol{\Delta}^T\boldsymbol{\Delta}
\end{aligned}$$

involves second derivatives of f but, at $\widehat{\boldsymbol{b}}_i$, the contribution of

$$\left.\frac{\partial^2 \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i}\left[\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\right]$$

is usually negligible compared to that of $\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)/\partial \boldsymbol{b}_i|_{\widehat{\boldsymbol{b}}_i}\ \partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)/\partial \boldsymbol{b}_i^T\big|_{\widehat{\boldsymbol{b}}_i}$ (Bates and Watts, 1980). Therefore, we use the approximation

$$\begin{aligned}
g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right) &\simeq \boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right) \\
&= \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\right|_{\widehat{\boldsymbol{b}}_i} \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i} + \boldsymbol{\Delta}^T\boldsymbol{\Delta}.
\end{aligned} \qquad (7.18)$$

This approximation is similar to that used in the Gauss–Newton algorithm for nonlinear least squares and has the advantage of requiring only the first-order partial derivatives of f with respect to the random effects. These are usually available as a by-product of the estimation of $\widehat{\boldsymbol{b}}_i$, which is a penalized least squares problem, for which standard and reliable code is available.

The modified Laplacian approximation to the log-likelihood of the single-level NLME model (7.1) is then given by

$$\begin{aligned}
\ell_{\mathrm{LA}}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \mid \boldsymbol{y}\right) = &-\frac{N}{2}\log\left(2\pi\sigma^2\right) + M\log|\boldsymbol{\Delta}| \\
&-\frac{1}{2}\left\{\sum_{i=1}^{M}\log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right| + \sigma^{-2}\sum_{i=1}^{M} g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right)\right\}. \qquad (7.19)
\end{aligned}$$

Because $\widehat{\boldsymbol{b}}_i$ does not depend on σ^2, for given $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$ the maximum likelihood estimate of σ^2 (based upon ℓ_{LA}) is

$$\widehat{\sigma}^2 = \widehat{\sigma}^2\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}\right) = \sum_{i=1}^{M} g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right)/N.$$

We can profile ℓ_{LA} on σ^2 to reduce the dimension of the optimization problem, obtaining

$$\begin{aligned}
&\ell_{\mathrm{LAp}}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}\right) = \\
&\quad -\frac{N}{2}\left[1 + \log\left(2\pi\right) + \log\left(\widehat{\sigma}^2\right)\right] + M\log|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right|.
\end{aligned}$$

If the model function f is linear in the random effects, then the modified Laplacian approximation is exact because the second-order Taylor expansion in (7.17) is exact when $\boldsymbol{f}_i\left(\boldsymbol{\beta}, \boldsymbol{b}_i\right) = \boldsymbol{f}_i\left(\boldsymbol{\beta}\right) + \boldsymbol{Z}_i\left(\boldsymbol{\beta}\right)\boldsymbol{b}_i$.

There does not yet seem to be a straightforward generalization of the concept of restricted maximum likelihood to NLME models. The difficulty is that restricted maximum likelihood depends heavily upon the linearity of the fixed effects in the model function, which does not occur in nonlinear models. Lindstrom and Bates (1990) circumvented that problem by using an approximation to the model function $\boldsymbol{f}$ in which the fixed effects, $\boldsymbol{\beta}$, occur linearly. This cannot be done for the Laplacian approximation, unless we consider yet another Taylor expansion of the model function, what would lead us back to something very similar to Lindstrom and Bates's approach.

The Laplacian approximation (7.19) can be extended to multilevel NLME models. For example, in a two-level NLME model, let

$$\boldsymbol{b}_i^{\text{aug}} = \begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{b}_{i1} \\ \vdots \\ \boldsymbol{b}_{iM_i} \end{bmatrix}, \qquad i = 1, \dots, M$$

denote the *augmented* random effects vector for the ith first-level group, containing the first-level random effects $\boldsymbol{b}_i$ and all the second-level random effects $\boldsymbol{b}_{ij}$ pertaining to first-level i. The two-level NLME likelihood can then be expressed as

$$p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) = \int \left(2\pi\sigma^2\right)^{-(n_i+q_1+M_i q_2)/2} |\boldsymbol{\Delta}_1| \, |\boldsymbol{\Delta}_2|^{M_i} \\ \times \exp\left[-g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}})/2\sigma^2\right] d\boldsymbol{b}_i^{\text{aug}},$$

where $g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}})$ is the objective function for the PNLS step in the alternating algorithm for two-level NLME models, defined in (7.14). We proceed as in the single-level case and define

$$\begin{aligned} \widehat{\boldsymbol{b}}_i^{\text{aug}} = \widehat{\boldsymbol{b}}_i^{\text{aug}}(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i) &= \arg\min_{\boldsymbol{b}_i^{\text{aug}}} g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}}), \\ g'(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}}) &= \frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}})}{\partial \boldsymbol{b}_i^{\text{aug}}}, \\ g''(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}}) &= \frac{\partial^2 g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}})}{\partial \boldsymbol{b}_i^{\text{aug}} \partial \left(\boldsymbol{b}_i^{\text{aug}}\right)^T}, \end{aligned}$$

to obtain the second-order approximation

$$g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}}) \simeq g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i^{\text{aug}}\right) \\ + \frac{1}{2}\left[\boldsymbol{b}_i^{\text{aug}} - \widehat{\boldsymbol{b}}_i^{\text{aug}}\right]^T g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i^{\text{aug}}\right)\left[\boldsymbol{b}_i^{\text{aug}} - \widehat{\boldsymbol{b}}_i^{\text{aug}}\right].$$

We note that $\partial^2 g(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \boldsymbol{b}_i^{\text{aug}})/\partial \boldsymbol{b}_{ij} \partial \boldsymbol{b}_{ik}^T = \boldsymbol{0}$ for any $j \neq k$ and use the same reasoning as in (7.18), to approximate the matrix $g''(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)$ by

$$g''\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i\right) \simeq \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i) = \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 \\ \boldsymbol{G}_2^T & \boldsymbol{G}_3 \end{bmatrix}, \quad \text{where}$$

$$\begin{aligned} \boldsymbol{G}_1 &= \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i^{\text{aug}})}{\partial \boldsymbol{b}_i}\right|_{\widehat{\boldsymbol{b}}_i} \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i^{\text{aug}})}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i^{\text{aug}}} + \boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1, \\ \boldsymbol{G}_2 &= \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i^{\text{aug}})}{\partial \boldsymbol{b}_i}\right|_{\widehat{\boldsymbol{b}}_i} \left.\left[\frac{\partial \boldsymbol{f}_{i1}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{i1})}{\partial \boldsymbol{b}_{i1}^T} \quad \cdots \quad \frac{\partial \boldsymbol{f}_{iM_i}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{iM_i})}{\partial \boldsymbol{b}_{iM_i}^T}\right]\right|_{\widehat{\boldsymbol{b}}_i^{\text{aug}}}, \\ \boldsymbol{G}_3 &= \bigoplus_{j=1}^{M_i} \left\{ \left.\frac{\partial \boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij})}{\partial \boldsymbol{b}_{ij}}\right|_{\widehat{\boldsymbol{b}}_i, \widehat{\boldsymbol{b}}_{ij}} \left.\frac{\partial \boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij})}{\partial \boldsymbol{b}_{ij}^T}\right|_{\widehat{\boldsymbol{b}}_i, \widehat{\boldsymbol{b}}_{ij}} + \boldsymbol{\Delta}_2^T \boldsymbol{\Delta}_2 \right\}, \\ \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i^{\text{aug}}) &= \begin{bmatrix} \boldsymbol{f}_{i1}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{i1}) \\ \vdots \\ \boldsymbol{f}_{iM_i}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{iM_i}) \end{bmatrix}. \end{aligned}$$

The modified, profiled Laplacian approximation to the log-likelihood of the two-level NLME model is then given by

$$\ell_{\text{LAp}}(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) = -\frac{N}{2}\left[1 + \log(2\pi) + \log\left(\widehat{\sigma}^2\right)\right] + M \log|\boldsymbol{\Delta}_1| \\ + \sum_{i=1}^{M} M_i \log|\boldsymbol{\Delta}_2| - \frac{1}{2}\sum_{i=1}^{M} \log|\boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i)|,$$

where $\widehat{\sigma}^2 = \sum_{i=1}^{M} g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i^{\text{aug}}\right)/N$. This formulation can be extended to multilevel NLME models with an arbitrary number of levels.

