ABSTRACT

HARTFORD, ALAN HUGHES. Computational approaches for maximum likelihood estimation for nonlinear mixed models. (Under the direction of Marie Davidian and John F. Monahan).

The nonlinear mixed model is an important tool for analyzing pharmacokinetic and other repeated-measures data. In particular, these models are used when the measured response for an individual, $y_i$, has a nonlinear relationship with unknown, random, individual-specific parameters, $\beta_i$. Ideally, the method of maximum likelihood is used to find estimates for the parameters of the model after integrating out the random effects in the conditional likelihood. However, closed form solutions to the integral are generally not available. As a result, methods have been previously developed to find approximate maximum likelihood estimates for the parameters in the nonlinear mixed model. These approximate methods include First Order linearization, Laplace’s approximation, importance sampling, and Gaussian quadrature. The methods are available today in several software packages for models of limited sophistication; constant conditional error variance is required for proper utilization of most software. In addition, distributional assumptions are needed. This work investigates how robust two of these methods, First Order linearization and Laplace’s approximation, are to these assumptions. The finding is that Laplace’s approximation performs well, resulting in better estimation than first order linearization when both models
A method must provide good estimates of the likelihood at points in the parameter space near the solution. This work compares this ability among the numerical integration techniques, Gaussian quadrature, importance sampling, and Laplace’s approximation. A new “scaled” and “centered” version of Gaussian quadrature is found to be the most accurate technique. In addition, the technique requires evaluation of the integrand at only a few abscissas. Laplace’s method also performs well; it is more accurate than importance sampling with even 100 importance samples over two dimensions. Even so, Laplace’s method still does not perform as well as Gaussian quadrature. Overall, Laplace’s approximation performs better than expected, and is shown to be a reliable method while still computationally less demanding.

This work also introduces a new method to maximize the likelihood. This method can be sharpened to any desired level of accuracy. Stochastic approximation is incorporated to continue sampling until enough information is gathered to result in accurate estimation. This new method is shown to work well for linear mixed models, but is not yet successful for the nonlinear mixed model.
COMPUTATIONAL APPROACHES FOR MAXIMUM LIKELIHOOD ESTIMATION FOR NONLINEAR MIXED MODELS

by

ALAN HUGHES HARTFORD

A dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy

in

STATISTICS

in the

GRADUATE SCHOOL
at
NORTH CAROLINA STATE UNIVERSITY
2000

Professor M Davidian
Co-chair of Advisory Committee

Professor P Gremaud

Professor C Savage

Professor J Monahan
Co-chair of Advisory Committee

Professor S Pantula

Sassy J. Pantula
To my brother and sisters,

Tim, Jeanne, and Jill.
Biography

Alan Hartford was born in Omaha, Nebraska, to parents Richard and Ardith Hartford on November 16, 1966. He earned his Bachelor of Science degree in Mathematics with a minor in Physics in May of 1991, a Master of Science degree in Mathematics with a minor in Applied Mathematics in May of 1994, and a Master of Science degree in Biometry in May of 1995, all at the University of Nebraska–Lincoln. He began his Ph.D. studies at North Carolina State University in August of 1995, and his thesis defense was April 3, 2000. After graduating, he will begin a career at Merck Research Laboratories as a Biometrician.
Acknowledgements

Thanks to all of the following people. Without you this would not have been possible.

- To my advisors, Dr. John Monahan and Dr. Marie Davidian, for your instruction and patience.
- To my committee members, Professors Sastry Pantula, Pierre Gremaud, and Carla Savage, for actually reading all of this.
- To Jenny Langdon who has been a constant, unwaivering support.
- To the friends I have made at NCSU and UNL: Erin Blankenship, Ann Oberg, Sarah Hardy, Jennifer Hoff, Dawn Haines, Jennifer Mueller, Valerie Shostrom, Carlos Silva, Anne Parkhurst and most especially, Gordon Brown.
- To Terry Byron and Nalin Dahyabhai for going beyond the call of duty. They made my computing life so much easier.
- To Dr. Carol Gotway, Dr. J.-C. Lu, and Stew Williams for the wonderful research opportunities they shared with me.
- Special thanks to my parents, grandparents, brother, sisters, and the rest of my family for encouraging me along the way.
Contents

List of Figures vii
List of Tables ix

1 Introduction 1
  1.1 Motivation 1
  1.2 Model 3
  1.3 Approximate Methods 4
    1.3.1 First Order Linearization 5
    1.3.2 Laplace’s Approximation 7
    1.3.3 Importance Sampling 11
    1.3.4 Gaussian Quadrature 13
  1.4 Preview of Results 15

2 Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models 17
  2.1 Introduction 19
  2.2 Nonlinear mixed effects model 23
  2.3 Approximate methods 27
  2.4 Design of simulations 34
  2.5 Results 40
  2.6 Discussion 48

3 Comparison of Numerical Integration Methods 67
  3.1 Introduction 67
  3.2 Integration Methods 70
    3.2.1 Laplace’s Approximation 70
    3.2.2 Importance Sampling 74
    3.2.3 Gaussian Quadrature 79
  3.3 Models 81
3.4 Sampling the Parameter Space .................................................. 85
3.5 Results ...................................................................................... 87
  3.5.1 Analysis for Model 1 ............................................................. 90
  3.5.2 Analysis for Model 2 ............................................................. 94
  3.5.3 Analysis for Model 3 ............................................................. 99
  3.5.4 Summary ............................................................................. 101

4 Computational Framework for Maximum Likelihood Estimation 120
  4.1 Introduction ............................................................................. 120
  4.2 The Model ............................................................................. 124
  4.3 Integration Method ................................................................. 125
  4.4 Initial Estimation ..................................................................... 129
  4.5 Optimization via a Heuristic Variation on Stochastic Approximation 132
  4.6 Simulation for a Linear Model ................................................. 134
    4.6.1 Model ............................................................................. 134
    4.6.2 Results ............................................................................. 135
  4.7 Simulation for a NLMM ......................................................... 137
    4.7.1 Model ............................................................................. 137
    4.7.2 Results ............................................................................. 139
  4.8 Discussion .............................................................................. 139

5 Conclusions and Future Work ....................................................... 141

6 Bibliography ................................................................................. 145

A Appendices .................................................................................... 147
  A.1 “Δ”-Method for Finding Approximate Variances ...................... 147
  A.2 Standard Errors for Approximations using Importance Sampling 150
  A.3 Additional Tables for Chapter 2 .............................................. 152
List of Figures

2.1 Perspective plots of the joint density functions corresponding to the asymmetric (A) and bimodal (B) random effects distributions. .......................... 55
2.2 Boxplots of estimates of $\beta_1$ from data sets for which both methods converged in each panel for each random effects distribution, correct inter-individual model fitted. Estimates have been centered and scaled by the true value so that 0 represents no bias and the vertical axis represents bias relative to the truth. Multiplying the numbers on the vertical axis will result in relative percentages. The white bar represents the median, the horizontal bars represent the extremes. In some cases, outliers that distorted the comparison have been deleted; the relative difference of these estimates from the truth is given numerically. .......................... 56
2.3 Same as Fig. 2.2 for $\beta_2$. ............................................................ 57
2.4 Boxplots of estimates of $D_{11}$ from data sets for which both methods converged in each panel for each random effects distribution, correct inter-individual model fitted. Estimates have been centered and scaled as in Fig. 2. The white bar represents the median, the horizontal bars represent actual estimates. In some cases, outliers that distorted the comparison have been deleted; the relative difference of these estimates from the truth is given numerically. ............................................ 58
2.5 Same as Fig. 2.4 for $D_{12}$. ............................................................ 59
2.6 Same as Fig. 2.4 for $D_{22}$. ............................................................ 60
2.7 Same as Fig. 2.4 for $\sigma$. ............................................................ 61
3.1 Model 1, Means of Relative Error for the Laplacian Methods .............. 102
3.2 Model 1, Means of Relative Error for the Importance Sampling Methods103
3.3 Model 1, Means of Relative Error for the Gaussian Quadrature Methods104
3.4 Model 1, Means of Relative Error for the Gaussian Quadrature Methods without $gq$ ............................................................ 105
3.5 Model 1, Means of Relative Error for the Best Methods ...................... 106
3.6 Model 2, $\sigma = 0.15$, Means of Relative Error for the Laplacian Methods 107
3.7 Model 2, $\sigma = 0.15$, Means of Relative Error for the Importance Sampling Methods ................................................................. 108
3.8 Model 2, $\sigma = 0.15$, Means of Relative Error for the Gaussian Quadrature Methods ................................................................. 109
3.9 Model 2, $\sigma = 0.15$, Means of Relative Error for the Best Methods .... 110
3.10 Model 2, $\sigma = 0.30$, Means of Relative Error for the Laplacian Methods 111
3.11 Model 2, $\sigma = 0.30$, Means of Relative Error for the Importance Sampling Methods ................................................................. 112
3.12 Model 2, $\sigma = 0.30$, Means of Relative Error for the Gaussian Quadrature Methods ................................................................. 113
3.13 Model 2, $\sigma = 0.30$, Means of Relative Error for the Best Methods .... 114
3.14 Model 3, Means of Relative Error for the Laplacian Methods ............. 115
3.15 Model 3, Means of Relative Error for the Importance Sampling Methods 116
3.16 Model 3, Means of Relative Error for the Gaussian Quadrature Methods 117
3.17 Model 3, Means of Relative Error for the Best Gaussian Quadrature Methods ................................................................. 118
3.18 Model 3, Means of Relative Error for the Best Methods .................... 119

4.1 Flowchart for Algorithm ................................................................. 123
List of Tables

2.1 Rate of nonconvergence out of 1000 trials for each distribution, sampling scheme, and fitting method, correct inter-individual model.

2.2 Rate of nonconvergence out of 1000 trials for each distribution, sampling scheme, incorrect inter-individual model.

2.3 Relative Efficiency of the First Order method relative to the Laplace method, correct inter-individual model with $\beta_2 = 0$. Each entry is based on all data sets for which convergence was achieved for both methods. Entries are given in italics in cases where both methods performed poorly (see Figs. 2–7).

2.4 Performance of Wald and Likelihood Ratio Tests on trials that converged for both the full and reduced models, correct inter-individual model.

2.5 Performance of Wald and Likelihood Ratio Tests on trials that converged for both the full and reduced models, incorrect inter-individual model.

3.1 Abbreviations for the Methods

3.2 Model 1, mean relative error.

3.3 Analysis of Variance for Model 1

3.4 Model 2 ($\sigma = 0.15$), mean relative error.

3.5 Analysis of Variance for Model 2 with $\sigma = 0.15$

3.6 Model 2 ($\sigma = 0.3$), mean relative error.

3.7 Analysis of Variance for Model 2 with $\sigma = 0.30$

3.8 Model 3, mean relative error.

3.9 Analysis of Variance for Model 3

4.1 Results for a Linear Model, $R^2$ and Eigenvalues

4.2 Results for a Linear Model, Estimates

4.3 Results for a Linear Model, Grid Spacings
Chapter 1

Introduction

1.1 Motivation

The nonlinear mixed model (NLMM) is a widely used tool in modeling repeated measurements on different individuals. NLMMs are used when the mean response of an individual is believed to be a nonlinear function of individual–specific random effects. In pharmacokinetics, for example, NLMMs are used to describe the variation of drug disposition of patients. Here, absorption or elimination rates of a drug specific to an individual may be modeled as functions of fixed and random effects, where the random effects allow for inter–subject differences in kinetics. In most cases, the marginal distribution of the response given the random effects cannot be specified in a closed form. The inability to specify in a closed form has restricted inference procedures from exact likelihood methods due to intense computational needs to
numercially integrate out the random effects and maximize the resulting likelihood function.

The work presented here investigates the performance of methods that approximate the log-likelihood of the NLMM. In Chapter 2, two popular methods for finding estimates for the NLMM are examined. Each method requires normal distributional assumptions on the random effects and intra-individual error. A computer simulation experiment was performed to discover how robust the methods are to these and other assumptions. (Chapter 2 was written as a separate work and is incorporated here in its original form as accepted for publication in Computational Statistics and Data Analysis. Additional tables were added in Appendix A.3 that were not previously included. These tables are not referenced in Chapter 2.)

A second computer simulation, reviewed in Chapter 3, compares several methods used in common practice to approximate the log-likelihood in terms of accuracy and computing time. With the advance of today’s computing, another, more computationally intense method can be considered. This method is based on maximum likelihood and does not attempt to find a closed form solution of the likelihood. The likelihood can be numerically integrated and maximized using stochastic approximation, resulting in estimation of parameters based on the true likelihood and not on approximations. This method is described in Chapter 4.

In Chapter 1, an introduction to NLMMs and the approximate methods is given. A Hierarchical NLMM is described in Section 1.2 to motivate existing approximate
methods. These approximate methods, which result in closed form approximate solutions to the likelihood, are described in Section 1.3. A preview of the results of Chapters 2–4 is given in Section 1.4 and a final summary of conclusions and a description of future work is made in Chapter 5.

1.2 Model

Consider the following Hierarchical NLMM,

\[ y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \]  \hspace{1cm} (1.1)

\[(e_i | \beta_i) \sim [0, R_i (\beta_i, \xi)]\]

\[ \beta_i = d(a_i, \beta, b_i) \]

\[ b_i \sim (0, D) \]

where \( y_{ij} \) is the \( j \)th response for the \( i \)th individual, \( i = 1, ..., M; j = 1, ..., n_i \); \( x_{ij} \) is the covariate for the \( i \)th individual for the \( j \)th response; \( \beta_i \) is a \( p \times 1 \) random effects vector for the \( i \)th individual; \( f(x_{ij}, \beta_i) \) is a nonlinear scalar mean function which will also be denoted as \( f_j (\beta_i) \), and as a vector mean function \( f(\beta_i) = [f_1 (\beta_i), ..., f_{n_i} (\beta_i)]^T \); the error for the \( j \)th response of the \( i \)th individual is \( e_{ij} \); and \( R_i \) is an \( n_i \times n_i \) matrix function of random effects \( \beta_i \) and fixed effects \( \xi (q \times 1) \). Dependence of \( \beta_i \) on covariates \( a_i \) at the individual level is modeled by \( d(a_i, \beta, b_i) \), a nonlinear \( p \times 1 \) function of fixed
effects $\beta \ (r \times 1)$ and latent random effects $b_i \ (k \times 1)$. The covariance of $b_i$ is the $(k \times k)$ matrix $D$.

Let the conditional density of $y_i$ given $b_i$ be $p(y_i|b_i)$ and the density of $b_i$ be $\pi(b_i)$.

Combining the unknown fixed parameters into one vector, let $\phi = [\beta^T, \xi^T, \text{vech} \ (D)^T]^T$. Then the maximum likelihood estimates for $\phi$ can be found by maximizing in $\phi$

$$L(\phi) = \prod_{i=1}^{M} \int p(y_i|b_i) \pi(b_i) \ db_i$$ (1.2)

where $M$ is the number of individuals.

### 1.3 Approximate Methods

Because the likelihood (1.2) cannot be maximized with a closed form solution for mean functions that are nonlinear in $b_i$, current inferential methods have approximated the likelihood by either approximating the integrand before integrating or approximating the integral with a finite sum (i.e. a numerical integration). Both strategies result in closed form expressions for $L(\phi)$ with computable values. The methods discussed in this section include First Order Linearization, Laplace’s Method, Importance Sampling, and Gaussian Quadrature. These methods all require the assumption that the intra–individual errors and random effects are normally distributed, i.e.

$$(e_i|\beta_i) \sim N \ [0, R_i \ (\beta_i, \xi)] \text{ and } b_i \sim N \ (0, D).$$
1.3.1 First Order Linearization

Sheiner and Beal (1980) approximated the integral (1.2) by approximating the integrand with a Taylor series expansion for \( f \left[ d \left( a_i, \beta, b_i \right) \right] \) about \( b_i = 0 \) before integration. Note that here \( \beta_i \) is denoted by \( d(a_i, \beta, b_i) \) to show explicitly how the expansion will be taken in \( b_i \). From (1.1), let \( e_i = R_i^{1/2} \left[ d \left( a_i, \beta, b_i \right), \xi \right] \epsilon_i \) where \( R_i^{1/2} \left[ d \left( a_i, \beta, b_i \right), \xi \right] \) is the Cholesky decomposition of \( R_i \left[ d \left( a_i, \beta, b_i \right), \xi \right] \), \( E(\epsilon_i) = 0 \) and \( \text{Var}(\epsilon_i) = I_{n_i} \) where \( e_i \) and \( b_i \) are independent. Then

\[
y_i = f_i \left[ d \left( a_i, \beta, b_i \right) \right] + R_i^{1/2} \left[ d \left( a_i, \beta, b_i \right), \xi \right] \epsilon_i
\]

by Taylor series expansion, where \( F_i(\beta, b_i^*) = \partial/\partial b_i^* f_i(\beta_i)\big|_{\beta_i=d(a_i,\beta,b_i^*)} \) and

\[
\Delta_{b_i}(\beta, b_i^*) = \partial/\partial b_i d(a_i, \beta, b_i)\big|_{b_i=b_i^*}. \quad \text{All terms quadratic in } e_i \text{ or } b_i \text{ and cross-product terms are ignored.}
\]

Next, let \( Z_i(\beta, b_i) = F_i(\beta, b_i) \Delta_{b_i}(\beta, b_i) \) and \( e_i^* = R_i^{1/2} \left[ d \left( a_i, \beta, 0 \right), \xi \right] \epsilon_i \). Then

\[
y_i \approx f \left[ d \left( a_i, \beta, 0 \right) \right] + Z_i(\beta, 0) b_i + e_i^*
\]

If the approximation in (1.3) is taken to be exact, then

\[
E(y_i) = f \left[ d \left( a_i, \beta, 0 \right) \right] \\
\text{Var}(y_i) = R_i \left[ d \left( a_i, 0, \xi \right) \right] + Z_i(\beta, 0) D Z_i^T(\beta, 0) \quad (1.4)
\]

\[
= V_i(\beta, 0, \omega)
\]

where \( \omega = [\xi^T, \text{vech}(D)^T]^T \).
Beal and Sheiner proceeded by constructing the likelihood from this approximate model where \( b_i \) and \( e_i^* \) are normally distributed. They performed the joint maximum likelihood estimation of \( \beta \) and \( \omega \) by maximizing the approximate likelihood

\[
L_{FO}(\phi) = \prod_{i=1}^{M} \int \frac{1}{(2\pi)^{n_i/2}} |R_i^{-1}[d(a_i, \beta, 0), \xi]|^{1/2} \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, 0)] - Z_i(\beta, 0)b_i \}^T R_i^{-1}[d(a_i, \beta, 0), \xi] \{y_i - f[d(a_i, \beta, 0)] - Z_i(\beta, 0)b_i \} \right) \frac{1}{(2\pi)^{k/2}} |D^{-1}|^{1/2} \exp \left( -\frac{1}{2} b_i^T D^{-1}b_i \right) \, db_i
\]

\[
= \prod_{i=1}^{M} \frac{1}{(2\pi)^{n_i/2}} |R_i^{-1}[d(a_i, \beta, 0), \xi]|^{1/2} |D^{-1}|^{1/2} \exp \left[ -\frac{1}{2} \{y_i - f[d(a_i, \beta, 0)] - Z_i(\beta, 0)b_i \}^T R_i^{-1}[d(a_i, \beta, 0), \xi] \{y_i - f[d(a_i, \beta, 0)] - Z_i(\beta, 0)b_i \} + b_i^T D^{-1}b_i \right] \, db_i
\]

which has the following closed form expression after completing the square,

\[
= \prod_{i=1}^{M} \frac{1}{(2\pi)^{n_i/2}} |V_i^{-1}(\beta, 0, \omega)|^{1/2} \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, 0)] \}^T V_i^{-1}(\beta, 0, \omega) \{y_i - f[d(a_i, \beta, 0)] \} \right).
\]

This likelihood results in the objective function

\[
l_{FO} = \sum_{i=1}^{M} \left( \ln |V_i(\beta, 0, \omega)| + \{y_i - f_i[d(a_i, \beta, 0)] \}^T V_i^{-1}(\beta, 0, \omega) \{y_i - f_i[d(a_i, \beta, 0)] \} \right)
\]

\[(1.5)\]

which is minus two times the marginal normal log-likelihood under (1.3).
Different numerical techniques can be implemented to minimize (1.5) including the Newton algorithm. The software packages NONMEM (Beal and Sheiner, 1998) and the SAS NLMIXED Procedure (Wolfinger, 1999) include minimization procedures based on the First Order approximation to the likelihood function. NONMEM maximizes the resulting normal log-likelihood (1.5) through the use of two quadratic estimating equations for $\beta$ and $\omega$. The NLMIXED Procedure in SAS provides several choices for algorithms which minimize (1.5) in $(\beta, \omega)$, including Newton-Raphson and quasi-Newton methods but requires $R_i (\beta_i, \xi) = \sigma^2 I_{n_i}$.

Another approach is Generalized Least Squares (GLS) assuming that the moments in (1.4) are exact. Davidian and Giltinan (1995) describe the GLS method in detail. A software package that uses this GLS approach as one of its several choices, is the SAS macro NLINMX also written by Wolfinger (Littell et al., 1996, Wolfinger and Lin, 1997). This macro allows more complex models for $R_i$ by utilizing a weighted least squares approach.

### 1.3.2 Laplace’s Approximation

Another method for approximating (1.2) uses Laplace’s Approximation,

$$
\int e^{n_i \ell_i(b_i)} \, db_i \approx \left( \frac{2\pi}{n_i} \right)^{k/2} | - \ell_i''(\hat{b}_i) |^{-1/2} e^{n_i \ell_i(\hat{b}_i)},
$$

where $b_i$ is $(k \times 1)$, $\hat{b}_i$ maximizes $\ell_i(b_i)$, and $\ell_i''(\hat{b}_i) = \partial^2 / \partial b_i \partial b_i^T \ell_i(b_i) |_{b_i = \hat{b}_i}$ is a $(k \times k)$ matrix.
In order to use Laplace’s approximation for our likelihood, we first let \( \ell_i(b_i) = \frac{1}{n_i} \ln[p(y_i|b_i)p(b_i)] \). Because normal assumptions have been made,

\[
\ell_i(b_i) = \frac{1}{n_i} \ln \left\{ \frac{1}{(2\pi)^{n_i/2}} |D^{-1}|^{1/2} |R_i^{-1}| d(a_i, \beta, b_i), \xi|^{1/2} \exp \left[ -\frac{1}{2} \left\{ y_i - f[d(a_i, \beta, b_i)] \right\}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \} + b_i^T D^{-1} b_i \right] \right\}
\]

\[
= \frac{1}{n_i} \ln \left\{ \frac{1}{(2\pi)^{n_i/2}} |D^{-1}|^{1/2} |R_i^{-1}| d(a_i, \beta, b_i), \xi|^{1/2} \right\} - \frac{1}{2n_i} \left\{ y_i - f[d(a_i, \beta, b_i)] \right\}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \} + b_i^T D^{-1} b_i
\]

If \( R_i[d(a_i, \beta, b_i), \xi] \) is of a simpler form that does not depend on \( b_i \), then

\[
\ell_i'(b_i) = \frac{\partial}{\partial b_i} \left[ -\frac{1}{2n_i} \left\{ y_i - f[d(a_i, \beta, b_i)] \right\}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \} + b_i^T D^{-1} b_i \right]
\]

\[
= -\frac{1}{n_i} \left( Z_i^T (\beta, b_i) R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \} + D^{-1} b_i \right)
\]

and

\[
\ell_i''(b_i) = \frac{\partial}{\partial b_i} \left[ \frac{1}{2n_i} \left\{ y_i - f[d(a_i, \beta, b_i)] \right\}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \} + b_i^T D^{-1} b_i \right]
\]

\[
= -\frac{1}{n_i} \left( Z_i^T (\beta, b_i) R_i^{-1} Z_i(\beta, b_i) + D^{-1} \right) + \frac{\partial^2 f[d(a_i, \beta, b_i)]}{\partial b_i \partial b_i} R_i^{-1} \{ y_i - f[d(a_i, \beta, b_i)] \}
\]

\[
\approx -\frac{1}{n_i} \left( Z_i^T (\beta, b_i) R_i^{-1} Z_i(\beta, b_i) + D^{-1} \right)
\]
Here the Hessian term is dropped since the effect of 

\[ \frac{\partial^2 f[d(a_i, \beta, \hat{b}_i)]}{\partial b_i \partial b_i^*} \bigg|_{b_i = \hat{b}_i} R_i^{-1} \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \} \] 

is much less than 

\[ \frac{\partial f[d(a_i, \beta, \hat{b}_i)]}{\partial b_i} R_i^{-1} \frac{\partial f[d(a_i, \beta, \hat{b}_i)]}{\partial b_i^*} \bigg|_{b_i = \hat{b}_i} \] 

on the likelihood (Bates and Watts, 1980).

Let \( \hat{Z}_i = Z_i(\beta, \hat{b}_i) \). The resulting approximation for minus two times the log-likelihood using Laplace’s approximation is

\[
-2 \ln L_i \approx -2 \left[ \frac{k}{2} \ln(2\pi) - \frac{k}{2} \ln(n_i) - \frac{1}{2} \ln \left| -\ell''(\hat{b}_i) \right| + n_i \ell_i(\hat{b}_i) \right]
\]

\[
\approx -k \ln(2\pi) + k \ln(n_i) + \ln \left| \frac{1}{R_i} (\hat{Z}_i^T R_i^{-1} \hat{Z}_i + D^{-1}) \right|
\]

\[
-2 \ln \left\{ \frac{1}{(2\pi)^{n_i+k/2}} |D^{-1}|^{1/2} |R_i|^{1/2} \right\}
\]

\[
+ \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \} + \hat{b}_i^T D^{-1} \hat{b}_i
\]

\[
\approx -k \ln(2\pi) + k \ln(n_i) - k \ln(n_i) + \ln |\hat{Z}_i^T R_i^{-1} \hat{Z}_i + D^{-1}|
\]

\[
+ n_i \ln(2\pi) + k \ln(2\pi) - \ln |D^{-1}| - \ln |R_i|^{-1}
\]

\[
+ \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \} + \hat{b}_i^T D^{-1} \hat{b}_i
\]

\[
\approx n_i \ln(2\pi) + \ln |\hat{Z}_i^T R_i^{-1} \hat{Z}_i + D^{-1}| + \ln |D| + \ln |R_i|
\]

\[
+ \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \}^T R_i^{-1} \{ y_i - f[d(a_i, \beta, \hat{b}_i)] \} + \hat{b}_i^T D^{-1} \hat{b}_i
\]

When \( R_i \) does depend on \( b_i \) but under conditions relevant to applications like pharmacokinetics where \( \sigma \) is “small” (see Ko and Davidian, 1999), the approximation is still appropriate.

From the approximate log-likelihood, the approximate marginal moments of \( y_i \) are given by

\[
E(y_i) \approx f_i[d(a_i, \beta, \hat{b}_i)] - \hat{Z}_i
\]
\[
\text{Var}(y_i) \approx \tilde{Z}_i D \tilde{Z}_i^T + R_i [d(a_i, \beta, \hat{b}_i), \xi]
\]

Note that these are the exact marginal moments of \( y_i \) if \( f_i \) is linear in \( b_i \).

These and other similar results using Laplace's approximation are the basis for several algorithms used to fit NLMMs in current software, e.g. the First Order Conditional Expectation method (FOCE) in NONMEM (Beal and Sheiner, 1998), the algorithm of Lindstrom and Bates (1990) implemented in the S–Plus function nlinme (Pinheiro and Bates, 1995), and the procedure obtained by using the EBLUP option in the SAS macro NLINMIX (Littell et al., 1996, Wolfinger and Lin, 1997).

Some of these methods were developed for somewhat simpler models. For example, in nlinme, the secondary structure modeling \( \beta_i \) takes the specific, simplified form

\[
d(a_i, \beta, b_i) = A_i \beta + B_i b_i
\]

where \( A_i \) \((p \times r)\) and \( B_i \) \((p \times k)\) are design matrices for the fixed and random effects, respectively. Note also that this approximation was derived for the model where \( R_i \) does not depend on \( b_i \). An approach based on Laplace's approximation that does not make this assumption is derived in Section 3.2.1.

The log of the above approximation is maximized in \( \beta, D \), and \( \sigma^2 \) to find approximate MLE's. In this manner, Pinheiro and Bates have approximated the likelihood. This method is exact when the mean function \( f[x_{ij}, d(a_i, \beta, b_i)] \) is linear in \( b_i \).
1.3.3 Importance Sampling

Importance sampling is a numerical integration technique that takes advantage of the fact that any integral can be thought of as an expectation, e.g., \( \int xg(x)dx = E(X) \), where \( g(X) \) is the probability distribution function (p.d.f.) of \( X \). Suppose \( \int h(x)dx \) is to be numerically integrated where \( h(x) \) can be written as \( h(x) = f(x)g(x) \) and \( g(x) \) is a p.d.f. Then, if a random sample \( x_1, x_2, ..., x_N \) can be taken from the distribution with p.d.f. \( g(x) \), a numerical approximation for \( \int h(x)dx \) is the sample mean of the \( f(x_i) \). In this context, the p.d.f \( g(x) \) is called the importance distribution. To actually compute a value for the integral, there must be no unknown parameters in the importance distribution or the function \( f(x) \).

Pinheiro and Bates (1995) describe how they used importance sampling to approximate the likelihood (1.2). Software has been provided in statlib to be called into S–Plus to perform their version of importance sampling. The model assumed is

\[
y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \tag{1.8}
\]

\[
(e_i | \beta_i) \sim N(0, \sigma^2 I)
\]

\[
\beta_i = A_i \beta + B_i b_i
\]

\[
b_i \sim N\left(0, \sigma^2 D\right)
\]

Note that the matrix \( D \) has been rescaled with \( \sigma^2 \). This was done so that the \( \sigma^2 \) drops out of the sums of squares later in the minimization. The \( \sigma^2 D \) in this model is
the same as $D$ as developed earlier in Equation 1.1. A random sample $z_1^*, z_2^*, ..., z_N^*$ was taken from $N(0, I_k)$ and transformed by

$$b_{ij}^* = \hat{b}_i + \sigma [G(\beta, D, y_i)]^{-1/2} z_j^*$$

so that

$$b_{ij}^* \sim N \left[ \hat{b}_i, \sigma^2 G^{-1}(\beta, D, y_i) \right] \quad (1.9)$$

where

$$\hat{b}_i = \arg \min_{b_i} [y_i - f(x_{ij}, A_i \beta + B_i b_i)]^T [y_i - f(x_{ij}, A_i \beta + B_i b_i)] + b_i^T D b_i$$

and $G(\beta, D, y_i) = \hat{Z}_i^T \hat{Z}_i + D^{-1}$.

Next, a closed form expression which approximates the likelihood was derived by using (1.9) as an importance distribution.

$$\prod_{i=1}^{M} \int p(y_i | b_i, \beta, D, \sigma^2) \pi(b_i) db_i =$$

$$= \prod_{i=1}^{M} \int \frac{1}{(2\pi \sigma^2)^{(m+k)/2}} \left| D^{-1} \right|^{1/2} \exp \left( -\frac{1}{2\sigma^2} \left[ \{y_i - f(A_i \beta + B_i b_i)\}^T \{y_i - f(A_i \beta + B_i b_i)\} + b_i^T D^{-1} b_i \right] \right) db_i$$

$$= \prod_{i=1}^{M} \int \frac{1}{(2\pi \sigma^2)^{n/2}} \left| D^{-1} \right|^{1/2} \exp \left( -\frac{1}{2\sigma^2} \left[ \{y_i - f(A_i \beta + B_i b_i)\}^T \{y_i - f(A_i \beta + B_i b_i)\} + b_i^T D^{-1} b_i \right] + \frac{1}{2\sigma^2} (b_i - \hat{b}_i)^T G(\beta, D, y_i)(b_i - \hat{b}_i) \right) db_i$$

$$= \frac{1}{(2\pi \sigma^2)^{k/2}} \left| G(\beta, D, y_i) \right|^{1/2} \exp \left[ -\frac{1}{2\sigma^2} (b_i - \hat{b}_i)^T G(\beta, D, y_i)(b_i - \hat{b}_i) \right] db_i$$
Similarly, Wolflinger has included importance sampling as an optional method in the SAS procedure NLMIXED (Wolflinger 1999).

### 1.3.4 Gaussian Quadrature

Another numerical integration method that Pinheiro and Bates have made available through statlib is Gaussian quadrature (Pinheiro and Bates, 1995, where the model assumed is 1.8). This method approximates integrals of functions with weighted averages of the integrand evaluated at abscissas which are predetermined by the quadrature rule (see Abramowitz and Stegun, 1967).

Similarly to importance sampling, suppose \( f h(x)dx \) is to be numerically integrated where \( h(x) \) can be written as \( h(x) = f(x)w(x) \) and \( w(x) \) is the importance distribution. The integral is approximated as

\[
\int h(x)dx = \int f(x)w(x)dx \approx \sum_{i=1}^{n} w_i f(x_i).
\]
Pinheiro and Bates derived an approximation to the likelihood using the method of Gauss–Hermite quadrature where they let $z_j, w_j \ j = 1, ..., N$ be the abscissas and weights, respectively, where $w(x)$ is the density function for the standard normal distribution (see Davis and Rabinowitz, 1984). Then

$$
\int p(y_i | b) \pi(b) \, db_i = \frac{1}{(2\pi \sigma^2)^{n_i/2}} \left| D \right|^{-1/2} \exp \left\{ \frac{-1}{2\sigma^2} \left[ y_i - f(A_i \beta + B_i b_i) \right]^T [y_i - f(A_i \beta + B_i b_i)] \right\} \, db_i
$$

$$
= \frac{1}{(2\pi \sigma^2)^{n_i/2}} \int \frac{1}{(2\pi)^{k/2}} \exp \left\{ \frac{-1}{2\sigma^2} \left[ y_i - f(A_i \beta + B_i \sigma D^{1/2} z^* \right) \right\] \left[ y_i - f(A_i \beta + B_i \sigma D^{1/2} z^* \right] \, dz^* 
$$

$$
\approx \frac{1}{(2\pi \sigma^2)^{n_i/2}} \sum_{j_1=1}^N \sum_{j_2=1}^N \cdots \sum_{j_k=1}^N \left( \exp \left\{ \frac{-1}{2\sigma^2} \left[ y_i - f(A_i \beta + B_i \sigma D^{1/2} z^*_j \right] \right\] \left[ y_i - f(A_i \beta + B_i \sigma D^{1/2} z^*_j \right] \right) \prod_{\ell=1}^k w_{j_\ell}
$$

where $z^*_j = (z^*_1, z^*_2, ..., z^*_k)^T$. Using this result to approximate $L_i$, they minimized $-2 \sum_{i=1}^M \ln L_i(\phi)$ in $\phi$ to get approximate maximum likelihood estimates.

Another, similar approach is also available in nlme which uses an alternative centering, $b_i = \hat{b}_i + \sigma [G(\beta, D, y_i)]^{1/2} z^*$. This approach results in the same kernel distribution as the importance distribution in Section 1.3.3 and has a more efficient numerical integration than with the first choice of kernel. Pinheiro and Bates called this alternate method “adaptive Gaussian quadrature” and it had results very similar to importance sampling (see Pinheiro and Bates, 1995). Adaptive Gaussian quadrature is included in SAS Proc NL MIXED as the default method (Wolfgang, 1999).
1.4 Preview of Results

Chapters 2–4 report the findings of three simulation studies. The first experiment tested which method was more robust to misspecifications of assumptions in the nonlinear mixed effects model: First Order linearization or Laplace’s approximation. The convergence rates of the methods were an issue; when both methods converged, Laplace’s approximation performed better, resulting in more precise parameter estimates. But the convergence rate of Laplace’s approximation was affected more by model misspecification. This resulted in a new question. What is different about the likelihood surfaces that causes these differences in the convergence rates?

As a result, a second study was designed to begin to examine how different methods vary in their approximations of the likelihood. A simulation experiment compared several methods to see which approximated the likelihood function best in a region “close” to its maximum. The methods included were Laplace’s approximation, importance sampling, and Gaussian quadrature. A small number of models were used to simulate data. Models with both constant and non-constant error variance were included. Existing methods were enhanced to deal properly with the non-constant error variance. New versions of importance sampling and Gaussian quadrature were introduced which are “scaled” and “centered” differently, leading to more accurate estimation. Laplace’s approximation again was a surprisingly good choice if a computationally low-intensive method is desired. The overall best performer was Gaussian quadrature. This study is described in Chapter 3. A future study is needed that
includes examination of methods on the non–converging data sets.