The Laplacian approximation generally gives more accurate estimates than the alternating algorithm, as it uses an expansion around the estimated random effects only, while the LME approximation in the alternating algorithm uses an expansion around the estimated fixed and random effects. Because it requires solving a different penalized nonlinear least-squares problem for each group in the data and its objective function cannot be profiled on the fixed effects, the Laplacian approximation is more computationally intensive than the alternating algorithm. The algorithm for calculating the Laplacian approximation can be easily parallelized, because the individual PNLS problems are independently optimized.

Adaptive Gaussian Approximation

Gaussian quadrature rules are used to approximate integrals of functions with respect to a given kernel by a weighted average of the integrand evaluated at predetermined abscissas. The weights and abscissas used in Gaussian quadrature rules for the most common kernels can be obtained from the tables of Abramowitz and Stegun (1964) or by using an algorithm proposed by Golub (1973) (see also Golub and Welsch (1969)). Gaussian quadrature rules for multiple integrals are known to be numerically complex (Davis and Rabinowitz, 1984), but using the structure of the integrand in the nonlinear mixed-effects model we can transform the problem into successive applications of simple one-dimensional Gaussian quadrature rules. We consider initially the single-level NLME model.

A natural candidate for the kernel function for the quadrature rule in the single-level NLME model is the marginal distribution of the random effects, that is, the $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ density. The Gaussian quadrature rule in this case can be viewed as a deterministic version of a Monte Carlo integration algorithm in which random samples of the random effects, $\boldsymbol{b}_i$, are generated from the $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ distribution. The samples and the weights in the Gaussian quadrature rule are fixed beforehand, while in Monte Carlo integration they are left to random choice. Because importance sampling tends to be much more efficient than simple Monte Carlo integration (Geweke, 1989), we consider an importance sample version of the Gaussian quadrature rule, which we denote by *adaptive Gaussian* quadrature.

The critical step for the success of importance sampling is the choice of an importance distribution that approximates the integrand. For the single-level NLME model the integrand is proportional to

$$\exp\left[-g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i\right)/2\sigma^2\right],$$

which is approximated by a $\mathcal{N}(\widehat{\boldsymbol{b}}_i, \sigma^2 \boldsymbol{G}^{-1}(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i))$ density, with $\widehat{\boldsymbol{b}}_i$ defined as in (7.16) and $\boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i)$ defined as in (7.18). This is the importance distribution used in the adaptive Gaussian quadrature, so that the grid of abscissas in the $\boldsymbol{b}_i$ scale is centered around the conditional modes $\widehat{\boldsymbol{b}}_i$ and $\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)$ is used for scaling. Letting z_j, w_j $j = 1, \ldots, N_{GQ}$ denote, respectively, the abscissas and the weights for the (one-dimensional) Gaussian quadrature rule with N_{GQ} points based on the $\mathcal{N}(0, 1)$ kernel, the adaptive Gaussian quadrature rule is given by

$$\begin{aligned}
\int \exp\left[-g\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \boldsymbol{b}_i\right)/2\sigma^2\right] d\boldsymbol{b}_i = \int & \sigma^q \left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right|^{-1/2} \\
& \times \exp\left\{-g\left[\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i + \sigma \boldsymbol{G}^{-1/2}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\boldsymbol{z}\right]/2\sigma^2 + \left\|\boldsymbol{z}\right\|^2/2\right\} \\
& \times \exp\left(-\left\|\boldsymbol{z}\right\|^2/2\right) d\boldsymbol{z} \\
& \simeq \sigma^q \left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right|^{-1/2} \\
& \times \sum_{j_1=1}^{N_{GQ}} \cdots \sum_{j_q=1}^{N_{GQ}} \exp\left(-g\left\{\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i + \sigma \boldsymbol{G}^{-1/2}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\boldsymbol{z}_j\right\}/2\sigma^2\right. \\
& \left. + \left\|\boldsymbol{z}_j\right\|^2/2\right) \prod_{k=1}^{q} w_{j_k},
\end{aligned}$$

where $\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right]^{1/2}$ denotes a square root of $\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)$ and $\boldsymbol{z}_j = \left(z_{j_1}, \ldots, z_{j_q}\right)^T$.

The adaptive Gaussian approximation to the log-likelihood function in the single-level NLME model is then

$$\begin{aligned}
\ell_{AGQ}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta} \mid \boldsymbol{y}\right) = & -\tfrac{N}{2}\log\left(2\pi\sigma^2\right) + M \log\left|\boldsymbol{\Delta}\right| - \tfrac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\right| \\
+ & \sum_{i=1}^{M}\log\left(\sum_{j}^{N_{GQ}}\exp\left\{-g\left[\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i + \sigma \boldsymbol{G}^{-1/2}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{y}_i\right)\boldsymbol{z}_j\right]/2\sigma^2\right.\right. \\
+ & \left.\left.\left\|\boldsymbol{z}_j\right\|^2/2\right\}\prod_{k=1}^{q} w_{j_k}\right).
\end{aligned}$$

The one point (i.e., $N_{GQ} = 1$) adaptive Gaussian quadrature approximation is simply the modified Laplacian approximation (7.19), because in this case $z_1 = 0$ and $w_1 = 1$. The adaptive Gaussian quadrature also gives the exact log-likelihood when the model function f is linear in the random effects $\boldsymbol{b}$.

The adaptive Gaussian approximation can be made arbitrarily accurate by increasing the number of abscissas, N_{GQ}. However, because N_{GQ}^q grid points are used to calculate the adaptive Gaussian quadrature for each group, it quickly becomes prohibitively computationally intensive as the number of abscissas increases. In practice $N_{GQ} \leq 7$ generally suffices and $N_{GQ} = 1$ often provides a reasonable approximation (Pinheiro and Bates, 1995).

The adaptive Gaussian approximation can be generalized to multilevel NLME models, using the same steps as in the multilevel Laplacian approximation. For example, the adaptive Gaussian approximation to the log-likelihood of a two-level NLME model is

$$\begin{aligned}
\ell_{AGQ}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \mid \boldsymbol{y}\right) = & -\frac{N}{2}\log\left(2\pi\sigma^2\right) + M \log\left|\boldsymbol{\Delta}_1\right| + \sum_{i=1}^{M} M_i \log\left|\boldsymbol{\Delta}_2\right| \\
& -\frac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i\right)\right| + \sum_{i=1}^{M}\log\left[\sum_{j}^{N_{GQ}}\left(\exp\left\{-g\left[\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i,\right.\right.\right.\right. \\
& \left.\left.\left.\left.\widehat{\boldsymbol{b}}_i^{\text{aug}} + \sigma \boldsymbol{G}^{-1/2}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{y}_i\right)\boldsymbol{z}_j\right]/2\sigma^2 + \left\|\boldsymbol{z}_j\right\|^2/2\right\}\prod_{k=1}^{q_1+M_i q_2} w_{j_k}\right)\right],
\end{aligned}$$

where $\boldsymbol{z}_j = \left(z_{j_1}, \ldots, z_{j_{q_1+Mq_2}}\right)^T$. In this case, the number of grid points for the ith first-level group is $N_{GQ}^{q_1+M_i q_2}$, so that the computational complexity of the calculations increases exponentially with the number of second-level groups. This formulation can be extended to multilevel NLME models with arbitrary number of levels.

7.2.2 Inference and Predictions

Because the alternating algorithm of Lindstrom and Bates (1990) is the only estimation algorithm implemented in the `nlme` function, we restrict ourselves for the rest of this section to inference and predictions using this estimation algorithm. The results described here can be easily extended to other likelihood estimation approaches, such as the Laplacian and adaptive Gaussian approximations described in §7.2.1.

Inference

Inference on the parameters of an NLME model estimated via the alternating algorithm is based on the LME approximation to the log-likelihood function, defined in §7.2.1. Using this approximation at the estimated values of the parameters and the asymptotic results for LME models described in §2.3 we obtain standard errors, confidence intervals, and hypothesis tests for the parameters in the NLME model. We use the single-level NLME model of §7.1.1 to illustrate the inference results derived from the LME approximation to the log-likelihood. Extensions to multilevel NLME models are straightforward.

Under the LME approximation, the distribution of the (restricted) maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$ of the fixed effects is

$$\widehat{\boldsymbol{\beta}} \stackrel{.}{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left[\sum_{i=1}^{M} \widehat{\boldsymbol{X}}_i^T \boldsymbol{\Sigma}_i^{-1} \widehat{\boldsymbol{X}}_i\right]^{-1}\right), \tag{7.20}$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{I} + \widehat{\boldsymbol{Z}}_i \boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^{-T}\widehat{\boldsymbol{Z}}_i^T$, with $\widehat{\boldsymbol{X}}_i$ and $\widehat{\boldsymbol{Z}}_i$ are defined as in (7.11). The standard errors included in the `summary` method for nlme objects are obtained from the approximate variance–covariance matrix in (7.20). The t and F tests reported in the `summary` method and in the `anova` method with a single argument are also based on (7.20). The degrees-of-freedom for t and F tests are calculated as described in §2.4.2.

Now let $\boldsymbol{\theta}$ denote an unconstrained set of parameters that determine the precision factor $\boldsymbol{\Delta}$. The LME approximation is also used to provide an approximate distribution for the (RE)ML estimators $(\widehat{\boldsymbol{\theta}}, \log\widehat{\sigma})^T$. We use $\log\sigma$ in place of σ^2 to give an unrestricted parameterization for which the normal approximation tends to be more accurate.

$$\begin{bmatrix}\widehat{\boldsymbol{\theta}}\\ \log\widehat{\sigma}\end{bmatrix} \stackrel{.}{\sim} \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}\\ \log\sigma\end{bmatrix}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta},\sigma)\right),$$
$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\sigma) = -\begin{bmatrix}\partial^2 \ell_{\text{LMEp}}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T & \partial^2 \ell_{\text{LMEp}}/\partial\log\sigma\partial\boldsymbol{\theta}^T\\ \partial^2 \ell_{\text{LMEp}}/\partial\boldsymbol{\theta}\partial\log\sigma & \partial^2 \ell_{\text{LMEp}}/\partial^2\log\sigma\end{bmatrix}, \tag{7.21}$$

where $\ell_{\text{LMEp}} = \ell_{\text{LMEp}}(\boldsymbol{\Delta},\sigma)$ denotes the LME approximation to the log-likelihood, profiled on the fixed effects, and $\boldsymbol{\mathcal{I}}$ denotes the empirical information matrix. The same approximate distribution is valid for the REML estimators with ℓ_{LMEp} replaced by the log-restricted-likelihood ℓ_{LME}^R defined in (7.13).

In practice, $\boldsymbol{\Delta}$ and σ^2 are replaced by their respective (RE)ML estimates in the expressions for the approximate variance–covariance matrices in (7.20) and (7.21). The approximate distributions for the (RE)ML estimators are used to produce the confidence intervals reported in the `intervals` method for nlme objects.

The LME approximate log-likelihood is also used to compare nested NLME models through likelihood ratio tests, as described in §2.4.1. In the case of REML estimation, only models with identical fixed and random-effects structures can be compared, because the $\widehat{\boldsymbol{X}}_i$ matrices depend on both $\widehat{\boldsymbol{\beta}}$ and the $\widehat{\boldsymbol{b}}_i$. The same recommendations stated in §2.4.1, on the use of likelihood ratio tests for comparing LME models, remain valid for likelihood ratio tests (based on the LME approximate log-likelihood) for comparing NLME models. Hypotheses on the fixed effects should be tested using t and F tests, because likelihood ratio tests tend to be "anticonservative." Likelihood ratio tests for variance–covariance parameters tend to be somewhat conservative, but are generally used to compare NLME models with nested random effects structures. Information criterion statistics, for example, AIC and BIC, based on the LME approximate log-likelihood are also used to compare NLME models.

The inference results for NLME models based on the LME approximation to the log-likelihood are "approximately asymptotic," making them less reliable than the asymptotic inference results for LME models described in §2.3.

Predictions

As with LME models, fitted values and predictions for NLME models may be obtained at different levels of nesting, or at the population level. Population-level predictions estimate the expected response when the random effects are equal to their mean value, $\boldsymbol{0}$. For example, letting $\boldsymbol{x}_h$ represent a vector of fixed-effects covariates and $\boldsymbol{v}_h$ a vector of other model covariates, the population prediction for the corresponding response y_h estimates $f(\boldsymbol{x}_h^T\boldsymbol{\beta}, \boldsymbol{v}_h)$.

Predicted values at the kth level of nesting estimate the conditional expectation of the response given the random effects at levels $\leq k$ and with the random effects at higher levels of nesting set to zero. For example, letting $\boldsymbol{z}_h(i)$ denote a vector of covariates corresponding to random effects associated with the ith group at the first level of nesting, the *level-1* predictions estimate $f(\boldsymbol{x}_h^T\boldsymbol{\beta} + \boldsymbol{z}_h(i)^T\boldsymbol{b}_i, \boldsymbol{v}_h)$. Similarly, letting $\boldsymbol{z}_h(i,j)$ denote a covariate vector associated with the jth level-2 group within the ith level-1 group, the *level-2* predicted values estimate $f(\boldsymbol{x}_h^T\boldsymbol{\beta} + \boldsymbol{z}_h(i)^T\boldsymbol{b}_i + \boldsymbol{z}_h(i,j)^T\boldsymbol{b}_{ij}, \boldsymbol{v}_h)$. This extends naturally to an arbitrary level of nesting.

The (RE)ML estimates of the fixed effects and the conditional modes of the random effects, which are estimated Best Linear Unbiased Estimates (*BLUPs*) of the random effects in the LME approximate log-likelihood, are used to obtain predicted values for the response. For example, the population, level-1, and level-2 predictions for y_h are

$$\begin{aligned}
\widehat{y}_h &= f(\boldsymbol{x}_h^T\widehat{\boldsymbol{\beta}}, \boldsymbol{v}_h),\\
\widehat{y}_h(i) &= f(\boldsymbol{x}_h^T\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_h(i)^T\widehat{\boldsymbol{b}}_i, \boldsymbol{v}_h),\\
\widehat{y}_h(i,j) &= f(\boldsymbol{x}_h^T\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_h(i)^T\widehat{\boldsymbol{b}}_i + \boldsymbol{z}_h(i,j)^T\widehat{\boldsymbol{b}}_{i,j}, \boldsymbol{v}_h).
\end{aligned}$$

7.3 Computational Methods

In this section we describe efficient computational methods for estimating the parameters in an NLME model using the alternating algorithm presented in §7.2.1. The LME step of the alternating algorithm consists of optimizing a linear mixed-effects log-likelihood, or log-restricted-likelihood, for which efficient computational algorithms are discussed in §2.2.8. Therefore, we concentrate here on computational methods for the PNLS step of the alternating algorithm, focusing initially on the single-level NLME model.