A new method was designed (see Chapter 4) to address two issues from the previous studies: how can maximization of the true likelihood be performed instead of using approximations and how can the optimization be performed with random input? By using numerical integration techniques, the likelihood can be approximated to any specified level of accuracy. In particular, a random numerical integration method, e.g. importance sampling, can report standard errors for the estimation. But this random method leads to random input for an optimization algorithm that results in different values of the likelihood when re–evaluated at the same point in the parameter space. Most optimization procedures were not designed to handle this kind of input. The new method described in Chapter 4 addresses these optimization issues. A simulation study with data from a linear mixed model was promising, resulting in accurate estimation. At this time the method has not shown success for a NLMM. Issues dealing with initial estimation used for starting values have blocked success and are addressed in Chapter 4.

More specific conclusions for all three studies are given in Chapters 2–4 with a final summary in Chapter 5. Future work to be conducted is also described in Chapter 5.
Chapter 2

Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models
Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models

Alan Hartford*, Marie Davidian

Department of Statistics, Box 8203, North Carolina State University, Raleigh, North Carolina 27695-8203

(919) 515-2489, hartford@stat.ncsu.edu, (919) 515-1940, davidian@stat.ncsu.edu

Abstract

The nonlinear mixed effects model provides a framework for inference in a number of applications, most notably pharmacokinetics and pharmacodynamics, but also in HIV and other disease dynamics and in a host of other longitudinal-data settings. In these models, to characterize population variation, individual-specific parameters are modeled as functions of fixed effects and mean-zero random effects. A standard assumption is that of normality of the random effects, but this assumption may not always be realistic, and, because the random effects are not observed, it may be difficult to verify. An additional issue is specifying the form of the function relating individual-specific parameters to fixed and random effects. Again, because this relationship is not observed explicitly, it may be difficult to specify. Popular methods for fitting these models are predicated on the normality assumption, and past studies evaluating their performance have assumed that normality and the form of the
model are correct specifications. We investigate the consequences for population inferences using these methods when the normality assumption is inappropriate and/or the model is misspecified.

**Key words:** Random Effects, Nonnormality, Laplace approximation, Linearization

* Corresponding author

### 2.1 Introduction

Nonlinear mixed effects models have become a routine tool in biomedical applications to represent repeated measurement data on each of several individuals. A key area where nonlinear mixed effects models have seen widespread use is population pharmacokinetics, where data on drug concentration and time from a number of subjects are to be used to characterize drug disposition and its variation in the population of patients (Beal and Sheiner, 1982, 1985). Other applications include pharmacodynamics (Rosner and Müller, 1994), joint pharmacokinetic-pharmacodynamic modeling (Holford and Sheiner, 1981), modeling of markers of disease progression (Morrell et al., 1995) and modeling of within-host HIV dynamics following potent antiviral treatment (Wu and Ding, 1999). The US Food and Drug Administration has encouraged use of these models as the basis for population pharmacokinetic analyses to be submitted for regulatory review (US FDA, 1997).

In all of these applications, mean response for a particular individual is thought to
be a nonlinear function of physically meaningful parameters; e.g. in pharmacokinetics, a function derived by compartmental modeling, depending in a nonlinear way on individual-specific absorption, elimination, and distribution parameters, is assumed to characterize the concentration-time profile for a given subject. These parameters in turn are assumed to depend, through some functional relationship, on fixed effects, covariates, and individual-specific random effects whose distribution provides a model for random variation of the parameters in the population of individuals; a standard assumption is that this distribution is multivariate normal. Thus, formally, the conditional (on the random effects) mean of the response for a given individual depends on the random effects in a nonlinear fashion. This nonlinear dependence requires multidimensional integration over the distribution of the random effects to obtain the marginal distribution or moments of the data, an operation that almost never can be carried out in closed form.

In order to avoid the computational burden and instability associated with complex numerical integration, the most common approaches to fitting nonlinear mixed models involve approximation of the integral involved in the marginal likelihood, assuming the conditional distribution of the response within individuals is normal. Several approaches to such approximation and algorithms to implement them have been proposed; an admittedly inexhaustive list includes Beal and Sheiner (1982), Goldstein (1991), Breslow and Clayton (1993), Hirst et al. (1991), Lindstrom and Bates (1990), Solomon and Cox (1992), Vonesh and Carter (1992), Wolfinger (1993),
and Vonesh (1996). A number of popular software implementations are available, including NONMEM (Beal and Sheiner, 1998), the nlme suite of S–Plus functions (Pinheiro and Bates, 1995), and the SAS NLMIX macro (Littell et al., 1996). A feature of all of these methods is that they are derived assuming that the random effects are normally distributed.

Other approaches involve evaluation of the integrals via numerical integration or using Markov chain Monte Carlo techniques, including Davidian and Gallant (1993), Pinheiro and Bates (1995), and Wakefield (1996), software is available (e.g. Davidian and Gallant, 1994) or emerging (SAS version 7.0 introduces the new NLMIXED procedure); however, considerable experience, certainly with population pharmacokinetics, suggests that approximate methods provide reliable inferences in many instances with less computational challenge. Numerous simulation studies (e.g. Sheiner and Beal, 1980; Pinheiro and Bates, 1995; Wolfinger and Lin, 1997) have exhibited reliable performance in many contexts.

Although considerable such evidence is available, almost all investigations of performance of approximate techniques have been based on the assumptions that (i) the random effects are normally distributed, an assumption upon which the methods are predicated, and (ii) the model used to relate individual-specific parameters to fixed and random effects and covariates is correctly specified. It is widely acknowledged that the assumption of normality of the random effects may not always be appropriate, see Mallet (1986), Davidian and Gallant (1992, 1993), Fattinger, Sheiner, and
Verotta (1995). Complicating matters is the fact that the random effects are unobservable latent model components; thus, no straightforward diagnostic to evaluate the validity of the assumption of normality is available, and analysts typically appeal to the methods implemented in the software without formal evidence that the underlying assumption of normality is appropriate. It is thus natural to be concerned whether these methods yield reliable inferences when the normality assumption is not correct. Moreover, although graphical displays may be used to guide the analyst in modeling of individual-specific parameters as a function of covariates (Davidian and Gallant, 1992; Wakefield, 1996) evidence may not support one model at the exclusion of others. Thus, there is potential for the relationship to be incorrectly specified.

The consequences of misspecifying the random effects distribution have been discussed for linear mixed effects models (Butler and Louis, 1992; Verbeke and Lesaffre, 1996, 1997; Muthen and Shedden, 1999; Tao et al., 1999), and simulations have shown that estimation of fixed parameters may not be severely compromised (e.g. Verbeke and LeSaffre, 1997). However, because of the complicated way in which the random effects enter a nonlinear mixed model, the different ways in which this dependence may be modeled, and the use of approximate inferential methods, it is not clear that conclusions for the linear case carry over to this setting.

We describe results of a simulation study performed to elucidate the robustness of two popular approximate methods to nonnormality and model misspecification under different situations. First, we review the nonlinear mixed effects framework. The two
approximate inferential methods are discussed in the next section. We then describe the simulation study and summarize the results.

2.2 Nonlinear mixed effects model

The general form of the nonlinear mixed effects model is given by

\[ y_i = f_i(\beta_i) + e_i, \quad (e_i \mid b_i) \sim \{0, \sigma^2 \Lambda_i(\beta_i, \gamma)\} \]  \hspace{1cm} (2.1)

\[ \beta_i = d(a_i, \beta, b_i), \quad b_i \sim (0, D). \]  \hspace{1cm} (2.2)

Here, \( y_i \) is the \((n_i \times 1)\) vector of observations \( y_{ij} \) on the \( i \)th subject, \( i = 1, \ldots, m \), at times \( t_{ij}, j = 1, \ldots, n_i \); \( f_i(\beta_i) \) is an \((n_i \times 1)\) vector of nonlinear functions with \( j \)th element \( f(x_{ij}, \beta_i) \), where \( f \) is the nonlinear function characterizing within-subject (conditional) mean response as a function of the individual-specific regression parameter \( \beta_i \) \((p \times 1)\), and \( x_{ij} \) contains \( t_{ij} \) and other covariates; \( e_i \) is a vector of random intra-individual errors for the \( i \)th subject; and \( y_i \) has covariance matrix \( \sigma^2 \Lambda_i(\gamma, \beta_i) \) with parameters \( \sigma^2 \) and \( \gamma \) \((q \times 1)\) common to all subjects. Dependence of \( \beta_i \) on an individual-level vector of covariates \( a_i \) is represented by the \( p \)-dimensional function \( d \) of \( a_i \), a \((r \times 1)\) vector of fixed effects \( \beta \), and a \( k \)-dimensional vector of random effects \( b_i \), where \( b_i \) are independent and identically distributed with covariance matrix \( D \). Thus the inter-individual model, Equation (2.2), characterizes variation among individual-specific parameters \( \beta_i \) as a parametric function of covariates and random phenomena.
In applications such as pharmacokinetics, the analyst may be confident about the form of the individual-specific response function \( f \); in a population study, a plausible choice for \( f \) may have been established by previous studies where subjects were intensively sampled following single or multiple doses. For example, for a drug given as an intravenous bolus, the one compartment model

\[
f(x_{ij}, \beta_i) = \frac{D_i}{V_i} \exp \left( -\frac{C_l_i t_{ij}}{V_i} \right)
\]

(2.3)

may provide a suitable characterization for within-subject plasma concentration at time \( t_{ij} \) following dose \( D_i \) for subject \( i \), where \( C_l_i \) and \( V_i \) represent subject-specific clearance and volume of distribution, respectively, and \( \beta_i = (C_l_i, V_i)' \). Moreover, the nature of within-subject variation, represented by \( e_i \) in Equation (4.6), may be well-understood from prior experience with intensive data. One standard model that is thought to provide an accurate characterization of intra-individual variation is to assume \( e_i \) conditional on \( b_i \) are normally distributed, where the covariance matrix \( \sigma^2 \Lambda_i(\beta_i, \gamma) \) is diagonal with \( j \)th diagonal element given by

\[
\text{var}(e_{ij} | b_i) = \sigma^2 f^2(x_{ij}, \beta_i).
\]

(2.4)

Note that this specification assumes that observations are taken sufficiently far apart in time that intra-subject autocorrelation is negligible. A related model is to assume that \( \log y_{ij} \) are conditionally independent and lognormally distributed with mean \( \log f(x_{ij}, \beta_i) \) and \( \Lambda_i(\beta_i, \gamma) \) an \( (n_i \times n_i) \) identity matrix. These two models are very similar when intra-subject variation is “small” relative to the range of the mean, as
is typical in pharmacokinetics (e.g. Davidian and Giltinan, 1995, Chap. 9).

The analyst often has a more difficult challenge in specifying the inter-individual model in Equation (2.2). There are two main aspects, specifying the form of the model and the assumption on the distribution of the random effects \( b_i \), and these are intertwined. For example, it is widely recognized that individual pharmacokinetic parameters may exhibit skewed distributions in the population of subjects; thus, it is popular to postulate loglinear models for \( Cl_i \) and \( V_i \) of the form

\[ Cl_i = \exp (\beta_1 + \beta_2 a_i + b_{1i}) \], \[ V_i = \exp (\beta_3 + b_{2i}) \]  

so that if \( b_i = (b_{1i}, b_{2i})' \) is normally distributed, \((Cl_i, V_i)\) are jointly lognormal. Equation (2.5) also allows for the possibility that variation among \( Cl_i \) is in part a systematic consequence of association with a covariate \( a_i \), e.g. a measure of renal function or body weight; thus, the effect of this covariate is taken to enter the model in a loglinear fashion. An alternative to the model in Equation (2.5) is one that allows random effects and covariates to enter linearly, e.g.

\[ Cl_i = \beta_1 + \beta_2 a_i + b_{1i}, \quad V_i = \beta_3 + b_{2i}. \]  

Depending on the situation, it might be difficult to distinguish between a linear and loglinear specification, as linear and exponential functions may be similar over a range of values of \( a_i \). Note, however, that the random effects \( b_i = (b_{1i}, b_{2i})' \) enter Equation (2.6) linearly; thus under the assumption of normality, even if the fixed effects parts of the linear and exponential functions in Equations (2.5) and (2.6) look similar as
a function of $a_i$, the distinction between the two models is more profound: Equation (2.5) takes $(Cl_i, V_i)$ to be jointly lognormal, while Equation (2.6) assumes they are jointly normal in the population.

Indeed, although the assumption of normality of random effects is routinely adopted, it simply may not be appropriate. A number of scenarios are possible. For example, it may be that $(Cl_i, V_i)$ are neither jointly normally nor lognormally distributed; rather, in a model like that in Equation (2.5), it may be that although log $Cl_i$ and log $V_i$ are roughly symmetrically distributed, individuals far from the “center” (in terms of log $Cl_i$ and/or log $V_i$) occur more frequently in the population than would be expected in a normally distributed population. In this case, the true distribution of $b_i$ might have heavier tails than a normal with the same variation. Alternatively, although a study may have been carried out targeting a specific population of subjects (e.g. through exclusion criteria), subjects who are in truth not members of this population may be inadvertently included. In this case, the apparent distribution of $b_i$ from such data may actually have heavier tails than the normal as well as greater dispersion than that corresponding only to the population of interest. It might instead be that the log transformation in Equation (2.5) may be an over- or under-transformation in terms of making the resulting logarithms of $Cl_i$ and $V_i$ look jointly normally; instead, the joint distribution may be skewed. Still another possibility is that the apparent distribution of $b_i$ may not be unimodal. For example, if an important covariate is unavailable, the model attributes systematic variation due to this covariate to random
causes, incorporating this variation in \( b_i \). The result may be an apparent bimodal distribution (see Davidian and Giltinan, 1995, Chap. 7) that is not well-approximated by the assumption of normality.

The objective of an analysis is usually to estimate the fixed effects \( \beta \) in order to characterize the typical population values of the individual-specific parameters and to assess the relationship between \( \beta_i \) and covariates. In the context of the inter-individual models in Equations (2.5) and (2.6), interest may focus in particular on whether \( \beta_2 = 0 \). Estimation of the covariance parameters \( D \), representing the magnitudes of inter-subject variation, is also of interest; how these parameters are interpreted depends on the structure of the model \( d \). For instance, in Equation (2.6) the diagonal elements of \( D \) would be interpreted as the variances of \( Cl_i \) and \( V_i \) in the population; in Equation (2.5), these would be interpreted as approximate coefficients of variation.

### 2.3 Approximate methods

Of the common approximations used for fitting nonlinear mixed effects models, we concentrate on two of the most popular: First Order expansion and Laplace's Approximation. Variations on these approximations form the basis for the fitting methods implemented in software such as \( \text{NONMEM} \) (Beal and Sheiner, 1998), the S-Plus function \( \text{nlme} \) (Pinheiro and Bates, 1995), and the SAS macro \( \text{NLMIX} \) (Littell et al., 1996). We give only a brief review of these methods; more complete derivations and discussion may be found in Wolfinger and Lin (1997).
The need for approximate methods may be appreciated by inspection of the form of the marginal distribution of $y_i$ implied by Equations (4.6) and (2.2). Letting the conditional density of $y_i$ given $b_i$ be $p(y_i|b_i)$ and the density of $b_i$ be $p(b_i)$, the marginal distribution of $y_i$ is given by

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i. \quad (2.7)$$

Note that, even if both $p(y_i|b_i)$ and $p(b_i)$ are $n_i$— and $k$—dimensional normal densities, respectively, $p(y_i)$ need not be normal; moreover, this integral will almost always be analytically intractable. Thus, if inference based on the likelihood of the observed data is desired, this is complicated by inability to express this likelihood in closed form. Note further that writing down the marginal moments of $y_i$ will be plagued by the same issue; thus, inference based on the solution of generalized estimating equations predicated on the first two marginal moments $E(y_i)$ and $\text{var}(y_i)$ implied by Equations (4.6) and (2.2) will encounter similar difficulties. Consequently, a standard approach is to use approximations to (2.7) or similar quantities as the basis for inference.

**First Order Expansion**

The First Order expansion suggested by Sheiner and Beal (1980) may be motivated in a number of ways; the basic approach is to obtain an approximation to the integral in Equation (2.7) by expansion about $b_i = 0$ before integration is carried out. One way to obtain this approximation is to expand Equation (4.6) and retain only nonnegligible terms linear in $b_i$. Expanding the conditional mean vector $f_i(\beta_k)$ in a linear Taylor
series about \( b_i = 0 \) yields

\[
  f_i(\beta_i) = f_i\{d(a_i, \beta, b_i)\} \approx f_i\{d(a_i, \beta, 0)\} + Z_i\{d(a_i, \beta, 0)\}b_i,  \tag{2.8}
\]

where \( Z_i\{d(a_i, \beta, b^*)\} = \partial / \partial b_i f_i\{d(a_i, \beta, b_i)\}\big|_{b_i=b^*}. \) Further, representing the error term in Equation (4.6) as \( e_i = \sigma \Lambda_i^{1/2}(\beta_i, \gamma) \epsilon_i, \) where \( \epsilon_i \sim (0, I_{n_i}) \) and is independent of \( b_i \) and \( \Lambda_i^{1/2} \) is a square root matrix of \( \Lambda_i, \) expansion about the \( \ell \)th element of \( b_i = 0 \) yields for each \( \ell = 1, \ldots, k \)

\[
e_i \approx \sigma \Lambda_i^{1/2}\{d(a_i, \beta, 0), \gamma\} \epsilon_i + \sigma A_i\{d(a_i, \beta, 0), \gamma\} b_i,  \tag{2.9}
\]

where \( A_i \) is the partial derivative of \( \Lambda_i^{1/2} \) with respect to \( b_i, \ell. \) Combining Equations (2.8) and (2.9) and noting that the product \( b_i \epsilon_i \) in (2.9) is negligible relative to \( b_i \) gives an approximation that is linear in \( b_i, \) namely

\[
y_i \approx f_i\{d(a_i, \beta, 0)\} + Z_i\{d(a_i, \beta, 0)\}b_i + \sigma \Lambda_i^{1/2}\{d(a_i, \beta, 0), \gamma\} \epsilon_i.
\]

This allows approximate marginal moments to be derived readily, i.e.

\[
\begin{align*}
  E(y_i) & \approx f_i\{d(a_i, \beta, 0)\} \\
  \text{var}(y_i) & \approx Z_i\{d(a_i, \beta, 0)\} DZ_i\{d(a_i, \beta, 0)\} + \sigma^2 \Lambda_i\{d(a_i, \beta, 0), \gamma\} = V_i\{d(a_i, \beta, 0)\}. \tag{2.10}
\end{align*}
\]

The standard approach is then to assume that these approximate marginal moments are exact assuming that \( b_i \sim N(0, D) \) and \( \epsilon_i \sim N(0, I_{n_i}) \) in Equation (4.6), so that \( y_i \) is approximately normal with these moments. Wolfinger and Lin (1997) derive this same approximate model by direct approximation of the marginal likelihood
The well-known First Order (FO) method implemented in the NONMEM (Beal and Sheiner, 1998) software maximizes the resulting normal log-likelihood in $\beta$, $\sigma$, and $D$. This results in the solution of a set of two estimating equations for $\beta$ and $(D, \sigma)$, both of which are quadratic functions of the data. Alternatively, the ZERO option in the SAS macro NLINMIX fits the same approximate model using a simpler estimating equation for $\beta$ that is linear in the data, ignoring the additional information on $\beta$ that may be available in $Z_i$ and $\Lambda_i$. This approach is in the spirit of solving standard generalized estimating equations for $\beta$, $D$, and $\sigma$ assuming the first two moments given in (2.10) as described by Prentice and Zhao (1991). This approach is appealing because, as noted in Davidian and Giltinan (1995, Chaps. 2, 6), it offers more robustness to nonnormality and misspecification of the covariance model, both of which are of some concern here as both the assumptions that $y_i$ is normal and the form of the moments in Equation (2.10) are only approximations. Even if the approximation is good, for applications such as pharmacokinetics, where intra-subject variation is small relative to that among subjects, the two strategies yield very similar results. See Wolfinger and Lin (1997) and Davidian and Giltinan (1995, Chap. 6) for further details.