The objective function optimized in the PNLS step of the alternating algorithm for a single-level NLME model is the penalized sum of squares

$$\sum_{i=1}^{M}\left[\|\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2\right]. \tag{7.22}$$

By adding "pseudo" observations to the data, we can convert (7.22) into a simple nonlinear sum of squares. Define the augmented response and model function vectors

$$\tilde{\boldsymbol{y}}_i = \begin{bmatrix}\boldsymbol{y}_i\\ \boldsymbol{0}\end{bmatrix}, \qquad \tilde{\boldsymbol{f}}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) = \begin{bmatrix}\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\\ \boldsymbol{\Delta}\boldsymbol{b}_i\end{bmatrix}.$$

The penalized sum of squares (7.22) can then be re-expressed as

$$\sum_{i=1}^{M}\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{f}}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\|^2. \tag{7.23}$$

It follows from (7.23) that, conditional on $\boldsymbol{\Delta}$, the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ in the PNLS step can be regarded as a standard nonlinear least-squares problem. A common iterative estimation method for standard nonlinear least-squares problems is the Gauss–Newton method (Bates and Watts, 1988, §2.2) wherein a nonlinear model $\boldsymbol{f}(\boldsymbol{\alpha})$ is replaced by a first-order Taylor series approximation about current estimates $\widehat{\boldsymbol{\alpha}}^{(w)}$ as

$$\boldsymbol{f}(\boldsymbol{\alpha}) \approx \boldsymbol{f}(\widehat{\boldsymbol{\alpha}}^{(w)}) + \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\alpha}^T}\right|_{\widehat{\boldsymbol{\alpha}}^{(w)}}\left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{(w)}\right).$$

The parameter increment $\widehat{\boldsymbol{\delta}}^{(w+1)} = \widehat{\boldsymbol{\alpha}}^{(w+1)} - \widehat{\boldsymbol{\alpha}}^{(w)}$ for the wth iteration is calculated as the solution of the least-squares problem

$$\left\|\left[\boldsymbol{y} - \boldsymbol{f}(\widehat{\boldsymbol{\alpha}}^{(w)})\right] - \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\alpha}^T}\right|_{\widehat{\boldsymbol{\alpha}}^{(w)}}\left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{(w)}\right)\right\|^2.$$

Step-halving is used at each Gauss–Newton iteration to ensure that the updated parameter estimates result in a decrease of the objective function. That is, the new estimate is set to $\widehat{\boldsymbol{\alpha}}^{(w)} + \widehat{\boldsymbol{\delta}}^{(w+1)}$ and the corresponding value of the objective function is calculated. If it is less than the value at $\widehat{\boldsymbol{\alpha}}^{(w)}$, the value is retained and the algorithm proceed to the next step, or declares convergence. Otherwise, the new estimate is set to $\widehat{\boldsymbol{\alpha}}^{(w)} + \widehat{\boldsymbol{\delta}}^{(w+1)}/2$ and the procedure is repeated, with the increment being halved until a decrease in the objective function is observed or some predetermined minimum step size is reached.

The Gauss–Newton algorithm is used to estimate $\boldsymbol{\beta}$ and the $\boldsymbol{b}_i$ in the PNLS step of the alternating algorithm. Because of the "loosely coupled" structure of the PNLS problem (Soo and Bates, 1992), efficient nonlinear least-squares algorithms can be employed.

The derivative matrices for the Gauss–Newton optimization of (7.23) are, for $i = 1, \ldots, M$,

$$\begin{aligned}
\left.\frac{\partial \tilde{\boldsymbol{f}}_i(\boldsymbol{\beta}, \boldsymbol{b}_i|\boldsymbol{\Delta})}{\partial \boldsymbol{\beta}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}} &= \tilde{\boldsymbol{X}}_i^{(w)} = \begin{bmatrix}\widehat{\boldsymbol{X}}_i^{(w)}\\ \boldsymbol{0}\end{bmatrix},\\
\left.\frac{\partial \tilde{\boldsymbol{f}}_i(\boldsymbol{\beta}, \boldsymbol{b}_i|\boldsymbol{\Delta})}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}} &= \tilde{\boldsymbol{Z}}_i^{(w)} = \begin{bmatrix}\widehat{\boldsymbol{Z}}_i^{(w)}\\ \boldsymbol{\Delta}\end{bmatrix},
\end{aligned}$$

with $\widehat{\boldsymbol{X}}_i^{(w)}$ and $\widehat{\boldsymbol{Z}}_i^{(w)}$ defined as in (7.11). The least-squares problem to be solved at each Gauss–Newton iteration is

$$\sum_{i=1}^{M}\left\|\left[\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{f}}_i\left(\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}\right)\right] - \tilde{\boldsymbol{X}}_i^{(w)}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(w)}\right) - \tilde{\boldsymbol{Z}}_i^{(w)}\left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i^{(w)}\right)\right\|^2$$

or, equivalently,

$$\sum_{i=1}^{M}\left\|\tilde{\boldsymbol{w}}_i^{(w)} - \tilde{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i^{(w)}\boldsymbol{b}_i\right\|^2, \quad \text{where} \quad \tilde{\boldsymbol{w}}_i^{(w)} = \begin{bmatrix}\widehat{\boldsymbol{w}}_i^{(w)}\\ \boldsymbol{0}\end{bmatrix}, \tag{7.24}$$

with $\widehat{\boldsymbol{w}}^{(w)}$ defined as in (7.11).

We use the same matrix decomposition methods as in §2.2.3 to obtain an efficient algorithm for solving (7.24). Consider first the orthogonal-triangular decomposition

$$\begin{bmatrix}\widehat{\boldsymbol{Z}}_i^{(w)} & \widehat{\boldsymbol{X}}_i^{(w)} & \widehat{\boldsymbol{w}}_i^{(w)}\\ \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_{1(i)}\begin{bmatrix}\boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)}\\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)}\end{bmatrix}, \tag{7.25}$$

where the reduction to triangular form is halted after the first q columns. The numbering scheme used for the components in (7.25) is the same introduced for the LME model in §2.2.3. Because $\boldsymbol{\Delta}$ is assumed to be of full rank, so is the upper-triangular matrix $\boldsymbol{R}_{11(i)}$ in (7.25).

Forming another orthogonal-triangular decomposition

$$\begin{bmatrix}\boldsymbol{R}_{00(1)} & \boldsymbol{c}_{0(1)}\\ \vdots & \vdots\\ \boldsymbol{R}_{00(M)} & \boldsymbol{c}_{0(M)}\end{bmatrix} = \boldsymbol{Q}_0\begin{bmatrix}\boldsymbol{R}_{00} & \boldsymbol{c}_0\\ \boldsymbol{0} & \boldsymbol{c}_{-1}\end{bmatrix} \tag{7.26}$$

and noticing that the $\boldsymbol{Q}_{1(i)}$ and $\boldsymbol{Q}_0$ are orthogonal matrices, we can rewrite (7.24) as

$$\sum_{i=1}^{M}\left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i - \boldsymbol{R}_{10(i)}\boldsymbol{\beta}\right\|^2 + \left\|\boldsymbol{c}_0 - \boldsymbol{R}_{00}\boldsymbol{\beta}\right\|^2 + \left\|\boldsymbol{c}_{-1}\right\|^2. \tag{7.27}$$

We assume that the $\boldsymbol{R}_{00}$ is of full-rank, in which case (7.27) is uniquely minimized by the least-squares estimates

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \boldsymbol{R}_{00}^{-1}\boldsymbol{c}_0,\\ \widehat{\boldsymbol{b}}_i &= \boldsymbol{R}_{11(i)}^{-1}\left(\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}\right), \quad i = 1, \dots, M.\end{aligned} \tag{7.28}$$

The Gauss–Newton increments are then obtained as the difference between the least-squares estimates (7.28) and the current estimates, $\widehat{\boldsymbol{\beta}}^{(w)}$ and $\widehat{\boldsymbol{b}}_i^{(w)}$.

Step-halving is used to ensure that the new estimates result in a decrease of the objective function (7.23).

The efficient Gauss–Newton algorithm described above can be extended to multilevel PNLS optimization problems. For example, in the two-level NLME model, the PNLS step consists in optimizing

$$\sum_{i=1}^{M}\left\{\sum_{j=1}^{M_i}\left[\|\boldsymbol{y}_{ij} - \boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij})\|^2 + \|\boldsymbol{\Delta}_2\boldsymbol{b}_{ij}\|^2\right] + \|\boldsymbol{\Delta}_1\boldsymbol{b}_i\|^2\right\} \tag{7.29}$$

over $\boldsymbol{\beta}$, $\boldsymbol{b}_i$, and $\boldsymbol{b}_{ij}$.

The Gauss–Newton iteration is implemented by solving the least-squares problem

$$\sum_{i=1}^{M}\left\{\sum_{j=1}^{M_i}[\|\widehat{\boldsymbol{w}}_{ij}^{(w)} - \widehat{\boldsymbol{X}}_{ij}^{(w)}\boldsymbol{\beta} - \widehat{\boldsymbol{Z}}_{i,j}^{(w)}\boldsymbol{b}_i - \widehat{\boldsymbol{Z}}_{ij}^{(w)}\boldsymbol{b}_{ij}\|^2 + \|\boldsymbol{\Delta}_2\boldsymbol{b}_{ij}\|^2] + \|\boldsymbol{\Delta}_1\boldsymbol{b}_i\|^2\right\}, \tag{7.30}$$

with $\widehat{\boldsymbol{w}}_{ij}^{(w)}$, $\widehat{\boldsymbol{X}}_{ij}^{(w)}$, $\widehat{\boldsymbol{Z}}_{i,j}^{(w)}$, and $\widehat{\boldsymbol{Z}}_{ij}^{(w)}$ defined as in (7.15). To solve it efficiently, we first consider the orthogonal-triangular decomposition

$$\begin{bmatrix}\widehat{\boldsymbol{Z}}_{ij}^{(w)} & \widehat{\boldsymbol{Z}}_{i,j}^{(w)} & \widehat{\boldsymbol{X}}_{ij}^{(w)} & \widehat{\boldsymbol{w}}_{ij}^{(w)}\\ \boldsymbol{\Delta}_2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_{2(ij)}\begin{bmatrix}\boldsymbol{R}_{22(ij)} & \boldsymbol{R}_{21(ij)} & \boldsymbol{R}_{20(ij)} & \boldsymbol{c}_{2(ij)}\\ \boldsymbol{0} & \boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)}\end{bmatrix},$$

where the reduction to triangular form is halted after the first q_2 columns. Because $\boldsymbol{\Delta}_2$ is assumed of full rank, so is $\boldsymbol{R}_{22(ij)}$. We then form a second orthogonal-triangular decomposition for each first-level group

$$\begin{bmatrix}\boldsymbol{R}_{11(i1)} & \boldsymbol{R}_{10(i1)} & \boldsymbol{c}_{1(i1)}\\ \vdots & \vdots & \vdots\\ \boldsymbol{R}_{11(iM_i)} & \boldsymbol{R}_{10(iM_i)} & \boldsymbol{c}_{1(iM_i)}\\ \boldsymbol{\Delta}_1 & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_{1(i)}\begin{bmatrix}\boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)}\\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)}\end{bmatrix},$$

where the reduction to triangular form is stopped after the first q_1 columns. The $\boldsymbol{\Delta}_1$ matrix is assumed of full rank and, as a result, so is $\boldsymbol{R}_{11(i)}$. A final orthogonal-decomposition, identical to (7.26), is then formed.

Because the matrices $\boldsymbol{Q}_{2(ij)}$, $\boldsymbol{Q}_{1(i)}$, and $\boldsymbol{Q}_0$ are orthogonal, (7.30) can be re-expressed as

$$\begin{aligned}\sum_{i=1}^{M}\left\{\sum_{j=1}^{M_i}\left[\left\|\boldsymbol{c}_{2(ij)} - \boldsymbol{R}_{20(ij)}\boldsymbol{\beta} - \boldsymbol{R}_{21(ij)}\boldsymbol{b}_i - \boldsymbol{R}_{22(ij)}\boldsymbol{b}_{ij}\right\|^2\right]\right.&\\ \left. + \left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\right\|^2\right\} &+ \left\|\boldsymbol{c}_0 - \boldsymbol{R}_{00}\boldsymbol{\beta}\right\|^2,\end{aligned}$$

which is uniquely minimized by the least-squares estimates

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \boldsymbol{R}_{00}^{-1}\boldsymbol{c}_0,\\
\widehat{\boldsymbol{b}}_i &= \boldsymbol{R}_{11(i)}^{-1}\left(\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}\right), \quad i = 1,\dots,M,\\
\widehat{\boldsymbol{b}}_{ij} &= \boldsymbol{R}_{22(ij)}^{-1}(\boldsymbol{c}_{2(ij)} - \boldsymbol{R}_{21(ij)}\widehat{\boldsymbol{b}}_i - \boldsymbol{R}_{20(i)}\widehat{\boldsymbol{\beta}}),\ i = 1,\dots,M,\ j = 1,\dots,M_i.
\end{aligned} \tag{7.31}$$

The Gauss–Newton increments are then obtained as the difference between the least-squares estimates (7.31) and the current estimates $\widehat{\boldsymbol{\beta}}^{(w)}$, $\widehat{\boldsymbol{b}}_i^{(w)}$, and $\widehat{\boldsymbol{b}}_{ij}^{(w)}$, with step-halving used to ensure that the objective function (7.29) decreases. This extends naturally to multilevel NLME models with arbitrary number of levels.

The efficiency of the Gauss–Newton algorithm described in this section derives from the fact that, at each iteration, the orthogonal-triangular decompositions are performed separately for each group and then once for the fixed effects. This allows efficient memory allocation for storing intermediate results and reduces the numerical complexity of the decompositions. Also, the matrix inversions required to calculate the Gauss–Newton increments involve only upper-triangular matrices of small dimension, which are easy to invert. (In fact, although (7.28) and (7.31) are written in terms of matrix inverses, such as $\boldsymbol{R}_{00}^{-1}$, the actual calculation performed is the solution of the triangular system of equations $\boldsymbol{R}_{00}\widehat{\boldsymbol{\beta}} = \boldsymbol{c}_0$, which is even simpler.)

7.4 Extending the Basic NLME Model

The nonlinear mixed-effects model formulation used so far in this chapter allows considerable flexibility in the specification of the random-effects structure, but restricts the within-group errors to be independent and to have constant variance. This *basic* NLME model provides an adequate model for a broad range of applications, but there are many cases in which the within-group errors are *heteroscedastic* (i.e., have unequal variances) or are *correlated* or are both heteroscedastic and correlated.

This section extends the basic NLME model to allow heteroscedastic, correlated within-group errors. We show how the estimation methods of §7.2 and the computational methods of §7.3 can be adapted to the extended model formulation.

7.4.1 General Formulation of the Extended NLME Model

As described in §7.1.1, the basic single-level NLME model (7.3) assumes that the within-group errors $\boldsymbol{\epsilon}_i$ are independent $\mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$ random vectors.

The extended single-level NLME model relaxes this assumption by allowing heteroscedastic and correlated within-group errors, being expressed for $i = 1,\dots,M$ as

$$\begin{gathered}
\boldsymbol{y}_i = \boldsymbol{f}_i\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\phi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,\\
\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_i\right).
\end{gathered} \tag{7.32}$$

The $\boldsymbol{\Lambda}_i$ are positive-definite matrices parametrized by a fixed, generally small, set of parameters $\boldsymbol{\lambda}$. As in the basic NLME model, the within-group errors $\boldsymbol{\epsilon}_i$ are assumed to be independent for different i and to be independent of the random effects $\boldsymbol{b}_i$. The σ^2 is factored out of the $\boldsymbol{\Lambda}_i$ for computational reasons (it can then be eliminated from the profiled likelihood function).