Inference based on this so-called First Order approximation is often favored in the case of sparse data on each subject, as is typical in population pharmacokinetics. Choice of this method in practice is often the consequence of failure to achieve convergence with sparse data of the fitting algorithm for the more complicated Laplace’s
approximation, which we discuss next.

Laplace’s Approximation

A potential drawback of the First Order approximation is that it may be a poor representation of the true marginal distribution and its moments. An alternative strategy is to base the approximation on something “closer” to \( b_i \) than its mean, zero. A common way to motivate this approach is to consider applying Laplace’s approximation to the integral (2.7) under the assumption that both \( p(y_i|b_i) \) and \( p(b_i) \) are normal densities. In particular, the general form of Laplace’s approximation is, in our context, with \( b_i \) \((k \times 1)\),

\[
\int e^{n_i \xi_i(b_i)} \, db_i \approx \left( \frac{2\pi}{n_i} \right)^{k/2} \left| -\xi''_i(\hat{b}_i) \right|^{-1/2} e^{n_i \xi_i(\hat{b}_i)},
\]

where \( \hat{b}_i \) maximizes \( \xi_i(b_i) \); the approximation is derived from expanding \( \xi_i(b_i) \) to quadratic terms about \( \hat{b}_i \). It may be shown (e.g. Vonesh, 1996; Wolfinger and Lin, 1997) that, when \( \Lambda_i \) does not depend on \( \beta_i \), application of Equation (2.11) to the integral in (2.7) with \( \xi_i(b_i) = \log \{ p(y_i|b_i) p(b_i) \} / n_i \) yields an approximation for \( p(y_i) \) that is an \( n_i \)-variate normal distribution. Even if \( \Lambda_i \) does depend on \( \beta_i \), as in Equation (4.6), it may be shown (Ko and Davidian, 1999) that this result is unchanged under conditions relevant to applications like pharmacokinetics, so that \( p(y_i) \) satisfies

\[
-2 \log p(y_i) \approx \log |\hat{V}_i| + (\hat{w}_i - \hat{f}_i)' \hat{V}_i^{-1} (\hat{w}_i - \hat{f}_i) + n_i \log(2\pi),
\]

where \( \hat{w}_i = y_i + \hat{Z}_i \hat{b}_i \), \( \hat{f}_i = f_i \{ d(a_i, \beta, \hat{b}_i) \} \), \( \hat{Z}_i = Z_i \{ d(a_i, \beta, \hat{b}_i) \} \), and \( \hat{V}_i = \hat{Z}_i D \hat{Z}_i' + \sigma^2 \Lambda_i \{ d(a_i, \beta, \hat{b}_i), \gamma \} \). Equation (2.12) indicates that the approximate marginal mo-
ments of \( y_i \) are then given by

\[
E(y_i) \approx f_i\{d(a_i, \beta, \hat{b}_i)\} - Z_i \{d(a_i, \beta, \hat{b}_i)\}
\]

\[
\text{var}(y_i) \approx Z_i \{d(a_i, \beta, \hat{b}_i)\} D Z_i^\prime \{d(a_i, \beta, \hat{b}_i)\} + \sigma^2 A_i \{d(a_i, \beta, \hat{b}_i), \gamma \} \equiv V_i \{d(a_i, \beta, \hat{b}_i), (2.4.3)\}
\]

These results and variations on them are used to justify a number of fitting methods, e.g. the so-called First Order Conditional Expectation method (FOCE) in NONMEM (Beal and Sheiner, 1998), the algorithm of Lindstrom and Bates (1990) implemented in the S–Plus function nlm (Pinheiro and Bates, 1995), and the procedure obtained by using the EBLUP option in the SAS macro NLINMIX (Littell et al., 1996). The latter two methods again ignore the dependence of \( Z_i \) and \( \Lambda_i \) on \( \beta \), thus relying on estimating equations that are linear in \( y_i \) to estimate \( \beta \), while NONMEM maximizes the full normal log-likelihood in (2.12), thus using an estimating equation for \( \beta \) that is quadratic in the data. As with the First Order approximation, in applications like pharmacokinetics, all methods yield similar results. More detail on the derivation of these approximations and specifics of implementation may be found in Wolfinger and Lin (1997). An important difference between this and the First Order approximations may be appreciated by comparing Equations (2.10) and (2.13). The form of the approximate moments under the former does not strictly rely on the assumption of normality, as \( \hat{b}_i \) is simply set to zero. In contrast, the moments in (2.13) rely on normality in the sense that the form of \( \hat{b}_i \) results from the normality assumptions for \( p(y_i|b_i) \) and \( p(b_i) \) that make up \( \ell(b_i) \). From another perspective, the approximate moments in (2.13) might be thought to be more accurate approximations to
the true moments than those in (2.10), because they attempt to take into account individual-specific differences through $\hat{b}_i$.

**Previous Simulation Evidence**

Previous simulation studies of these methods and variations upon them appear both in the pharmacokinetic and statistical literature; an incomplete list includes Sheiner and Beal (1980), Davidian and Giltinan (1993), Roe (1997) and the ensuing discussion, Pinheiro and Bates (1995), and Wolfinger and Lin (1997). For example, these last authors compare these methods as implemented in the SAS macro `NLINMIX` and conclude that both produce reliable estimates, with the latter slightly outperforming the former at the cost of greater computing times and instability. In general, a theme that emerges from much of this work is that methods based on the Laplace approximation tend to yield more reliable results than those based on the First Order approximation when sampling within individuals is intensive. When data on individuals are sparse, there is no consensus regarding which type of method is to be preferred. All of these previous studies were carried out under the assumption of normal random effects and correctly specified model $d$, however, so whether these findings carry over to the cases of nonnormal random effects or misspecified models is unclear.
2.4 Design of simulations

Overview

In order to study the robustness of the two approximate methods to the assumptions of normal random effects and specification of the model \( d \) in Equation (2.2), we carried out simulation studies in which several factors were varied. As described in greater detail below, we considered a simple pharmacokinetic model from which data were generated for each of a number of subjects under different true distributions of the random effects and different sampling schemes for each subject, intensive or sparse. We also considered fitting in the case where the assumed inter-individual model \( d \) is different from that generating the data. Performance is evaluated with respect to bias and precision of estimation of the fixed parameters \( \beta \), \( D \), and \( \sigma \) and reliability of common hypothesis testing procedures to detect covariate effects (i.e. size and power). In all cases, the First Order and Laplace approximations were carried out using the implementations available in the SAS NLINMIX macro for SAS version 6.12; that is, the First Order approximation was implemented using the \texttt{ZERO} option and the Laplace using the \texttt{EBLUP} option.

In all cases, we focused on the one-compartment model \( f \) given in Equation (2.3) with dose \( D_i \equiv 100 \) for all subjects. This simple model was chosen so that, hopefully, model identification and parameterization issues would be less likely to confound interpretation of performance. For each of \( m = 50 \) subjects in each simulated data set, data were generated at \( n_i \) time points \( t_{ij} \) described below according to the model
in Equations (4.6) and (2.2), where, conditional on random effects \( b_i \) as described shortly, \( e_i \) was normally distributed with mean zero and covariance matrix \( \Lambda_i(\beta_i, \gamma) \), a diagonal matrix with diagonal elements as in (2.4). In all cases, the subject-specific parameters \( Cl_i \) and \( V_i \) were generated according to the model \( d \) given by

\[
\log Cl_i = \beta_1 + \frac{\beta_2}{100} a_i + b_{1i}, \quad a_i \sim N(70, 20^2), \quad \log V_i = \beta_3 + b_{2i},
\]

(2.14)

where the covariate \( a_i \) was generated from a normal distribution for each subject as shown and was meant to represent a covariate like creatinine clearance. True values of the parameters were \( \beta_1 = \log 1.25, \beta_3 = \log 10, \sigma = 0.15 \), and

\[
D = \begin{bmatrix}
0.086 & 0.052 \\
0.052 & 0.086
\end{bmatrix};
\]

this matrix was specified so that coefficient of variation in each of \( Cl_i \) and \( V_i \) is roughly 0.3 with correlation 0.6. The parameter \( \beta_2 \) characterizing the relationship between \( \log Cl_i \) and the covariate was taken to be equal to either zero or 0.6, depending on whether size of hypothesis testing procedures or power and quality of estimation were the focus.

Distribution of Random Effects

Several different distributions for \( b_i \) were used to generate random effects:

N. A normal distribution, \( b_i \sim N(0, D) \)

t. A heavier tailed distribution, \( b_i \sim t_5(0, D) \), the bivariate Student’s \( t \) distribution with 5 degrees of freedom and covariance matrix \( D \)
C1. A mildly contaminated normal distribution, $b_i \sim (1 - \alpha)N(0, D) + \alpha N(0, D^*)$, with contamination fraction $\alpha = 0.05$ and $D^*$ chosen as described below.

C2. A moderately contaminated normal distribution, $b_i \sim (1 - \alpha)N(0, D) + \alpha N(0, D^*)$, with contamination fraction $\alpha = 0.10$ and $D^*$ chosen as described below.

A. An asymmetric distribution for $\log Cl_i$ obtained by a mixture of normals given as $b_i \sim (1 - \alpha)N(\mu, D) + \alpha N(-\mu, D)$, with mixing proportion $\alpha = 0.3$ and $\mu$ chosen as described below.

B. A bimodal, symmetric distribution for $\log Cl_i$ obtained by a mixture of normals given as $b_i \sim (1 - \alpha)N(\mu, D) + \alpha N(-\mu, D)$, with mixing proportion $\alpha = 0.5$ and $\mu$ chosen as below.

Distribution N represents the case where the usual assumption of normality on the random effects is valid. Distribution t represents a situation where the true distribution of the random effects has heavier tails than expected from a normal distribution but with the same variability in the population. Distributions C1 and C2 are meant to characterize the situation where a proportion $\alpha$ of the subjects sampled are actually from a second population not of interest; distribution C1 represents mild (5%) such contamination, distribution C2 represents moderate (10%) contamination. In each case, the second population has covariance matrix

$$D^* = \begin{bmatrix} 0.223 & 0.134 \\ 0.134 & 0.223 \end{bmatrix}.$$
chosen so that the overall coefficients of variation in each component of the resulting
distribution of $Cl_i$ and $V_i$ is roughly 0.5 rather than 0.3. Thus, these distributions
represent cases where the apparent inter-subject variation is greater than that in
the population of interest due to errors in sampling. Distribution A is intended to
represent the situation where random effects are not symmetrically distributed; as
discussed earlier, this could be because the log transformation of model parameters
does not induce both linearity and normality. Distribution B characterizes a situation
where the underlying distribution is in fact symmetric but bimodal; as previously
noted, this might be the case if a key covariate is not available. For each distribution,
writing $D = R'R$ for $R$ an upper triangular matrix, $\mu = \{(sep/2)\sqrt{(R_{11}^2 + R_{22}^2)}, 0\}'$, 
where $sep = 2.5$ (4.0) for distribution A (B), which yields a resulting distribution for
$b_{1i}$ that is nonnormal while that for $b_{2i}$ is $N(0, D)$. Plots of distributions A and B
are shown in Fig. 2.1. In all, these six different distributions were used to investigate
varying violations of the assumption of normality of the random effects.

*Intra-subject Sampling*

Two different sampling schemes were used to investigate whether violations of
the assumptions of normality of $b_i$ and the nature of the model $d$ may have different
effects depending on whether individual sampling is intensive or sparse; the latter
situation is likely in a population pharmacokinetic study. For intensive sampling,
responses were generated for each of the $m = 50$ individuals at $n_i = 8$ time points
$t_{ij} = (0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 12.0, 24.0)$. For sparse sampling, where each subject
is potentially observed at only a few time points, responses were generated according to the following procedure, which resulted in anywhere from $n_i = 1$ to 8 samples post-dose, with an average of 5 per subject: For each of $m = 50$ subjects, the first observation was chosen at a time generated according to a uniform distribution on $(0.1, 0.3)$. The remaining $n_i - 1$ time points for the individual were generated by choosing $n_i - 1$ from a uniform distribution on $(1, 7)$. Then, $n_i - 1$ times were sampled from $(0.5, 1.0, 2.0, 4.0, 8.0, 12.0, 24.0)$. To determine which of these times were observed, an approximate Poisson distribution with $\lambda = 1$ was sampled without replacement for each observation. The approximate Poisson distribution used the following weights for the possible seven outcomes: $(0.3679, 0.3679, 0.1839, 0.0613, 0.0153, 0.0031, 0.0006).

**Inter-individual Model Misspecification**

In all cases, data were generated assuming the model given in (2.14). To address the issue of misspecification of the model $d$, simulated data were fitted two ways: assuming that model (2.14) holds, representing the case of correct specification, and assuming incorrectly that the relationship between the individual-specific parameters $(Cl_i, V_i)$ and covariates and random effects is instead given by

$$
Cl_i = \beta_1 + \frac{\beta_2}{100} a_i + b_{1i}, \quad V_i = \beta_3 + b_{2i}. \tag{2.15}
$$

Note that (2.15) represents both a misspecification in the manner in which clearance is associated with the covariate and the joint distribution of $(Cl_i, V_i)$. For the values of $\beta_1$ and $\beta_2 = 0.6$, the term $\exp(\beta_1 + \beta_2 a_i/100)$ in $Cl_i$ in the correct model (2.14) does not deviate too much from linearity; thus, (2.15) might reasonably be assumed in
practice as an alternative to (2.14) on the basis of graphical evidence from a fit with no
covariates as in Davidian and Gallant (1992) and Wakefield (1996). Because the First
Order approximation linearizes about zero, we might not expect the misspecification
to affect the results. In contrast, Laplace’s method depends on how random effects
enter the model, so it is apt to be more sensitive.

*Evaluation of Covariate Effects*

For each combination of distribution, sampling scheme, and specification of inter-
individual model, in addition to assessment of bias and precision of the estimators for
\(\beta, \sigma,\) and \(D,\) the performance of two standard hypothesis testing procedures to address
the null hypothesis \(H_0: \beta_2 = 0,\) representing the effect of the covariate in clearance,
was also evaluated. Both tests were conducted at nominal level of significance 0.05.
Both size and power of each test was evaluated. To evaluate size, data were generated
using \(\beta_2 = 0;\) for power, data were generated with \(\beta_2 = 0.6.\)

- **Wald test.** For the fit of the assumed model (2.14) or (2.15), the test statistic
  constructed by dividing the estimate of \(\beta_2\) divided by its estimated standard
  error as provided by the approximation used was compared to a t critical value
  with degrees of freedom as estimated in \texttt{PROC MIXED}.

- **Approximate Likelihood ratio test (LRT).** For each data set, it was necessary
to fit both the “reduced” model where \(\beta_2\) was taken to be equal to zero and
the “full” model where \(\beta_2\) was estimated. From these two fits for each method,
the approximate LRT statistic based on the approximate marginal likelihoods
routinely calculated by the software was formed. (Note that this is not quite a true LRT, even approximately, because fitting is carried out using a linear rather than quadratic estimating equation, ignoring the dependence of $Z_i$ and $\Lambda_i$ on $\beta$; however, this is standard practice.) The test statistic was compared to the appropriate critical value from a $\chi^2$ distribution.

**Simulation Scenarios**

A simulation scenario thus consisted of a particular choice of distribution, intra-subject sampling scheme (intensive or sparse), specification of model $d$ [correct as in (2.14) or incorrect as in (2.15)], and true covariate effect (null, $\beta_2 = 0$, for size of the tests, or alternative, $\beta_2 = 0.6$, for power of the tests). The full factorial of $6 \times 2 \times 2 \times 2 = 48$ possible combinations was investigated, where, for each scenario, 1000 Monte Carlo data sets, each with $m = 50$ subjects, were generated. For each data set in each scenario, fitting was carried out using both the First Order approximation and Laplace's approximation as described above.

### 2.5 Results

The simulations were conducted on several different platforms. The majority of the simulations were completed in the Unix environment on Sun Sparc 4, Sparc 5, Ultra 1, and Ultra 60 workstations. A few simulations were also performed on a Gateway Solo 9100, Pentium 166, laptop with 64 Meg ram, but this PC environment proved
inconvenient. Due to these complications, the PC environment was abandoned in favor of the Unix workstations. Consistency of results was examined across platforms to ensure that conclusions were the result of properties of the methods rather than numerical anomalies.

Considering all the factors [distribution (6), approximate method (2), full and reduced models (2), size and power of the test (2), intensively or sparsely sampled data (2), correctly or incorrectly specifying the model $d$ (2), number of data sets per scenario (1000)], there were a total of 192,000 calls to the SAS macro NLINMIX. As is frequently the case in fitting nonlinear mixed effects models by any method, convergence and other numerical problems were encountered. Whenever numerical errors occurred, the trial was repeated with efforts to identify and correct the problem. After these efforts were made, Of the 192,000 calls to NLINMIX, there were several nonconvergent data sets, and a floating point overflow error occurred in 21. These trials were categorized as nonconvergent. In all cases, starting values were taken as the true values generating the data to allow automation of this large number of simulations. As is familiar to analysts using these methods, it is sometimes the case that convergence can never be achieved. This may be the result of poor initial values, practical lack of identifiability with the particular data available, or other unknown factors; typical practice is to try several sets of starting values to address the first issue. Due to the large number of data sets processed, it was only possible to make limited attempts to emulate this “real” practice for initially nonconvergent data sets, which
did not improve convergence. The number of data sets (of 1000) for which satisfactory convergence ultimately was obtained for each scenario and method are shown in Table 2.1 for the case where the correct inter-individual model (2.14) was fitted and in Table 2.2 for the case where the incorrect model (2.15) was fitted. Consequently, the results reported below may be pessimistic with respect to possible convergence in practice and must be interpreted with this caution in mind. However, the way in which nonconvergence manifested itself across scenarios described below, suggests that it is not unreasonable to attribute much of the problem to model misspecification.