Similarly, the extended two-level NLME model generalizes the basic two-level NLME model (7.7) described in §7.1.2 by letting

$$\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_{ij}\right), \quad i = 1,\dots,M \quad j = 1,\dots,M_i,$$

where the $\boldsymbol{\Lambda}_{ij}$ are positive-definite matrices parametrized by a fixed $\boldsymbol{\lambda}$ vector. This readily generalizes to a multilevel model with Q levels of random effects. For simplicity, we concentrate for the remainder of this section on the extended single-level NLME model (7.32), but the results we obtain are easily generalizable to multilevel models with an arbitrary number of levels of random effects.

As described in §5.1.3, the variance–covariance structure of the within-group errors can be decomposed into two independent components: a *variance* structure and a *correlation* structure. Variance function models to represent the variance structure component of the within-group errors are described and illustrated in §5.2. Correlation models to represent the correlation structure of the within-group errors are presented and have their use illustrated in §5.3. The use of the `nlme` function to fit the extended NLME model is described in §8.3.

7.4.2 Estimation and Computational Methods

Because $\boldsymbol{\Lambda}_i$ is positive-definite, it admits an invertible square-root $\boldsymbol{\Lambda}_i^{1/2}$ (Thisted, 1988, §3), with inverse $\boldsymbol{\Lambda}_i^{-1/2}$, such that

$$\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i^{T/2}\boldsymbol{\Lambda}_i^{1/2} \quad \text{and} \quad \boldsymbol{\Lambda}_i^{-1} = \boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\Lambda}_i^{-T/2}.$$

Letting

$$\begin{aligned}
\boldsymbol{y}_i^* &= \boldsymbol{\Lambda}_i^{-T/2}\boldsymbol{y}_i,\\
\boldsymbol{f}_i^*\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right) &= \boldsymbol{\Lambda}_i^{-T/2}\boldsymbol{f}_i\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right),\\
\boldsymbol{\epsilon}_i^* &= \boldsymbol{\Lambda}_i^{-T/2}\boldsymbol{\epsilon}_i,
\end{aligned} \tag{7.33}$$

and noting that $\boldsymbol{\epsilon}_i^* \sim \mathcal{N}\left[\boldsymbol{\Lambda}_i^{-T/2}\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i^{-T/2}\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_i^{-1/2}\right] = \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}\right)$, we can rewrite (7.32) as

$$\begin{aligned} \boldsymbol{y}_i^* &= \boldsymbol{f}_i^*\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right) + \boldsymbol{\epsilon}_i^*, \\ \boldsymbol{\phi}_i &= \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \\ \boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), &\quad \boldsymbol{\epsilon}_i^* \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}\right). \end{aligned}$$

That is, $\boldsymbol{y}_i^*$ is described by a basic NLME model.

Because the differential of the linear transformation $\boldsymbol{y}_i^* = \boldsymbol{\Lambda}_i^{-T/2}\boldsymbol{y}_i$ is simply $d\boldsymbol{y}_i^* = \left|\boldsymbol{\Lambda}_i\right|^{-1/2} d\boldsymbol{y}_i$, the log-likelihood function corresponding to the extended NLME model (7.32) is expressed as

$$\begin{aligned} \ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}|\boldsymbol{y}\right) &= \sum_{i=1}^{M} \log p\left(\boldsymbol{y}_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}\right) \\ &= \sum_{i=1}^{M} \log p\left(\boldsymbol{y}_i^*|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}\right) - \frac{1}{2}\sum_{i=1}^{M} \log\left|\boldsymbol{\Lambda}_i\right| \\ &= \ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}|\boldsymbol{y}^*\right) - \frac{1}{2}\sum_{i=1}^{M} \log\left|\boldsymbol{\Lambda}_i\right|. \end{aligned}$$

The log-likelihood function $\ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}|\boldsymbol{y}^*\right)$ corresponds to a basic NLME model with model function $\boldsymbol{f}_i^*$ and, therefore, the approximations presented in §7.2.1 can be applied to it. The inference results described in §7.2.2 also remain valid.

Alternating Algorithm

The PNLS step of the alternating algorithm for the extended NLME model consists of minimizing, over $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, $i = 1, \ldots, M$, the penalized nonlinear least-squares function

$$\sum_{i=1}^{M}\left[\left\|\boldsymbol{y}_i^* - \boldsymbol{f}_i^*(\boldsymbol{\beta}, \boldsymbol{b}_i)\right\|^2 + \left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2\right] = \\ \sum_{i=1}^{M}\left\{\left\|\boldsymbol{\Lambda}_i^{-T/2}\left[\boldsymbol{y}_i - \boldsymbol{f}_i\left(\boldsymbol{\beta}, \boldsymbol{b}_i\right)\right]\right\|^2 + \left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2\right\}.$$

The derivative matrices and working vector used in the Gauss–Newton algorithm for the PNLS step and also in the LME step are defined as

$$\begin{aligned} \widehat{\boldsymbol{X}}_i^{*(w)} &= \left.\frac{\partial \boldsymbol{f}_i^*}{\partial \boldsymbol{\beta}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}} = \boldsymbol{\Lambda}_i^{-T/2}\widehat{\boldsymbol{X}}_i^{(w)}, \\ \widehat{\boldsymbol{Z}}_i^{*(w)} &= \left.\frac{\partial \boldsymbol{f}_i^*}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}} = \boldsymbol{\Lambda}_i^{-T/2}\widehat{\boldsymbol{Z}}_i^{(w)}, \\ \widehat{\boldsymbol{w}}_i^{*(w)} &= \boldsymbol{y}_i^* - \boldsymbol{f}_i^*(\widehat{\boldsymbol{\beta}}^{(w)}, \widehat{\boldsymbol{b}}_i^{(w)}) + \widehat{\boldsymbol{X}}_i^{*(w)}\widehat{\boldsymbol{\beta}}^{(w)} + \widehat{\boldsymbol{Z}}_i^{*(w)}\widehat{\boldsymbol{b}}_i^{(w)} = \boldsymbol{\Lambda}_i^{-T/2}\widehat{\boldsymbol{w}}_i^{(w)}, \end{aligned}$$

with $\widehat{\boldsymbol{X}}_i^{(w)}$, $\widehat{\boldsymbol{Z}}_i^{(w)}$, and $\widehat{\boldsymbol{w}}_i^{(w)}$ defined as in (7.11).

The Gauss–Newton algorithm for the PNLS step is identical to the algorithm described in §7.3, with $\widehat{\boldsymbol{X}}_i^{(w)}$, $\widehat{\boldsymbol{Z}}_i^{(w)}$, and $\widehat{\boldsymbol{w}}_i^{(w)}$ replaced, respectively, by $\widehat{\boldsymbol{X}}_i^{*(w)}$, $\widehat{\boldsymbol{Z}}_i^{*(w)}$, and $\widehat{\boldsymbol{w}}_i^{*(w)}$. The LME approximation to the log-likelihood function of the extended single-level NLME model is

$$\ell_{\mathrm{LME}}^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda} \mid \boldsymbol{y}\right) = \ell_{\mathrm{LME}}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda} \mid \boldsymbol{y}^*\right) - \frac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{\Delta}_i\right|,$$

which has the same form as the log-likelihood of the extended single-level LME model described in §5.1. The log-restricted-likelihood for the extended NLME model is similarly defined.