Table 2.1 shows that when the inter-individual model is correctly specified, hence the only deviation from assumptions is in the distribution of $b_i$, the First Order approximation achieves very high rates of convergence in almost all cases, although this is more uniform for intensive sampling. When the data are sparse, this method exhibits difficulty in achieving convergence when the true random effects distribution is far from normality (A and B) when covariate effects are present. In contrast, the convergence of the Laplace method seems much more strongly tied to how the true distribution of random effects deviates from normality, and convergence seems to be more difficult to achieve. This is especially true with the most profound departures from normality, distributions A and B, both for sparse and intensive data. Intuitively, from the previous discussion, this effect of departure from normality is not surprising, as this method, in contrast to the First Order method, relies more heavily on normality in the approximation. Table 2.2 shows the same qualitative pattern when the inter-
individual model is misspecified; however, the failure to achieve convergence with the Laplace approximation is now more profound. A possible interpretation is that, unlike the First-Order approximation, the Laplace approximation attempts to make the approximate moments “individual-specific” [compare Equations (2.10) and (2.13)]. Perhaps when the way in which the individual-specific contribution \( b_i \) enters the model is misspecified, this feature becomes a drawback, as it results in an incorrect “individual-specification” arising from the interplay of the functional dependence on \( b_i \) and the incorrect implication about the distribution of \( Cl_i \) and \( V_i \).

These convergence results suggest that the First Order approximation, whose approximate moments do not depend strictly on normality, is computationally less sensitive to deviations from underlying assumptions than the Laplace approximation, regardless of sampling scheme. One might be tempted to infer in practice that failure to achieve convergence, in particular with the Laplace approximation, may be taken as evidence of violation of underlying assumptions. This may well be true in many instances; however, other issues (e.g. model identification and complexity) may also play a role. In any event, the poor convergence rates in many instances here must be borne in mind in interpreting the following results.

We next consider how the quality of estimation is affected by deviations from the normality assumption when the correct model is specified. Table 2.3 and Figs. 2.2-2.7 summarize performance in terms of bias, efficiency, and sampling distribution of estimators. All of these results refer to fitting of the model when data were generated
with $\beta_2 = 0$, the correct inter-individual model (2.14) was fitted (“reduced model”); this case was chosen over the case $\beta = 0.6$ owing to the overall higher convergence rates. Results fitting the model with $\beta_2 = 0.6$ and the correct model (2.14) were qualitatively similar. Figs. 2–7 show graphically the sampling distributions for the estimators of each fixed model parameter for both methods. Because of the poor performance of the Laplace approximation when this fitted inter-individual model is incorrect, we do not attempt to interpret these results.

The parameters $\beta_1$ and $\beta_3$ characterize the means of the distributions of $\log Cl_i$ and $\log V_i$, both of which enter the true model in a highly nonlinear fashion. The elements of $D$ characterize the variation in these distributions. A general implication of Figs. 2–6 is that estimation of central tendency (means) of the (log) pharmacokinetic parameters is little affected by deviation from normality of the random effects, while that of variation in these parameters is more profoundly affected. Intuitively, this is not too surprising; even in simple problems, estimation of means is relatively robust to underlying assumptions while that of variances is often not.

From Fig. 2.2, estimation of $\beta_1$ is mostly unbiased for the Laplace method and intensive data with the exception of the asymmetric distribution. Estimation using the First Order method yielded biased estimates overall. The assumption of normal sampling distributions seems plausible in most cases. Quality of estimation seems fairly insensitive to random effects distribution except for the distributions that deviate most profoundly from normality (A and B). Fig. 2.3 shows results for estimation
of $\beta_3$. Overall, estimates for the most part exhibit little bias (but recall the low convergence rate for the Laplace approximation in some cases). Table 2.3 gives numerical summaries of relative efficiency of the First Order approximation to the Laplace approximation for these parameters in the convergent cases. Interestingly, despite the latter approximation’s dependence on normality, this method dominates the First Order method in most cases, even when the distribution is far from normality. A possible interpretation is that it is better, at least for this setting, to use approximate moments that attempt to be “individual-specific,” even if the approach to calculating the individual-specific component ($\hat{b}_i$) is predicated on normality. As long as the true underlying distribution is still unimodal in particular, estimation is not terribly compromised. A referee has speculated that perhaps data sets exhibiting convergence for the Laplace approach are by chance “more Gaussian” than those that failed to achieve convergence with this method.

This same theme carries over to estimation of the covariance components $D$ and $\sigma$. From Figs. 4–6, note that the sampling distributions of the estimators of elements of $D$ are markedly skewed in many cases except in those where the $b_i$ are normal; moreover, deviations from normality often result in much more variable estimation for both methods. A similar pattern is evident for estimation of $\sigma$, see Fig. 2.7. Table 2.3 compares the two approximations on the basis of efficiency; these results should be interpreted with caution when both methods perform poorly. Again, the Laplace approximation yields greater efficiency even under departures from normal-
ity. One exception is in estimation of $D_{11}$, associated with variation in $Cl_i$, when the true underlying distribution is bimodal. The Laplace approximation, which uses the normality assumption to “estimate” the $b_i$, effectively must interpret the underlying bimodal distribution as normal, hence in principle is overlaying the bimodal distribution with a normal distribution with larger variance. The First Order method, on the other hand, need not make such a compromise. It is important to keep in mind that in some instances these results are based on fewer than 1000 data sets due to nonconvergence, so are likely optimistic in favor of the Laplace approximation (Table 2.1). The results do suggest, if convergence is achieved, efficiency of estimation is greater with the Laplace approximation, regardless of the true random effects distribution. It appears that obtaining more “individually-tailored” approximate moments is of some importance, even if normality is assumed, perhaps incorrectly, to do this.

Table 2.4 summarizes the performance of the Wald and LRT testing procedures for $H_0 : \beta_2 = 0$ when the correct inter-individual model (2.14) is fitted. The Wald test requires fitting with no restriction on $\beta_2$ (“full model”) while the LRT requires fitting both “full” and “reduced” models. Only trials where convergence was achieved for both “full” and “reduced” models, allowing implementation of both tests, are included. Of considerable interest is the poor performance of the LRT for both sampling schemes when it is based on the First Order approximation; the rejection rate is substantially greater than the nominal level of significance and becomes worse the further the true random effects distribution deviates from normality. In contrast,
rejection rates are more in line with the nominal level for the Laplace approximation, although the test is slightly liberal with sparse data. Intuitively, the behavior may not be so surprising. For both approximations, the LRT is based on assuming that the marginal distribution of $y_i$ is normal with first two moments given by (2.10) and (2.13). This may be reasonable for the “individually-tailored” Laplace approximation, even when the underlying $b_i$ are not normal ($p(y_i|b_i)$ still is, which may be enough). It is likely a poor approximation with $b_i \equiv 0$, even when the true random effects distribution is indeed normal. In contrast, the Wald test uses only the approximate sampling distribution of the estimator. From Table 2.4, the Wald test based on either approximation yields a rejection rate close to the nominal level with intensive data when the true random effects distribution is close to normal (N or t); otherwise, it is slightly liberal.

Table 2.4 also gives results for the power of these tests for detecting a moderate departure from $H_0$, $\beta_2 = 0.6$. Power in the case of the LRT based on the First Order approximation is not relevant, because of the high rejection rate under the null. Otherwise, for the Wald tests, power is in most instances slightly better using the Laplace approximation, with comparable power from using the LRT, which is valid in this case. As expected, power is higher overall when intensive sampling is used. Note that power drops substantially for both sampling schemes as the true distribution of the random effects deviates from normality.

In Table 2.5, we report the same information when the incorrect inter-individual
model (2.15) was fitted. Here, care must be taken in interpretation, as the Laplace approximation exhibited poor convergence under these conditions. The results concern only the ability to detect a qualitative covariate effect when the inter-individual model involves both incorrect dependence on covariate and incorrect implied distribution of $Cl_i$ and $V_i$. Recall that the form of the model in (2.14) for $Cl_i$ is not too far from linearity across the range of the covariate; thus, it is perhaps not too surprising that the results are similar to those in Table 2.4. Interestingly, the fact that $b_i$ enters the model incorrectly in every instance has only a slight effect on the results.

2.6 Discussion

We have reported the results of a simulation study undertaken to gain insight into the consequences of violation of assumptions underlying two popular approximations used in fitting nonlinear mixed effects models. Although it is not appropriate to draw general conclusions from a single simulation study, the results suggest some interesting features that may be worthy of further investigation. It appears that estimation of fixed parameters based on the Laplace approximation is fairly robust to mild deviations from normality of the random effects distribution, although it can show difficulty in achieving convergence. This method’s convergence is also sensitive to misspecification of the inter-individual model, whereas that of the First Order approximation is less so. In cases where both methods are feasible, the Laplace approximation seems to achieve more precise parameter estimation, and both Wald and
LRT tests for covariate effects are reliable, although they can be somewhat liberal when the random effects are not normal. It would be of interest to learn whether these observations hold true for other models $f$, magnitudes of variation, and other features. For example, previous studies for linear mixed models suggest that estimation of conditionally linear parameters in the present context may be less affected by nonnormality of random effects than parameters that enter $f$ nonlinearly. Future simulation studies extending this work should address these issues. Moreover, work is needed to gain insight into reasons for the observed convergence rates under departures from normality. For example, higher convergence rates for the Laplace approximation with sparse relative to intensive data may suggest that the estimated $\hat{b}_i$ may be near zero in these cases.

This study focused on estimation of fixed parameters, most of which $(\beta, D)$ characterize the population. An issue that this study has not addressed is the consequences for “individual estimation,” that is, for prediction of individual-specific parameters or conditional means. It is not unreasonable to expect that misspecification of the random effects distribution or inter-individual model may have more profound effects.

Overall, the results suggest some caveats. Inference based on these methods can be sensitive to underlying distributional and model misspecifications. Operating characteristics of hypothesis tests can be greatly affected. Overall the results highlight the inherent difficulty in specifying any type of complex model with latent unobservable components, a difficulty that suggests that caution is in order in interpreting both the
nature of computational problems and results in the event convergence is achieved. As such models see even greater widespread use, there is an urgent need for further research in this area. Of particular value will be to explore whether methods that do not rely on normality assumptions for the random effects (e.g. Davidian and Gallant, 1993) can produce more reliable inferences and be made sufficiently computationally stable and accessible to facilitate widespread use.

References


Figure 2.1: Perspective plots of the joint density functions corresponding to the asymmetric (A) and bimodal (B) random effects distributions.
Figure 2.2: Boxplots of estimates of $\beta_1$ from data sets for which both methods converged in each panel for each random effects distribution, correct inter-individual model fitted. Estimates have been centered and scaled by the true value so that 0 represents no bias and the vertical axis represents bias relative to the truth. Multiplying the numbers on the vertical axis will result in relative percentages. The white bar represents the median, the horizontal bars represent the extremes. In some cases, outliers that distorted the comparison have been deleted; the relative difference of these estimates from the truth is given numerically.
Figure 2.3: Same as Fig. 2.2 for $\beta_3$. 
Figure 2.4: Boxplots of estimates of $D_{11}$ from data sets for which both methods converged in each panel for each random effects distribution, correct inter-individual model fitted. Estimates have been centered and scaled as in Fig. 2. The white bar represents the median, the horizontal bars represent actual estimates. In some cases, outliers that distorted the comparison have been deleted; the relative difference of these estimates from the truth is given numerically.
Figure 2.5: Same as Fig. 2.4 for $D_{12}$. 
Figure 2.6: Same as Fig. 2.4 for $D_{22}$. 
Figure 2.7: Same as Fig. 2.4 for $\sigma$. 
Table 2.1: Rate of nonconvergence out of 1000 trials for each distribution, sampling scheme, and fitting method, correct inter-individual model.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Model</td>
<td>Red Model</td>
</tr>
<tr>
<td></td>
<td>Number Converging</td>
<td>Number Converging</td>
</tr>
<tr>
<td>Intensive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>First Order</td>
<td>1(^1)</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>1(^1)</td>
</tr>
<tr>
<td>t</td>
<td>First Order</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>27</td>
</tr>
<tr>
<td>C1</td>
<td>First Order</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>3</td>
</tr>
<tr>
<td>C2</td>
<td>First Order</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>82</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>First Order</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>50</td>
</tr>
<tr>
<td>Bimodal</td>
<td>First Order</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>3</td>
</tr>
<tr>
<td>Sparse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>First Order</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>3</td>
</tr>
<tr>
<td>t</td>
<td>First Order</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>41</td>
</tr>
<tr>
<td>C1</td>
<td>First Order</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>12</td>
</tr>
<tr>
<td>C2</td>
<td>First Order</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>103</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>First Order</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>114</td>
</tr>
<tr>
<td>Bimodal</td>
<td>First Order</td>
<td>61(^2)</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>305(^3)</td>
</tr>
</tbody>
</table>

1. A floating point error occurred for the 1 trial that did not converge.
2. A floating point error occurred in 11 trials out of the 61 that did not converge.
3. A floating point error occurred in 7 trials out of the 305 that did not converge.
4. A floating point error occurred in 1 trial out of the 275 that did not converge.
Table 2.2: Rate of nonconvergence out of 1000 trials for each distribution, sampling scheme, incorrect inter-individual model.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Intensive</th>
<th>Sparse</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size, $\beta_2 = 0$</td>
<td>Power, $\beta_2 = 0.6$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full Model</td>
<td>Reduced Model</td>
<td>Full Model</td>
</tr>
<tr>
<td></td>
<td>Number Converging</td>
<td>Number Converging</td>
<td>Number Converging</td>
</tr>
<tr>
<td>Normal</td>
<td>First Order</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>451</td>
<td>242</td>
</tr>
<tr>
<td>t</td>
<td>First Order</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>562</td>
<td>389</td>
</tr>
<tr>
<td>C1</td>
<td>First Order</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>558</td>
<td>298</td>
</tr>
<tr>
<td>C2</td>
<td>First Order</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>753</td>
<td>540</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>First Order</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>353</td>
<td>40</td>
</tr>
<tr>
<td>Bimodal</td>
<td>First Order</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>388</td>
<td>122</td>
</tr>
</tbody>
</table>

| Normal       | First Order | 2 | 1 | 0 | 0 |
|              | Laplace | 292 | 208 | 205 | 225 |
| t            | First Order | 10 | 5 | 6 | 2 |
|              | Laplace | 399 | 336 | 308 | 342 |
| C1           | First Order | 0 | 0 | 2 | 0 |
|              | Laplace | 371 | 283 | 276 | 303 |
| C2           | First Order | 2 | 0 | 14 | 7 |
|              | Laplace | 614 | 511 | 539 | 561 |
| Asymmetric   | First Order | 5 | 1 | 181 | 138 |
|              | Laplace | 265 | 176 | 199 | 165 |
| Bimodal      | First Order | 26 | 1 | 132 | 82 |
|              | Laplace | 299 | 275 | 201 | 115 |
Table 2.3: Relative Efficiency of the First Order method relative to the Laplace method, correct inter-individual model with $\beta_2 = 0$. Each entry is based on all data sets for which convergence was achieved for both methods. Entries are given in italics in cases where both methods performed poorly (see Figs. 2–7).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.45</td>
<td>0.43</td>
<td>0.47</td>
<td>0.36</td>
<td>4.87</td>
<td>0.76</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.78</td>
<td>0.74</td>
<td>0.84</td>
<td>0.69</td>
<td>0.35</td>
<td>0.87</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.75</td>
<td>0.02</td>
<td>0.52</td>
<td>0.11</td>
<td>1.12</td>
<td>3.16</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.73</td>
<td>0.04</td>
<td>0.60</td>
<td>0.16</td>
<td>0.36</td>
<td>0.92</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.71</td>
<td>0.06</td>
<td>0.61</td>
<td>0.23</td>
<td>0.52</td>
<td>0.91</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.15</td>
<td>0.04</td>
<td>0.12</td>
<td>0.03</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Sparse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.94</td>
<td>0.99</td>
<td>0.88</td>
<td>0.90</td>
<td>2.07</td>
<td>2.64</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.61</td>
<td>0.63</td>
<td>0.58</td>
<td>0.50</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.76</td>
<td>0.09</td>
<td>0.54</td>
<td>0.31</td>
<td>0.82</td>
<td>1.24</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.77</td>
<td>0.11</td>
<td>0.64</td>
<td>0.28</td>
<td>0.24</td>
<td>0.41</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.76</td>
<td>0.22</td>
<td>0.51</td>
<td>0.16</td>
<td>0.55</td>
<td>0.48</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.34</td>
<td>0.30</td>
<td>0.29</td>
<td>0.09</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Values less than 1 for relative efficiency favor the use of the Laplace method.
Table 2.4: Performance of Wald and Likelihood Ratio Tests on trials that converged for both the full and reduced models, correct inter-individual model.

<table>
<thead>
<tr>
<th>Data Composition</th>
<th>Distribution</th>
<th>Number of Trials Converging</th>
<th>Number Rejecting</th>
<th>Percent Rejecting</th>
<th>Number of Trials Converging</th>
<th>Number Rejecting</th>
<th>Percent Rejecting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald/LRT</td>
<td>Wald/LRT</td>
<td></td>
<td>Wald/LRT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensive</td>
<td>Normal</td>
<td>First Order</td>
<td>999</td>
<td>96/270</td>
<td>5.6/27.0</td>
<td>1000</td>
<td>870/837</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>999</td>
<td>55/50</td>
<td>5.5/5.0</td>
<td>999</td>
<td>818/772</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>999</td>
<td>55/47</td>
<td>5.7/4.9</td>
<td>999</td>
<td>823/815</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>999</td>
<td>77/72</td>
<td>7.8/7.3</td>
<td>956</td>
<td>833/821</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>First Order</td>
<td>1000</td>
<td>75/283</td>
<td>7.5/28.3</td>
<td>1000</td>
<td>818/805</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>1000</td>
<td>74/336</td>
<td>7.4/33.6</td>
<td>997</td>
<td>685/715</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>1000</td>
<td>61/50</td>
<td>6.8/5.6</td>
<td>783</td>
<td>628/613</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>First Order</td>
<td>999</td>
<td>62/443</td>
<td>6.2/44.3</td>
<td>988</td>
<td>352/457</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>918</td>
<td>48/41</td>
<td>5.2/4.5</td>
<td>267</td>
<td>89/86*</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>Asymmetric</td>
<td>1000</td>
<td>63/306</td>
<td>6.3/30.7</td>
<td>999</td>
<td>585/538</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>998</td>
<td>60/55</td>
<td>6.0/5.5</td>
<td>955</td>
<td>585/560</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>997</td>
<td>86/74</td>
<td>8.6/7.4</td>
<td>999</td>
<td>709/686</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>First Order</td>
<td>999</td>
<td>93/206</td>
<td>9.3/20.6</td>
<td>999</td>
<td>660/684</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asymmetric</td>
<td>997</td>
<td>86/74</td>
<td>8.6/7.4</td>
<td>999</td>
<td>660/786</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>First Order</td>
<td>999</td>
<td>93/189</td>
<td>9.4/19.0</td>
<td>999</td>
<td>690/719</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>999</td>
<td>74/68</td>
<td>7.7/7.1</td>
<td>999</td>
<td>690/675</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>1000</td>
<td>96/216</td>
<td>9.6/21.6</td>
<td>996</td>
<td>653/668</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>996</td>
<td>81/69</td>
<td>8.2/7.0</td>
<td>991</td>
<td>692/669</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>Asymmetric</td>
<td>997</td>
<td>72/58</td>
<td>8.1/6.5</td>
<td>987</td>
<td>525/832</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>996</td>
<td>92/256</td>
<td>9.2/25.7</td>
<td>980</td>
<td>544/927</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>888</td>
<td>75/51</td>
<td>7.8/6.5</td>
<td>711</td>
<td>227/219*</td>
</tr>
<tr>
<td></td>
<td>Asymmetric</td>
<td>First Order</td>
<td>986</td>
<td>78/446</td>
<td>7.9/45.2</td>
<td>617</td>
<td>166/299*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>851</td>
<td>66/55</td>
<td>7.8/6.5</td>
<td>711</td>
<td>227/219*</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>First Order</td>
<td>851</td>
<td>66/55</td>
<td>7.8/6.5</td>
<td>711</td>
<td>227/219*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laplace</td>
<td>927</td>
<td>73/353</td>
<td>7.9/38.1</td>
<td>643</td>
<td>99/306*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First Order</td>
<td>650</td>
<td>42/40*</td>
<td>6.5/6.2</td>
<td>589</td>
<td>102/96*</td>
</tr>
</tbody>
</table>

* These values have been calculated from trials with low convergence rates.
Table 2.5: Performance of Wald and Likelihood Ratio Tests on trials that converged for both the full and reduced models, incorrect inter-individual model.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Distribution</th>
<th>Size of Test</th>
<th>Power of Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number of Trials Converging</td>
<td>Number of Trials Rejecting</td>
</tr>
<tr>
<td>Intensive</td>
<td>Normal</td>
<td>1000</td>
<td>62/282</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>549*</td>
<td>28/26*</td>
</tr>
<tr>
<td>t</td>
<td>First Order</td>
<td>1000</td>
<td>67/299</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>437*</td>
<td>22/19*</td>
</tr>
<tr>
<td>C.N. .05</td>
<td>First Order</td>
<td>1000</td>
<td>87/279</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>442*</td>
<td>43/38*</td>
</tr>
<tr>
<td>C.N. .10</td>
<td>First Order</td>
<td>996</td>
<td>87/350</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>247*</td>
<td>15/14*</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>First Order</td>
<td>1000</td>
<td>77/435</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>642*</td>
<td>39/33*</td>
</tr>
<tr>
<td>Bimodal</td>
<td>First Order</td>
<td>1000</td>
<td>67/301</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>612*</td>
<td>42/38*</td>
</tr>
<tr>
<td>Sparse</td>
<td>Normal</td>
<td>998</td>
<td>97/209</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>698*</td>
<td>58/47*</td>
</tr>
<tr>
<td>t</td>
<td>First Order</td>
<td>990</td>
<td>92/207</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>596*</td>
<td>46/39*</td>
</tr>
<tr>
<td>C.N. .05</td>
<td>First Order</td>
<td>1000</td>
<td>98/214</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>618*</td>
<td>47/38*</td>
</tr>
<tr>
<td>C.N. .10</td>
<td>First Order</td>
<td>998</td>
<td>110/254</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>369*</td>
<td>27/25*</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>First Order</td>
<td>994</td>
<td>80/433</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>705*</td>
<td>47/40*</td>
</tr>
<tr>
<td>Bimodal</td>
<td>First Order</td>
<td>974</td>
<td>80/329</td>
</tr>
<tr>
<td></td>
<td>Laplace</td>
<td>644*</td>
<td>47/43*</td>
</tr>
</tbody>
</table>

* These values have been calculated from trials with low convergence rates.
Chapter 3

Comparison of Numerical Integration Methods

3.1 Introduction

Because the likelihood for nonlinear mixed models must be approximated with one of several methods in practice today, it is appropriate to investigate how accurate these methods are. These methods were discussed in Chapter 1, and include Laplace’s approximation, importance sampling, and Gaussian quadrature.

Pinheiro and Bates (1995) derived estimating equations for the log–likelihood using these methods. They performed a simulation experiment to compare the performance of methods that estimate the parameters of a nonlinear mixed effects model. In their experiment, they chose only those methods based on maximum likelihood
and restricted maximum likelihood. They wrote computer code available in StatLib using S–Plus for each of the methods they compared.

Pinheiro and Bates implemented these methods by finding closed form expressions for approximate log-likelihoods. These expressions are maximized resulting in approximate maximum likelihood estimates. Of particular interest, they used the idea of centering for Gaussian quadrature resulting in their method called adaptive Gaussian quadrature. In this Chapter, the closed form expressions of Pinheiro and Bates are examined and new versions of these methods are derived which incorporate centering for not only Gaussian quadrature but also for importance sampling. Also, the idea of scaling for both importance sampling and Gaussian quadrature is implemented.

To expand on the simulation of Pinheiro and Bates, an experiment was performed to check the accuracy of their approximations, including those introduced here which center and scale. The closed form expressions were used to find approximations of the log-likelihood at different points in the parameter space. Pinheiro and Bates found the method of First Order Linearization to be not as accurate although it is not as computationally challenging. They concluded that it could be used for starting values for other, more accurate approximate methods. Due to their findings, First Order Linearization was not included in this simulation experiment. Pinheiro and Bates also concluded that of the methods they compared, the Laplacian and adaptive Gaussian quadrature methods gave the best mix of efficiency and accuracy. The goal
of this experiment was to test how further incorporation of centering and scaling could improve estimation of the log–likelihood.

Another concern in this experiment was the model. The model used by Pinheiro and Bates stipulated a constant error variance. The model used in this experiment allowed for error variance that is a function of the random coefficients, $\beta_i$. The closed form expressions used by Pinheiro and Bates do not take this into consideration and had to be modified. Unfortunately, software available today does not make this distinction and the user is forced to specify a constant error variance or use a weighted least squares approach for estimation. In this experiment, estimation of the log–likelihood was performed with both the “uncorrected” and “corrected” versions of the estimating equations to determine its effect.

Data were simulated from three NLMMs assuming normal distributions for the random effects. To approximate the log–likelihood at several locations in the parameter space, good starting values were needed to narrow the region of interest to a reasonable size. The first order method was used to obtain estimates for the parameters using the SAS procedure \texttt{NLMIXED}. Standard errors from this fitting were used in determining the region of points in the parameter space to be sampled.

Section 3.2 describes the various approximate methods which were investigated. The models and simulation are reported in Section 3.3. In Section 3.4 the method for sampling points from the parameter space is given. The results of the experiment are analyzed in Section 3.5.
3.2 Integration Methods

The integration methods investigated were Laplace’s approximation, importance sampling, and Gaussian quadrature. In this section the estimating equations for the methods used by Pinheiro and Bates are examined. New versions are derived that incorporate the ideas of both centering and scaling as well as correct for nonconstant error variance are derived.

3.2.1 Laplace’s Approximation

In Section 1.3.2, an approximation to minus two times the log-likelihood was derived for when $R_i$ does not depend on $b_i$. Furthermore, this derivation neglected a Hessian term. A second approximation was derived to show how neglecting this Hessian term affects the accuracy of the estimation. This comparison was done both for models with constant variance and nonconstant variance.

Assume a model with a more complicated structure for $R_i$, such as

$$R_i(\beta_i, \xi) = \sigma^2 \text{diag}\left\{ f_i^{2\beta}[d(a_i, \beta, b_i)] \right\}, \; \xi = (\sigma, \theta)^T. \tag{3.1}$$

Another version of Laplace’s approximation considered here takes into consideration how $R_i$ is a function of $b_i$ when differentiating to find $| - \ell''(\hat{b}_i) |$. The following is the derivation of $\ell''(\hat{b}_i)$ for this $R_i$ that also does not neglect the Hessian term as in 1.3.2.

Let

$$h_1 = \frac{1}{n_i} \ln \left\{ \frac{1}{(2\pi)^{n_i+k}/2} |D^{-1/2}|R_i^{-1}[d(a_i, \beta, b_i), \xi]|^{1/2} \right\}$$
\[ h_2 = \frac{1}{2n_i} \{ y_i - f [d(a_i, \beta, b_i)] \}^T R_i^{-1} [d(a_i, \beta, b_i), \xi] \{ y_i - f [d(a_i, \beta, b_i)] \} \]
\[ h_3 = \frac{1}{2n_i} b_i^T D^{-1} b_i \]

Then
\[ h'_1 = \frac{\partial}{\partial b_i} \frac{1}{n_i} \ln \left\{ \frac{1}{(2\pi)^{(n_i+k)/2}} |D^{-1}|^{1/2} |R_i^{-1} [d(a_i, \beta, b_i), \xi]|^{1/2} \right\} \]
\[ = \frac{\partial}{\partial b_i} \frac{1}{n_i} \ln |R_i^{-1} [d(a_i, \beta, b_i), \xi]|^{1/2} \]
\[ = - \frac{\partial}{\partial b_i} \frac{1}{2n_i} (n_i \ln \sigma^2 + \ln \prod_{j=1}^{n_i} f_{ij}^{20}) \]

where \( f_{ij} \) is the \( j \)th element of \( f_i [d(a_i, \beta, b_i)] \)
\[ = - \frac{\partial}{\partial b_i} \frac{1}{2n_i} 2\theta \sum_{j=1}^{n_i} \ln f_{ij} \]
\[ = - \theta \sum_{j=1}^{n_i} f_{ij}^{-1} \frac{\partial f_{ij}}{\partial b_i} \]

\[ h''_1 = - \frac{\theta}{n_i} \frac{\partial}{\partial b_i^T} \frac{\partial}{\partial b_i} \sum_{j=1}^{n_i} f_{ij}^{-1} \frac{\partial f_{ij}}{\partial b_i} \]
\[ = - \frac{\theta}{n_i} \sum_{j=1}^{n_i} \left( - f_{ij}^{-2} \frac{\partial f_{ij}}{\partial b_i} \frac{\partial f_{ij}}{\partial b_i} + f_{ij}^{-1} \frac{\partial^2 f_{ij}}{\partial b_i \partial b_i^T} \right) \]

\[ h_2' = \frac{\partial}{\partial b_i} \frac{1}{2n_i} \{ y_i - f [d(a_i, \beta, b_i)] \}^T R_i^{-1} [d(a_i, \beta, b_i), \xi] \{ y_i - f [d(a_i, \beta, b_i)] \} \]
\[ = \frac{1}{2n_i} \left( - \frac{2}{\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - f_{ij}) \left[ (1 - \theta) f_{ij}^{-20} + \theta y_{ij} f_{ij}^{-20-1} \right] \frac{\partial f_{ij}}{\partial b_i} \right) \]
\[ = - \frac{1}{n_i \sigma^2} \sum_{j=1}^{n_i} (y_{ij} - f_{ij}) \left[ (1 - \theta) f_{ij}^{-20} + \theta y_{ij} f_{ij}^{-20-1} \right] \frac{\partial f_{ij}}{\partial b_i} \]

\[ h''_2 = - \frac{1}{n_i \sigma^2} \frac{\partial}{\partial b_i^T} \sum_{j=1}^{n_i} (y_{ij} - f_{ij}) \left[ (1 - \theta) f_{ij}^{-20} + \theta y_{ij} f_{ij}^{-20-1} \right] \frac{\partial f_{ij}}{\partial b_i} \]
\[ h'_3 = \frac{\partial}{\partial b_i} - \frac{1}{2n_i} b_i^T D^{-1} b_i = \frac{1}{n_i} D^{-1} b_i \]

\[ h''_3 = \frac{\partial}{\partial b_i} - \frac{1}{n_i} \frac{1}{\partial b_i^T} D^{-1} b_i \]

\[ \ell''_i = h''_1 + h''_2 + h''_3 \]

\[ \ell''_i = -\frac{1}{n_i} \sum_{j=1}^{n_i} \left( \left[ \theta^2 \left( -2f_{ij}^2 + 4y_{ij}f_{ij} - 2y_{ij}^2 \right) \right] + \theta \left( 3f_{ij}^2 - 2y_{ij}f_{ij} - y_{ij}^2 \right) \right) \frac{\partial f_{ij}}{\partial b_i} \frac{\partial f_{ij}}{\partial b_i^T} + \left\{ \theta f_{ij}^{-1} + \frac{1}{\sigma^2} (y_{ij} - f_{ij}) \left[ (1 - \theta) f_{ij}^2 + \theta y_{ij}f_{ij} f_{ij}^{-2} \right] \frac{\partial^2 f_{ij}}{\partial b_i \partial b_i^T} \right\} + D^{-1} \]

When \( \theta = 1 \),

\[ \ell''_i = -\frac{1}{n_i} \sum_{j=1}^{n_i} \left\{ \left[ -f_{ij}^2 + \frac{1}{\sigma^2} (-3y_{ij} + 2f_{ij}) y_{ij} f_{ij}^{-4} \right] \frac{\partial f_{ij}}{\partial b_i} \frac{\partial f_{ij}}{\partial b_i^T} + \left[ f_{ij}^{-1} + \frac{1}{\sigma^2} (y_{ij} - f_{ij}) y_{ij} f_{ij}^{-3} \right] \frac{\partial^2 f_{ij}}{\partial b_i \partial b_i^T} \right\} + D^{-1} \)
The result is the following approximation of minus two times the log-likelihood, i.e. the deviance:

\[ k \ln(n_i) + n_i \ln(2\pi) + \ln\left| -\ell''(\hat{b}_i) \right| + \ln |D| + n_i \ln \sigma^2 + \sum_{j=1}^{n_i} \ln f_{ij}^2 + \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \hat{f}_{ij})^2 f_{ij}^{-2} + \hat{b}_i^T D^{-1} \hat{b}_i \]

This derivation illustrates how complex the approximation is for models straying from constant variance. But in practice, if numerical derivatives are used to find \( \ell''(\hat{b}_i) \) then analytic expressions become unnecessary.

For this experiment, both the version of Laplace’s approximation introduced in Section 1.3.2 and the “corrected” version were implemented. In general, the “corrected” version of Laplace’s approximation to the log-likelihood for the \( i \)th subject is

\[ \ln L_i \approx \ell_{i,\text{laplace}}^c = \frac{k}{2} \ln(2\pi) - \frac{k}{2} \ln n_i - \frac{1}{2} \ln \left| -\ell''(\hat{b}_i) \right| + n_i \ell_i(\hat{b}_i) \]

while the “uncorrected” version is

\[ \ln L_i \approx \ell_{i,\text{laplace}}^l = \frac{k}{2} \ln(2\pi) - \frac{k}{2} \ln n_i - \frac{1}{2} \ln \left| G(\beta, D, y_i) \right| + n_i \ell_i(\hat{b}_i) \]

where \( G(\beta, D, y_i) = \hat{Z}_i^T R_i^{-1} [d(a_i, \beta, \hat{b}_i), \xi] \hat{Z}_i + D^{-1} \). This form for \( G(\beta, D, y_i) \) was similarly specified in Section 1.3.3 except that the variance matrix \( R_i \) is included here due to our more complex model. This inclusion of \( R_i \) is an attempt to mimic the performance of other software that will not allow the specification of a nonconstant variance.
Note that if a model has constant variance, any resulting difference between the two versions of Laplace’s approximation is due to how the “uncorrected” version neglects the Hessian term, i.e. $\ell''(\hat{b}_i) = -\frac{1}{n_i}G(\beta, D, y_i) + \text{Hessian term.}$

### 3.2.2 Importance Sampling

In Section 1.3.3 a closed form expression was derived that approximates the log-likelihood using importance sampling. It was derived similarly by Pinheiro and Bates (1995). Here is the analogue for our models of the contribution to the log-likelihood for the $i$th subject:

$$L_i = \int \frac{1}{(2\pi)^{n_i/2}} \left| D^{-1} \right|^{1/2} \left| R_i^{-1}[d(a_i, \beta, b_i), \xi] \right|^{1/2} \exp \left( -\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_i), \xi] \{ y_i - f[d(a_i, \beta, b_i)] \} \right) \frac{g(b_i)}{g(b_i)} \, db_i$$

$$= \int \frac{1}{(2\pi)^{n_i/2}} \left| D^{-1} \right|^{1/2} \left| R_i^{-1}[d(a_i, \beta, b_i), \xi] \right|^{1/2} \exp \left( -\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_i), \xi] \{ y_i - f[d(a_i, \beta, b_i)] \} \right) \frac{g(b_i)}{g(b_i)} \, db_i$$

where $g(b_i)$ is the p.d.f. for $N(\hat{b}_i, H_i^{-1})$, $H_i$ is the Hessian of $-\ell_i(b_i)$ at $b_i = \hat{b}_i$. Then

$$L_i = \int \frac{1}{(2\pi)^{n_i/2}} \left| D^{-1} \right|^{1/2} \left| H^{-1} \right|^{1/2} \left| R_i^{-1}[d(a_i, \beta, b_i), \xi] \right|^{1/2} \exp \left( -\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_i), \xi] \{ y_i - f[d(a_i, \beta, b_i)] \} \right) \frac{g(b_i)}{g(b_i)} \, db_i$$

$$+ \frac{1}{2} \left( b_i - \hat{b}_i \right)^T H_i \left( b_i - \hat{b}_i \right) g(b_i) db_i$$
\[
= \frac{1}{(2\pi)^{n/2}} |D^{-1}|^{1/2} |H^{-1}|^{1/2} E \left[ |R_i^{-1}[d(a_i, \beta, b_i), \xi]|^{1/2} \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, b_i)]\}^T R_i^{-1}[d(a_i, \beta, b_i), \xi] \{y_i - f[d(a_i, \beta, b_i)]\} \right) \right]
- \frac{1}{2} b_i^T D^{-1}b_i \right]
\]
where \(b_i \sim N(\hat{b}_i, H_i^{-1})\)
\[
\approx \frac{1}{(2\pi)^{n/2}} |D^{-1}|^{1/2} |H^{-1}|^{1/2} \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} E \left[ |R_i^{-1}[d(a_i, \beta, b_{ij}^*), \xi]|^{1/2} \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, b_{ij}^*)]\}^T R_i^{-1}[d(a_i, \beta, b_{ij}^*), \xi] \{y_i - f[d(a_i, \beta, b_{ij}^*)]\} \right) \right]
- \frac{1}{2} b_{ij}^* D^{-1}b_{ij}^* + \frac{1}{2} z_j^* z_j^*
\]
where \(z_j^* \sim N(0, I_q)\), and \(b_{ij}^* = \hat{b}_i + H_i^{-1/2} z_j^*\), \(H_i^{-1/2}\) is the lower triangular, Cholesky decomposition of \(H_i^{-1}\).