Laplacian and Adaptive Gaussian Approximations

For the extended single-level NLME model, the objective function which is minimized to produce the conditional modes $\widehat{\boldsymbol{b}}_i$ used in the Laplacian and adaptive Gaussian approximations is

$$g^*\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i, \boldsymbol{b}_i\right) = \left\|\boldsymbol{\Lambda}_i^{-T/2}\left[\boldsymbol{y}_i - \boldsymbol{f}_i\left(\boldsymbol{\beta}, \boldsymbol{b}_i\right)\right]\right\|^2 + \left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2.$$

The corresponding approximation to the second-derivative matrix of g^* with respect $\boldsymbol{b}_i$ evaluated at $\widehat{\boldsymbol{b}}_i$ is

$$\left.\frac{\partial\partial^2 g^*(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i} \simeq \boldsymbol{G}^*\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i\right) = \\ \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\right|_{\widehat{\boldsymbol{b}}_i} \boldsymbol{\Lambda}_i^{-1} \left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^T}\right|_{\widehat{\boldsymbol{b}}_i} + \boldsymbol{\Delta}^T\boldsymbol{\Delta}.$$

The modified Laplacian approximation to the log-likelihood of the extended single-level NLME model is then given by

$$\ell_{\mathrm{LA}}^*\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \mid \boldsymbol{y}\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) + M\log\left|\boldsymbol{\Delta}\right| \\ -\tfrac{1}{2}\left\{\sum_{i=1}^{M}\log\left|\boldsymbol{G}^*\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i\right)\right| + \sigma^{-2}\sum_{i=1}^{M} g^*(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i)\right\} - \tfrac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{\Lambda}_i\right|$$

and the adaptive Gaussian approximation is given by

$$\begin{aligned}
&\ell^*_{\mathrm{AGQ}}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \mid \boldsymbol{y}\right) = \\
&- \quad \tfrac{N}{2}\log\left(2\pi\sigma^2\right) + M\log\left|\boldsymbol{\Delta}\right| - \tfrac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{G}^*\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i\right)\right| \\
&+ \quad \sum_{i=1}^{M}\log(\sum_{j}^{N_{GQ}}\exp\{-g^*[\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i + \sigma\left(\boldsymbol{G}^*\right)^{-\frac{1}{2}}\left(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\lambda}, \boldsymbol{y}_i\right)\boldsymbol{z}_j]/2\sigma^2 \\
&+ \quad \left\|\boldsymbol{z}_j\right\|^2/2\}\prod_{k=1}^{q} w_{j_k}) - \frac{1}{2}\sum_{i=1}^{M}\log\left|\boldsymbol{\Lambda}_i\right|.
\end{aligned}$$

The same comments and conclusions presented in §5.2 for the case when the within-group variance function depends on the fixed effects and/or the random effects also apply to the extended NLME model. As in the LME case, to keep the optimization problem feasible, an "iteratively reweighted" scheme is used to approximate the variance function. The fixed and random effects used in the variance function are replaced by their current estimates and held fixed during the log-likelihood optimization. New estimates for the fixed and random effects are then produced and the procedure is repeated until convergence. In the case of the alternating algorithm, the estimates for the fixed and random effects obtained in the PNLS step are used to calculate the variance function weights in the LME step. If the variance function does not depend on either the fixed effects or the random effects, then no approximation is necessary.

7.5 An Extended Nonlinear Regression Model

In many applications of nonlinear regression models to grouped data, one wishes to represent the within-group variance–covariance structure through the $\boldsymbol{\Lambda}_i$ matrices only, avoiding the use of random effects to account for within-group dependence. This results in a simplified version of the extended single-level NLME model (7.32), which we call the *extended nonlinear regression* model. In this section, we present the general formulation of the extended nonlinear regression model, describe methods for estimating its parameters, and present computational algorithms for implementing such estimation methods.

The modeling function `gnls` in the nlme library fits the extended nonlinear regression model using maximum likelihood. The use of this function is described and illustrated in §8.3.3.

7.5.1 *General Model Formulation*

The extended nonlinear regression model for the jth observation on the ith group, y_{ij}, is

$$\begin{aligned}
y_{ij} &= f\left(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}\right) + \epsilon_{ij}, \qquad i = 1, \ldots, M,\ j = 1, \ldots, n_i, \\
\boldsymbol{\phi}_{ij} &= \boldsymbol{A}_{ij}\boldsymbol{\beta}.
\end{aligned} \tag{7.34}$$

The real-valued function f depends on a group-specific parameter vector $\boldsymbol{\phi}_{ij}$ and a known covariate vector $\boldsymbol{v}_{ij}$. It is nonlinear in at least one component of $\boldsymbol{\phi}_{ij}$ and differentiable with respect to the group-specific parameters. M is the number of groups, n_i is the number of observations on the ith group, and ϵ_{ij} is a normally distributed error term.

The extended nonlinear regression model (7.34) is a two-stage model in which the second stage expresses the group-specific parameters $\boldsymbol{\phi}_{ij}$ as a linear function of a fixed set of parameters $\boldsymbol{\beta}$. The design matrices $\boldsymbol{A}_{ij}$ are known. We note that the coefficients $\boldsymbol{\beta}$ could be incorporated directly into the model function f, thus eliminating the need of a second stage in the model. However, there are advantages in having the second stage in (7.34), some of which are (i) group-specific parameters generally have a more natural interpretation in the model, (ii) inclusion and elimination of covariates in the model can be done at the second-stage model only, facilitating the understanding of the model building process, and (iii) because the $\boldsymbol{\phi}_{ij}$ are linear functions of the $\boldsymbol{\beta}$, derivatives with respect to $\boldsymbol{\phi}_{ij}$ are easily obtained from derivatives with respect to $\boldsymbol{\beta}$.

Using the same definitions of vectors and matrices given in (7.4), we can express the extended nonlinear regression in matrix form as

$$\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{f}_i\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right) + \boldsymbol{\epsilon}_i, \\
\boldsymbol{\phi}_i &= \boldsymbol{A}_i\boldsymbol{\beta}, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_i\right).
\end{aligned}$$

As in the extended NLME model of §7.4, the $\boldsymbol{\Lambda}_i$ matrices are determined by a fixed, generally small, set of parameters $\boldsymbol{\lambda}$.

Estimation and inference under this model has been studied extensively in the nonlinear regression literature (Carroll and Ruppert, 1988; Seber and Wild, 1989), usually assuming that the $\boldsymbol{\Lambda}_i$ matrices are known, being referred to as the generalized least-squares model (Seber and Wild, 1989, §2.1.4). We refer to it as the *generalized nonlinear least-squares* (GNLS) model to differentiate from the extended linear model described in §5.1.2.

Using the same transformations described in (7.33), the GNLS model (7.34) can be re-expressed as a "classic" nonlinear regression model:

$$\begin{aligned}
\boldsymbol{y}^*_i &= \boldsymbol{f}^*_i\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right) + \boldsymbol{\epsilon}^*_i, \\
\boldsymbol{\phi}_i &= \boldsymbol{A}_i\boldsymbol{\beta}, \quad \boldsymbol{\epsilon}^*_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right).
\end{aligned}$$

7.5.2 Estimation and Computational Methods

Different estimation methods have been proposed for the parameters in the GNLS model (Davidian and Giltinan, 1995, §2.5). We concentrate here on maximum likelihood estimation.

The log-likelihood function for the GNLS model is

$$\ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \middle| \boldsymbol{y}\right) = -\frac{1}{2}\left\{N \log\left(2\pi\sigma^2\right) + \sum_{i=1}^{M}\left[\frac{\left\|\boldsymbol{y}_i^* - \boldsymbol{f}_i^*\left(\boldsymbol{\beta}\right)\right\|^2}{\sigma^2} + \log\left|\boldsymbol{\Lambda}_i\right|\right]\right\}, \tag{7.35}$$

where N represents the total number of observations and, for simplicity, we use $\boldsymbol{f}_i^*\left(\boldsymbol{\beta}\right) = \boldsymbol{f}_i^*\left(\boldsymbol{\phi}_i, \boldsymbol{v}_i\right)$.