The following approximation to the log-likelihood is the result.
\[
\ell_i^{\text{cimp}} = -\frac{n_i}{2} \ln(2\pi) - \frac{1}{2} \ln |D| - \frac{1}{2} \ln |H_i| + \ln \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} E \left[ |R_i^{-1}[d(a_i, \beta, b_{ij}^*), \xi]|^{1/2} \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, b_{ij}^*)]\}^T R_i^{-1}[d(a_i, \beta, b_{ij}^*), \xi] \{y_i - f[d(a_i, \beta, b_{ij}^*)]\} \right) \right]
- \frac{1}{2} b_{ij}^* D^{-1}b_{ij}^* + \frac{1}{2} z_j^* z_j^*
\]
Note that the approximate log-likelihood is denoted as \(\ell_i^{\text{cimp}}\) for corrected importance sampling because the correct Hessian is used. Here the \(b_{ij}^*\) are \(\hat{b}_i + [-n_i \ell_i''(\hat{b}_i)]^{-1/2} z_j^*\). By replacing \(H_i = -\ell_i''(\hat{b}_i)\) with \(\frac{1}{n_i} G(\beta, D, y_i)\), the uncorrected version of the contribution to the approximate log-likelihood by the \(i^{th}\) subject for importance sampling is
\[
\ln L_i \approx \ell_i^{\text{cimp}}
= -\frac{n_i}{2} \ln(2\pi) - \frac{k}{2} \ln n_i - \frac{1}{2} \ln \left( \frac{1}{n_i} G(\beta, D, y_i) \right) - \frac{1}{2} \ln |D|
\]
\[-\frac{1}{2} \ln |R_i[d(a_i, \beta, b_{i,j}^*), \xi]| \]
\[+ \ln \sum_{j=1}^{N} \left\{ \frac{1}{N} \exp \left[ -\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_{i,j}^*), \xi] \times \{ y_i - f[d(a_i, \beta, b_i)] \} - \frac{1}{2} b_{i,j}^* D^{-1} b_{i,j}^* + \frac{1}{2} z_j^T z_j \right] \right\} \]
\[= -\frac{n_i}{2} \ln(2\pi) - \frac{1}{2} \ln |G(\beta, D, y_i)| - \frac{1}{2} \ln |D| - \frac{1}{2} \ln |R_i[d(a_i, \beta, b_{i,j}^*), \xi]| \]
\[+ \ln \sum_{j=1}^{N} \left\{ \frac{1}{N} \exp \left[ -\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_{i,j}^*), \xi] \times \{ y_i - f[d(a_i, \beta, b_i)] \} - \frac{1}{2} b_{i,j}^* D^{-1} b_{i,j}^* + \frac{1}{2} z_j^T z_j \right] \right\} \]

where $z_1^*, z_2^*, ..., z_N^*$ is a random sample from $N(0, I_k)$ and $b_{i,j}^* = \hat{b}_i + [G(\beta, D, y_i)]^{-1/2} z_j^*$ and $[G(\beta, D, y_i)]^{-1/2}$ is the lower triangular Cholesky decomposition of the inverse of $G(\beta, D, y_i)$. Both of these methods, $uimp$ and $cimp$, were used in this experiment to show the effect of this approximation to $\ell_i''(\hat{b}_i)$.

The performance of importance sampling is affected by the variance of the function to be integrated. If $\int f(x) g(x) dx$ is the integral, where $x$ has p.d.f. $g(x)$, then the performance of importance sampling is affected by the size of the variance of $f(x)$. If this variance is large, then the approximation to the integral will not be as accurate for a set number of samples. To ensure a smaller variance, the integrand can be scaled by an approximation to the function $f$. The next version of importance sampling derived here shows how this scaling is done. Let
\[
\psi_i(b_i) = -\frac{1}{2} \ln |R_i[d(a_i, \beta, b_i), \xi]| \]
\[-\frac{1}{2} \{ y_i - f[d(a_i, \beta, b_i)] \}^T R_i^{-1}[d(a_i, \beta, b_i), \xi] \{ y_i - f[d(a_i, \beta, b_i)] \} \]
\[-\frac{1}{2} b_i^T D^{-1} b_i \]
\[
L_i = \int \frac{1}{(2\pi)^{(m+q)/2}} |D|^{-1/2} |R_i[d(a_i, \beta, b_i), \xi]|^{-1/2} \exp \left( -\frac{1}{2} b_i^T D^{-1} b_i \right) \\
\quad - \frac{1}{2} \{y_i - f[d(a_i, \beta, b_i)]\}^T R_i^{-1} [d(a_i, \beta, b_i), \xi] \{y_i - f[d(a_i, \beta, b_i)]\} \right) db_i \\
= \int \frac{1}{(2\pi)^{(m+q)/2}} |D|^{-1/2} \exp (\psi_i) db_i
\]

Let \( \hat{b}_i \) be the mode of \( \psi_i(b_i) \) (which is the same \( \hat{b}_i \) as in Laplace’s approximation). Then a second order Taylor series expansion of \( \psi_i(b_i) \) about \( \hat{b}_i \) is

\[
\psi_i(b_i) \approx \psi_i^*(b_i) = \psi_i(\hat{b}_i) - \frac{1}{2} (b_i - \hat{b}_i) H_i (b_i - \hat{b}_i)
\]

where \( H_i \) is the negative Hessian of \( \psi_i(b_i) \) at \( \hat{b}_i \), 

\( H_i = -\nabla^2 \psi_i(b_i) \big|_{b_i=\hat{b}_i} \). (Note that because \( \psi_i(b_i) = \text{constant} + n_i \ell_i(b_i) \) the Hessian \( H_i \) is the same for both.) The linear term is neglected since \( E(b_i - \hat{b}_i) \approx 0 \).

Thus,

\[
L_i = \int \frac{1}{(2\pi)^{(m+q)/2}} |D|^{-1/2} \exp[\psi_i(b_i)] db_i \\
= \int \frac{1}{(2\pi)^{(m+q)/2}} |D|^{-1/2} \exp[\psi_i(b_i) - \psi_i^*(b_i)] \exp[\psi_i^*(b_i)] db_i
\]
\[
\begin{align*}
\ell_i^\text{simp} & = \int \frac{1}{(2\pi)^{n_i/2}} |D|^{-1/2} \exp[\psi_i(b_i) - \psi_i^*(b_i)] |R_i[d(a_i, \beta, \hat{b}_i), \xi]|^{-1/2} |H_i|^{-1/2} \\
& \quad \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\}^T R_i^{-1} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\} \right) \\
& \quad - \frac{1}{2} \hat{b}_i^T D^{-1} \hat{b}_i \\
& = \frac{1}{(2\pi)^{n_i/2}} |D|^{-1/2} \left| R_i[d(a_i, \beta, \hat{b}_i), \xi] \right|^{-1/2} |H_i|^{-1/2} \\
& \quad \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\}^T R_i^{-1} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\} \right) \\
& \quad - \frac{1}{2} \hat{b}_i^T D^{-1} \hat{b}_i \sum_{j=1}^N \frac{1}{N} \exp[\psi_i(b_{ij}) - \psi_i^*(b_{ij})] \\
& \approx \frac{1}{(2\pi)^{n_i/2}} |D|^{-1/2} \left| R_i[d(a_i, \beta, \hat{b}_i), \xi] \right|^{-1/2} |H_i|^{-1/2} \\
& \quad \exp \left( -\frac{1}{2} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\}^T R_i^{-1} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\} \right) \\
& \quad - \frac{1}{2} \hat{b}_i^T D^{-1} \hat{b}_i \sum_{j=1}^N \frac{1}{N} \exp[\psi_i(b_{ij}^*) - \psi_i^*(b_{ij}^*)]
\end{align*}
\]

This derivation results in the approximation to the log–likelihood of the \(i\)th subject based on an alternate scaling as

\[
\ln L_i \approx \ell_i^\text{simp}
\]

\[
\approx -\frac{n_i}{2} \ln(2\pi) - \frac{1}{2} \ln |D| - \frac{1}{2} \ln |R_i[d(a_i, \beta, \hat{b}_i), \xi]| - \frac{1}{2} \ln |H_i| \\
- \frac{1}{2} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\}^T R_i^{-1} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\} \\
- \frac{1}{2} \hat{b}_i^T D^{-1} \hat{b}_i + \frac{1}{N} \sum_{j=1}^N \exp[\psi_i(b_{ij}^*) - \psi_i^*(b_{ij}^*)]
\]

where \(b_{ij}^* = \hat{b}_i + H_i^{-1/2} z_j\), and \(z_j \sim N(0, 1)\) \((H_i^{-1/2})\) is once again the lower triangular Cholesky decomposition of \(H_i^{-1}\).

This scaled version of importance sampling \((simp)\) was included along with the “corrected” version \((cimp)\), and the “uncorrected” version \((uimp)\) in this experiment.
3.2.3 Gaussian Quadrature

In Section 1.3.4 the Gaussian quadrature method derived by Pinheiro and Bates was introduced. In this experiment this method was corrected to work for models with the more complex variance structure found in our models. This correction was done similarly as was done for correcting the methods based on Laplace’s approximation and importance sampling. Instead of also using an uncorrected version of Gaussian quadrature for comparison, efforts were aimed at including the adaptive Gaussian quadrature method also derived by Pinheiro and Bates (1995). The performance of a third version of Gaussian quadrature was also investigated which is based on the same “centering” scheme as centered importance sampling described in Section 3.2.2.

The first Gaussian quadrature method results in the following approximation to the log-likelihood:

\[
\ln L_i \approx \ell_i^{qq} = -\frac{1}{2} \ln(2\pi) + \ln \left[ \sum_{j_k=1}^{N_{GQ}} \cdots \sum_{j_2=1}^{N_{GQ}} \sum_{j_1=1}^{N_{GQ}} \left| R^{-1} \left[ d(a_i, \beta, b^*_i, j_{i, j_2, \ldots, j_k}) \right] \right|^{1/2} w_{j_1} w_{j_2} \cdots w_{j_k} \right]
\]

\[
\exp \left( -\frac{1}{2} \left[ y_i - f[d(a_i, \beta, b^*_i, j_{i, j_2, \ldots, j_k})] \right]^T R_t^{-1} [d(a_i, \beta, b^*_i, j_{i, j_2, \ldots, j_k}), \xi] \times \{ y_i - f[d(a_i, \beta, b^*_i, j_{i, j_2, \ldots, j_k})] \} \right)
\]

where \( z_1^*, z_2^*, \ldots, z_k^* \) and \( w_1, w_2, \ldots, w_k \) are the abscissas and weights for the N(0,1) quadrature rule, \( b^*_i, j_{i, j_2, \ldots, j_k} = D^{1/2} z^*_i, j_{i, j_2, \ldots, j_k} = D^{1/2} (z_{i, j_1, \ldots, j_k})^T \), and \( D^{1/2} \) is the lower triangular Cholesky decomposition of \( D \). This method is the analogue of that described by Pinheiro and Bates (1995), with the adjustment for a nonconstant variance.
The second Gaussian quadrature method is Pinheiro and Bates’ adaptive Gaussian quadrature:

$$\ln L_i \approx \ell_i^{ga}\n$$

$$\approx \frac{-n_i}{2} \ln(2\pi) - \frac{k}{2} \ln n_i - \frac{1}{2} \ln | - \ell_i^{g}(\hat{b}_i) | - \frac{1}{2} \ln |D|$$

$$+ \ln \sum_{j=1}^{N} \{ \left| R_i [d(a_i, \beta, b_{ij}^{*}), \xi] \right|^{-1/2} \right.$$

$$\exp \left[ -\frac{1}{2} \{ y_i - f [d(a_i, \beta, b_{ij}^{*})] \}^T R_i^{-1} [d(a_i, \beta, b_{ij}^{*}), \xi] \{ y_i - f [d(a_i, \beta, b_{ij}^{*})] \} \right.$$

$$\left. - \frac{1}{2} b_{ij}^{*T} D^{-1} b_{ij}^{*} + \frac{1}{2} z_{i}^{*T} z_{i}^{*} \right] w_1 w_2 \cdots w_k \}$$

where $b_{ij}^{*}$ is now $\hat{b}_i + [- n_i \ell_i^{g}(\hat{b}_i)]^{-1/2} z_{i}^{*}$ this approximate log–likelihood is very similar to the log–likelihood for the “corrected” importance sampling method. There is one major difference with adaptive Gaussian quadrature: the abscissas are determined by a quadrature rule for adaptive Gaussian quadrature while with importance sampling, they are randomly chosen.

Lastly, a final version of Gaussian quadrature was used which scales the integrand similarly to the scaled importance sampling method (simp). Consider,

$$L_i = \frac{1}{(2\pi)^{n_i/2}} |D|^{-1/2} \left| R_i [d(a_i, \beta, \hat{b}_i), \xi] \right|^{-1/2} \left| H_i \right|^{-1/2}$$

$$\exp \left( -\frac{1}{2} \{ y_i - f [d(a_i, \beta, b_{ij}^{*})] \}^T R_i^{-1} [d(a_i, \beta, b_{ij}^{*}), \xi] \{ y_i - f [d(a_i, \beta, \hat{b}_i)] \} \right.$$  

$$- \frac{1}{2} b_{ij}^{*T} D^{-1} b_{ij}^{*} \right) \exp[\psi_i(b_k) - \psi_i^{*}(b_k)]$$

$$\approx \frac{1}{(2\pi)^{n_i/2}} |D|^{-1/2} \left| R_i [d(a_i, \beta, \hat{b}_i), \xi] \right|^{-1/2} \left| H_i \right|^{-1/2}$$

$$\exp \left( -\frac{1}{2} \{ y_i - f [d(a_i, \beta, \hat{b}_i)] \}^T R_i^{-1} [d(a_i, \beta, \hat{b}_i), \xi] \{ y_i - f [d(a_i, \beta, \hat{b}_i)] \} \right.$$  

$$- \frac{1}{2} b_{ij}^{*T} D^{-1} b_{ij}^{*} \right)$$
\[-\frac{1}{2} b_i^T D^{-1} b_i \sum_{j_q=1}^{N} \cdots \sum_{j_1=1}^{N} \exp[\psi_i(b^*_{j_1j_2\ldots j_k}) - \psi_i^*(b^*_{j_1j_2\ldots j_k})]w_{j_1} \cdots w_{j_q}\]

where \( b^*_{j_1j_2\ldots j_k} = \hat{b}_i + H_i^{-1/2} z_{j_1j_2\ldots j_k} \), and \( z_{j_1j_2\ldots j_k} \) and \( w_{j_1} \cdots w_{j_q} \) are the appropriate abscissas and weights based on the \( N(0,1) \) Gaussian quadrature rule.

This results in the approximation:

\[
\ln L_i \approx \ell_i^{ga}
\]

\[
\approx -\frac{n_i}{2} \ln(2\pi) - \frac{1}{2} \ln |D| - \frac{1}{2} \ln |R_i|d(a_i, \beta, \hat{b}_i, \xi)| - \frac{1}{2} \ln |H_i|
\]

\[
- \frac{1}{2} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\} R_i^{-1} \{y_i - f[d(a_i, \beta, \hat{b}_i)]\}
\]

\[
- \frac{1}{2} b_i^T D^{-1} b_i + \sum_{j_q=1}^{N} \cdots \sum_{j_1=1}^{N} \exp[\psi_i(b^*_{ij}) - \psi_i^*(b^*_{ij})]w_{j_1} \cdots w_{j_q}
\]

### 3.3 Models

Data were simulated from three models chosen to satisfy the following general form.

\[
y_{ij} = f(x_{ij}, \beta_i) + e_{ij}
\]

\[
(e_i | \beta_i) \sim N[0, R_i(\beta_i, \xi)]
\]

\[
\beta_i = d(a_i, \beta, b_i)
\]

\[
b_i \sim N(0, D)
\]

Each data set contains information for at least 35 subjects. Two replicate data sets were simulated for Model 1 and Model 3. Model 2 was used to simulate data for
4 data sets, 2 at each of 2 different parameter settings. Overall, there were 8 sets of data simulated.

**Model 1.** The following is a one-compartmental model for a drug given as an intravenous bolus. It is similar to the model used in the simulation study in Chapter 2.

\[ y_{ij} = f(x_{ij}, \beta_i) + e_{ij} = \frac{100}{V_i} \exp \left( \frac{-C_l t_{ij}}{V_i} \right) + e_{ij} \]

\[ (e_{i} | \beta_i) \sim N(0, \sigma^2 I_{n_i}) \]

\[ C_l = \exp \left( \beta_1 + \frac{\beta_2}{100} a_i + b_{1i} \right) \]
\[ V_i = \exp \left( \beta_3 + b_{2i} \right) \]
\[ b_i \sim N(0, D) \]

Eight observations were simulated for each of the 35 individuals. The values for \( t_{ij} \) were 0.25, 0.50, 1.00, 2.00, 4.00, 8.00, 12.00, and 24.00. The covariate \( a_i \) for each individual was sampled independently from \( N(70, 20) \). The values for the fixed parameters were \( \beta = [\beta_1, \beta_2, \beta_3]^T = [\ln(1.25), 0.6, \ln(10)]^T = [0.223, 0.6, 2.3]^T. \) The random effects were sampled from \( N(0, D) \) where \( D \) was chosen to have C.V. = .3 and \( \rho = 0.6, \)

\[ D = \begin{bmatrix} 0.086 & 0.052 \\ 0.052 & 0.086 \end{bmatrix}. \]

The errors were sampled to have constant variance of \( \sigma^2 = (0.15)^2 = 0.0225. \)
For this model, the differences between the “corrected” and “uncorrected” methods should be slight due to the constant variance. Any difference observed will be due to approximating the Hessian with $G(\beta, D, y_i)$. The numerical derivatives used for the “corrected” method should not be an issue since they should be very accurate.