For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma}^2(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{M}\left\|\boldsymbol{y}_i^* - \boldsymbol{f}_i^*\left(\boldsymbol{\beta}\right)\right\|^2/N, \tag{7.36}$$

so that the profiled log-likelihood, obtained by replacing σ^2 with $\widehat{\sigma}^2(\boldsymbol{\beta}, \lambda)$ in (7.35), is

$$\ell\left(\boldsymbol{\beta}, \boldsymbol{\lambda} \middle| \boldsymbol{y}\right) = -\frac{1}{2}\left\{N\left[\log\left(2\pi/N\right) + 1\right] + \log\left(\sum_{i=1}^{M}\left\|\boldsymbol{y}_i^* - \boldsymbol{f}_i^*\left(\boldsymbol{\beta}\right)\right\|^2\right)\right. \\ \left. + \sum_{i=1}^{M}\log\left|\boldsymbol{\Lambda}_i\right|\right\}. \tag{7.37}$$

A Gauss–Seidel algorithm (Thisted, 1988, §3.11.2) is used with the profiled log-likelihood (7.37) to obtain the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Given the current estimate $\widehat{\boldsymbol{\lambda}}^{(w)}$ of $\boldsymbol{\lambda}$, a new estimate $\widehat{\boldsymbol{\beta}}^{(w)}$ for $\boldsymbol{\beta}$ is produced by maximizing $\ell(\boldsymbol{\beta}, \widehat{\boldsymbol{\lambda}}^{(w)})$. The roles are then reversed and a new estimate $\boldsymbol{\lambda}^{(k+1)}$ is produced by maximizing $\ell(\widehat{\boldsymbol{\beta}}^{(w)}, \boldsymbol{\lambda})$. The procedure iterates between the two optimizations until a convergence criterion is met.

It follows from (7.37) that, conditional on $\boldsymbol{\lambda}$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by solving an ordinary nonlinear least-squares problem

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{M}\left\|\boldsymbol{y}_i^* - \boldsymbol{f}_i^*\left(\boldsymbol{\beta}\right)\right\|^2,$$

for which we can use a standard Gauss–Newton algorithm. If k is the iteration counter for this algorithm and $\widehat{\boldsymbol{\beta}}^{(k)} = \widehat{\boldsymbol{\beta}}^{(k)}(\boldsymbol{\lambda}^{(w)})$ is the current estimate of $\boldsymbol{\beta}$, then the derivative matrices

$$\widehat{\boldsymbol{X}}_i^{(k)} = \left.\frac{\partial \boldsymbol{f}_i\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}^T}\right|_{\widehat{\boldsymbol{\beta}}^{(k)}}, \qquad \widehat{\boldsymbol{X}}_i^{*(k)} = \boldsymbol{\Lambda}_i^{-T/2}\widehat{\boldsymbol{X}}_i^{(k)},$$

provide the Gauss–Newton increment $\widehat{\boldsymbol{\delta}}^{(k+1)}$ for $\boldsymbol{\beta}$ as the (ordinary) least-squares solution of

$$\sum_{i=1}^{M}\left\|\widehat{\boldsymbol{w}}_i^{*(k)} - \widehat{\boldsymbol{X}}_i^{*(k)}\boldsymbol{\delta}\right\|^2,$$

where $\widehat{\boldsymbol{w}}_i^{*(k)} = \boldsymbol{y}_i^* - \boldsymbol{f}_i^*(\widehat{\boldsymbol{\beta}}^{(k)})$. Orthogonal-triangular decomposition methods similar to the ones described in §7.3 can be used to obtain a compact and numerically efficient implementation of the Gauss–Newton algorithm for estimating $\boldsymbol{\beta}$. The derivation is left to the reader as an exercise.

Inference on the parameters of the GNLS model generally relies on "classical" asymptotic theory for maximum likelihood estimation (Cox and Hinkley, 1974, §9.2), which states that, for large N, the MLEs are approximately normally distributed with mean equal to the true parameter values and variance–covariance matrix given by the inverse of the information matrix. Because $\mathrm{E}[\partial^2\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda})/\partial\boldsymbol{\beta}\partial\boldsymbol{\lambda}^T] = \boldsymbol{0}$ and $\mathrm{E}[\partial^2\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda})/\partial\boldsymbol{\beta}\partial\sigma^2] = \boldsymbol{0}$, the expected information matrix for the GNLS likelihood is block-diagonal and the MLE of $\boldsymbol{\beta}$ is asymptotically uncorrelated with the MLEs of $\boldsymbol{\lambda}$ and σ^2.

The approximate distributions for the MLEs in the GNLS model which are used for constructing confidence intervals and hypothesis tests are

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &\dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left[\sum_{i=1}^{M}\widehat{\boldsymbol{X}}_i^T\boldsymbol{\Lambda}_i^{-1}\widehat{\boldsymbol{X}}_i\right]^{-1}\right), \\ \begin{bmatrix}\widehat{\boldsymbol{\lambda}} \\ \log\widehat{\sigma}\end{bmatrix} &\dot{\sim} \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\lambda} \\ \log\sigma\end{bmatrix}, \boldsymbol{\mathcal{I}}^{-1}\left(\boldsymbol{\lambda}, \sigma\right)\right), \\ \boldsymbol{\mathcal{I}}\left(\boldsymbol{\lambda}, \sigma\right) &= -\begin{bmatrix}\partial^2\ell/\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^T & \partial^2\ell/\partial\log\sigma\partial\boldsymbol{\lambda}^T \\ \partial^2\ell/\partial\boldsymbol{\lambda}\partial\log\sigma & \partial^2\ell/\partial^2\log\sigma\end{bmatrix}, \end{aligned} \tag{7.38}$$

where $\widehat{\boldsymbol{X}}_i$ is the derivative matrix evaluated at the true parameter values. As in §7.2.2, $\log\sigma$ is used in place of σ^2 to give an unrestricted parameterization for which the normal approximation tends to be more accurate. In practice, the parameters in the approximate variance–covariance matrices in (7.38) are replaced by their respective MLEs.

To reduce the bias associated with the maximum likelihood estimation of σ^2, the following modified version of (7.36) is used,

$$\tilde{\sigma}^2 = \sum_{i=1}^{M}\left\|\widehat{\boldsymbol{\Lambda}}_i^{-T/2}\left[\boldsymbol{y}_i - \boldsymbol{f}_i\left(\widehat{\boldsymbol{\beta}}\right)\right]\right\|^2/(N-p),$$

with p denoting the length of $\boldsymbol{\beta}$. $(N-p)\tilde{\sigma}^2$ is approximately distributed as a $\sigma^2\chi^2_{N-p}$ random variable and is asymptotically independent of $\widehat{\boldsymbol{\beta}}$. This is used to produce approximate t and F tests for the coefficients $\boldsymbol{\beta}$. These tests tend to have better small sample properties than the tests obtained from the normal approximation (7.38) alone.

7.6 Chapter Summary

This chapter presents the theoretical foundations of the nonlinear mixed-effects model for single- and multilevel grouped data, including the general model formulation and its underlying distributional assumptions. Efficient computational methods for maximum likelihood estimation in the NLME model are described and discussed. Different approximations to the NLME model log-likelihood with varying degrees of accuracy and computational complexity are derived.

The basic NLME model with independent, homoscedastic within-group errors is extended to allow correlated, heteroscedastic within-group errors and efficient computational methods are described for maximum likelihood estimation of its parameters.

An extended class of nonlinear regression models, with correlated and heteroscedastic errors, but with no random effects, is presented. An efficient maximum likelihood estimation algorithm is described and approximate inference results for the parameters in this extended nonlinear regression are presented.