**Model 2.** This model is the same as Model 1 except that the errors have the following, slightly more complicated, yet common heteroscedastic structure:

$$(e_i | \beta_i) \sim N[0, \sigma^2 f^2(x_{ij}, \beta_i)]$$

For this model, since the error variance is not constant, the difference between the “corrected” and “uncorrected” methods should become more pronounced.

Two different data sets were sampled from this model. The first had all the same parameter settings as Model 1 including $\sigma^2 = 0.0225$ while the second had $\sigma^2 = (0.3)^2 = 0.09$. The reason for simulating data at two levels of $\sigma$ is that the Laplacian method is known to work well for “small” $\sigma$ while being not as successful for larger values (noisier data).

**Model 3.** The third model is based on the model used in the dose escalation study of argatroban, an anticoagulant. In this study, subjects received a four-hour intravenous infusion of one of several doses of argatroban. The objectives of this study were to characterize the pharmacokinetics of argatroban after intravenous infusion at different doses and to clarify the relationship between argatroban concentration and the pharmacodynamic endpoint aPTT (activated partial thromboplastin time). See Davidian and Giltinan (1995) for more details.
The concentration-time profile is modeled as

\[ y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \]

\[ = \frac{D_i}{\text{Cl}_i} \left[ \exp \left( -\frac{\text{Cl}_i}{V_i} t^* \right) - \exp \left( -\frac{\text{Cl}_i}{V_i} t_{ij} \right) \right] + e_{ij} \]

\[ t^* = \begin{cases} 0 & \text{for } t \leq t_{inf} \\ t - t_{inf} & \text{for } t > t_{inf} \end{cases} \]

\( (e|\beta_i) \sim (0, R_i (\beta_i, \xi)) \).

Here, \( y_{ij} \) is the \( j \)th concentration measurement taken at dose \( D_i \) and time \( t_{ij} \) for the \( i \)th subject, \( j = 1, \cdots, n_i \). \( \text{Cl}_i \) is the clearance and \( V_i \) is the volume for the \( i \)th subject.

Similarly to Model 2, this model also has nonconstant variance specified as

\[ R_i (\beta_i, \xi) = \sigma^2 \text{diag} \left[ f^2 (x_{i1}, \beta_i), \cdots, f^2 (x_{in_i}, \beta_i) \right], \]

\[ x_{ij} = [D_i, t_{ij}]^T, \]

\[ \xi = [\sigma, \theta]^T. \]

Once again, the “corrected” methods would be expected to outperform the “uncorrected” methods.

Further, the subject to subject variation is modeled

\[ \beta_i = [\beta_{i1}, \beta_{i2}]^T = [\ln(\text{Cl}_i), \ln(V_i)]^T \]

\[ \beta_i = \beta + b_i, \]

\[ b_i \sim (0, D), \]
\[ \beta = [\beta_1, \beta_2]^T \quad \text{and} \quad b_i = [b_{1i}, b_{2i}]^T. \]

Fourteen observations were simulated for each of the 36 individuals using Model 3. The values for the dose \( D_i \) were chosen to be 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. Each dose level was given to one of 4 subjects per replicate data set (4 subjects \( \times \) 9 dose levels = 36 subjects).

The fixed parameters were set at \( \beta = [\beta_1, \beta_2]^T = [−5.5, −1.9]^T, \)

\[
D = \begin{bmatrix} 0.14 & 0.01 \\ 0.01 & 0.01 \end{bmatrix},
\]

\( \sigma^2 = (0.1)^2 = 0.01 \), and \( \theta = 1 \). The parameter \( \theta \) was assumed to be known and was not included as a dimension in the parameter space. The values for \( t_{ij} \) were 30, 60, 90, 115, 160, 200, 240, 245, 250, 260, 275, 295, 320, and 360 and \( t_{inf} \) was 240.

### 3.4 Sampling the Parameter Space

The goal of the experiment was to compare the performance of the approximate methods in evaluating the log-likelihood at different points in the parameter space. In practice, a minimization procedure would be used to minimize the value of minus two times the log-likelihood to find the maximum likelihood estimates of the parameters. Then, approximations of the log-likelihood would be found at different points in the parameter space. These points need to be near the maximum likelihood solution and the approximations of the log-likelihood need to be “close” to the true values. To
test the performance of these approximate methods, points were sampled near the
true values of the parameters where the data were simulated. Two elliptical shells
about the truth were constructed: an inner shell and an outer shell. The outer shell
was twice the distance from the center than the inner shell. Twenty points on each
shell were then sampled; the points on the outer shell were along the same vector out
from the center as the points on the inner shell.

The choice for the distance from the center was made using standard errors from
initial estimates. These estimates were found using the new SAS procedure \texttt{NLMIXED}.
Let the true vector of parameters that were used to simulate the data be $\phi$, the
estimate from \texttt{NLMIXED} be $\hat{\phi}$, and the sample covariance matrix of the initial estimates
provided by \texttt{NLMIXED} be $\hat{\Omega}$. The deviance was used to create a modified covariance
matrix,

\[ \hat{\Omega}^* = \hat{\Omega} + (\phi - \hat{\phi}) (\phi - \hat{\phi})^T. \]

A random sample $z_1, z_2, ..., z_{20}$ was taken from $\mathcal{N}(0, I_p)$ ($p$ is the length of the vector
$\phi$). The sample from the inner elliptical shell was

\[ \phi^1_j = \phi + \hat{\Omega}^{1/2} z_j, \ j = 1, 2, ..., 20 \]

and the outer

\[ \phi^2_j = \phi + 2 \hat{\Omega}^{1/2} z_j, \ j = 1, 2, ..., 20 \]

where $\hat{\Omega}^{1/2}$ is the lower triangular Cholesky factorization of $\hat{\Omega}$.

In this manner, points were sampled either 1 or 2 standard deviations away from
the truth. The sampled values for the components of the matrix \( D \) occasionally result in a \( D \) matrix that was negative definite. When this occurred, the sampled point was discarded and the methods were only applied to the remaining points in the sample.

### 3.5 Results

Each method was used to calculate an approximate log-likelihood value for each of the two replicate data sets for each model. The Gaussian quadrature and importance sampling methods were used with a different number of abscissas. By including different numbers of abscissas, sampling intensities for accurately approximating the log-likelihood were investigated. The choices for the number of abscissas for the Gaussian Quadrature methods were 3, 7, and 10. Because the integration was performed over two dimensions \((k = 2 \text{ for these } 3 \text{ models})\) the resulting number of evaluations of the integrand were 9, 49, and 100. To match the intensity of the quadrature methods, the number of importance samples were chosen to also be 9, 49, and 100.

Unlike the other methods, the intensity of the methods based on Laplace’s approximation could not be modified. The goal was to discover how intensive the other methods needed to be to outperform Laplace’s method. If the intensity needed was extremely high, then the lower intensive Laplace’s method might be more advantageous due to its lower computational demands.

Even though the intensity of Laplace’s method cannot be increased, it still has a significant computational complexity. Each evaluation of an individual’s contribution
to the log–likelihood requires finding \( \hat{b}_i \), the value that maximizes \( \ell_i(b_i) \). This requires using an optimizing subroutine that is possibly not able to converge to a solution for \( \hat{b}_i \). But Laplace’s method is considered computationally “cheaper” than the other methods that center about \( \hat{b}_i \) since they also require searching for this value.

As the intensity of the Laplacian methods cannot be modified, and there are no evaluations at random values generated within the algorithm, these methods are fixed; they will always result in the same approximation for the log–likelihood. The quadrature methods are also fixed for any given intensity level, i.e. number of abscissas. But, the methods based on importance sampling are random. If a second random sample were taken, one would expect the result for the approximation to be slightly different. The choice of centering and scaling of the method will determine how large the variance of the sampling of the log–likelihood will be. In practice, these methods are linked to maximizing/minimizing subroutines to find the maximum of the log–likelihood (or minimum of minus two times the log–likelihood). This randomness of the importance sampling methods can wreak havoc with the max/min subroutines if care is not taken to ensure that the value of the log–likelihood remains fixed once it has been approximated. This safeguard can be managed by taking one sample from the importance distribution and reusing this sample at every point in the parameter space that the likelihood is to be evaluated. In this experiment, this safeguard was employed. In addition, the same random vector was used across the different random methods. This safeguard was employed to achieve a more “fair” comparison of the
methods.

The approximations using the scaled and centered Gaussian quadrature method with 100 abscissas were chosen to represent the “true” log-likelihood since it was expected that this method was to perform the best. The other methods were compared to this marker using relative error and absolute relative error for each approximation at each point. The relative error was calculated as

$$\frac{\text{(approximate log-likelihood) } - \text{ truth}}{\text{truth}}.$$  

The resulting data of absolute relative errors were analyzed as a two factor (method of approximation and radius), factorial experiment. The results for each model were analyzed separately as it was expected that there would be an interaction between model, the methods, and radii (distance of either 1 or 2 standard errors from the truth). Plots were constructed showing the mean relative error for each model, method, and radii combination. Mean relative error was chosen for the plots since the absolute relative error would not show whether a method was more often over or under approximating the log-likelihood. Since it is of interest which method is closer to the truth and not in which direction it misses, the absolute relative error was chosen as the response variable for the analysis of variance.

For convenience, the methods have been denoted with abbreviations as listed in Table 3.1.
Table 3.1: Abbreviations for the Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Version</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace’s Approximation</td>
<td>Uncorrected</td>
<td>ulaplace</td>
</tr>
<tr>
<td></td>
<td>Corrected</td>
<td>claplace</td>
</tr>
<tr>
<td>Importance Sampling</td>
<td>Uncorrected</td>
<td>uimp</td>
</tr>
<tr>
<td></td>
<td>Corrected</td>
<td>cimp</td>
</tr>
<tr>
<td></td>
<td>Centered and Scaled</td>
<td>simp</td>
</tr>
<tr>
<td>Gaussian Quadrature</td>
<td>Uncentered</td>
<td>gq</td>
</tr>
<tr>
<td></td>
<td>Adaptive</td>
<td>agg</td>
</tr>
<tr>
<td></td>
<td>Centered and Scaled</td>
<td>sgq</td>
</tr>
</tbody>
</table>

3.5.1 Analysis for Model 1

The mean values of relative error for Model 1 at each combination of method, radii, and replicate are given in Table 3.2. Plots of the mean relative errors are given in Figures 3.1–3.5. The plots are divided so that means for the Laplacian methods, importance sampling methods, and Gaussian quadrature methods are all shown separately. In each, the smaller symbols represent mean relative errors for points with radius one (one standard error from the truth) while larger symbols represent those with radius two. The $y$-axis is the relative error in each plot while the meaning of the $x$-axis is different among the plots. For importance sampling and Gaussian quadrature, the $x$-axis represents the number of abscissas sampled, i.e. the intensity of the method. For the Laplacian methods, there is no difference in intensity, so a value of one for intensity is set for these methods. In plots for the Laplacian methods the $x$-axis represents the replicate, one or two.

The radius by method interaction is significant when all methods are included in
the analysis. Therefore, the methods were initially culled to only \textit{ulaplace}, \textit{claplace}, \textit{simp}, and \textit{sgq}.

The means for the Laplacian methods are displayed in a scatter plot in Figure 3.1. The separation between the “corrected” and “uncorrected” versions of this method, with the “corrected” version so much closer to zero, shows how ignoring the Hessian term can affect the approximation. Since Model 1 has constant variance, we know that this separation is only caused by the removal of the Hessian term. When the scale of the axis is taken into consideration, this separation does not appear large but was still found to be statistically significant with a p-value less than 0.0001.

The means for all three importance sampling methods are given in Figure 3.2. The means for \textit{simp} are very close to zero showing good, stable approximations, while the other two methods appear less reliable. A large difference is not shown in the plot between the \textit{uimp} and \textit{cimp} methods. This is due to the constant variance of the

\begin{table}
\centering
\caption{Model 1, mean relative error.}
\begin{tabular}{cccccc}
\hline
r & rep & n & \text{Laplace’s Approx.} & \text{Importance Sampling} & \text{Gaussian Quadrature} \\
& & & \text{ulaplace} & \text{claplace} & \text{uimp} & \text{cimp} & \text{simp} & \text{gq} & \text{agq} & \text{sgq} \\
\hline
1 & 1 & 1 & 0.00011 & 0.0000002 & 0.09 & 0.09 & -0.000026 & 0.87 & 0.005 & 0.000000244 \\
1 & 1 & 9 & 0.09 & 0.08 & -0.000018 & 0.23 & 0.009 & -0.000000001 \\
1 & 1 & 100 & 0.09 & 0.09 & 0.000006 & 0.17 & 0.013 \\
2 & 1 & 1 & 0.00011 & 0.0000002 & 0.09 & 0.09 & -0.000028 & 1.05 & 0.005 & 0.000000278 \\
2 & 1 & 9 & 0.08 & 0.08 & -0.000019 & 0.31 & 0.009 & -0.000000001 \\
2 & 1 & 100 & 0.08 & 0.08 & 0.000006 & 0.21 & 0.013 \\
1 & 2 & 1 & 0.00003 & 0.000003 & 0.16 & 0.16 & -0.000051 & 2.11 & 0.008 & 0.000000527 \\
1 & 2 & 9 & 0.15 & 0.15 & -0.000035 & 0.43 & 0.015 & -0.000000001 \\
1 & 2 & 100 & 0.15 & 0.15 & 0.000011 & 0.28 & 0.023 \\
2 & 2 & 1 & 0.00003 & 0.000003 & 0.15 & 0.15 & -0.000048 & 2.51 & 0.007 & 0.000000512 \\
2 & 2 & 9 & 0.14 & 0.14 & -0.000033 & 0.72 & 0.014 & -0.000000001 \\
2 & 2 & 100 & 0.14 & 0.14 & 0.000010 & 0.46 & 0.022 \\
\hline
\end{tabular}
\end{table}
error term.

The means for all three Gaussian quadrature methods are in Figure 3.3. From the values in the plots, it is easily seen that $gq$ is the worst method for Model 1 (and continues to be so for the other models). Since $gq$ is separated quite noticeably from the others, a second plot was made without this method. This plot is in Figure 3.4 and it can now be seen how $agq$ and $sgq$ are different. A trend can also be seen that shows how $agq$ actually gets further from the truth when the intensity of the method is increased.

The radius by method interaction is significant when all methods are included in the analysis. Even though $ulaplace$ does fairly well for some replicates, there is no reason to keep considering it; this method has the same intensity as the more correct method, $claplace$, and does not cost any more computing time. Therefore, all but the best methods ($claplace$, $simp$, and $sgq$) were removed for a final analysis. A plot showing the means for just these methods is given in Figure 3.5. In this figure, the intensity for the $claplace$ method was arbitrarily given the value of zero. The means for $sgq$ can be seen to remain constant as the intensity increases. This validates the choice of $sgq$ with 10 (100) abscissas to represent the true log-likelihood since it shows convergence so quickly.

The Analysis of Variance for Model 1 is given in Table 3.3. There was no significant interaction between method and radius and the main effect of radius was also not significant. There were differences among the methods; the p-value for this
Table 3.3: Analysis of Variance for Model 1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13</td>
<td>1.09034E-7</td>
<td>8.38728E-9</td>
<td>106.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>532</td>
<td>4.18324E-8</td>
<td>7.86324E-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>545</td>
<td>1.50867E-7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Square</td>
<td>0</td>
<td>722.720</td>
<td>82.35105</td>
<td>8.8675E-6</td>
<td>0.00001</td>
</tr>
<tr>
<td>Coef Var</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>6</td>
<td>1.09034E-7</td>
<td>8.38728E-9</td>
<td>106.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>radius</td>
<td>1</td>
<td>2.24291E-11</td>
<td>2.24291E-11</td>
<td>0.29</td>
<td>0.5935</td>
</tr>
<tr>
<td>radius*trt</td>
<td>6</td>
<td>4.36021E-11</td>
<td>7.26709E-12</td>
<td>0.00</td>
<td>0.9071</td>
</tr>
<tr>
<td>sgq linear</td>
<td>1</td>
<td>6.36650E-12</td>
<td>6.36650E-12</td>
<td>0.08</td>
<td>0.7761</td>
</tr>
<tr>
<td>sgq quadratic</td>
<td>1</td>
<td>1.40092E-12</td>
<td>1.40092E-12</td>
<td>0.02</td>
<td>0.8906</td>
</tr>
<tr>
<td>simp linear</td>
<td>1</td>
<td>3.35566E-8</td>
<td>3.35566E-8</td>
<td>426.75</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>simp quadratic</td>
<td>1</td>
<td>1.39683E-9</td>
<td>1.39683E-9</td>
<td>17.76</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp10</td>
<td>1</td>
<td>1.31808E-8</td>
<td>1.31808E-8</td>
<td>16.76</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp7</td>
<td>1</td>
<td>2.20050E-8</td>
<td>2.20050E-8</td>
<td>279.85</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp3</td>
<td>1</td>
<td>4.92584E-8</td>
<td>4.92584E-8</td>
<td>626.44</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sgq10</td>
<td>1</td>
<td>2.35061E-10</td>
<td>2.35061E-10</td>
<td>2.99</td>
<td>0.0844</td>
</tr>
<tr>
<td>claplace vs sgq7</td>
<td>1</td>
<td>2.34931E-10</td>
<td>2.34931E-10</td>
<td>2.99</td>
<td>0.0845</td>
</tr>
<tr>
<td>claplace vs sgq3</td>
<td>1</td>
<td>1.66790E-10</td>
<td>1.66790E-10</td>
<td>2.12</td>
<td>0.1459</td>
</tr>
<tr>
<td>sgq7 vs simp7</td>
<td>1</td>
<td>2.67874E-8</td>
<td>2.67874E-8</td>
<td>340.67</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgq10 vs simp10</td>
<td>1</td>
<td>2.66639E-8</td>
<td>2.66639E-8</td>
<td>33.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgq3 vs simp3</td>
<td>1</td>
<td>5.51578E-7</td>
<td>5.51578E-7</td>
<td>701.46</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

comparison was less than 0.0001.

There were no linear or quadratic effects for sgq. This shows that for this model, sgq does a good job of estimating the log–likelihood even with only 3 abscissas.

On the other hand, there were significant linear and quadratic effects for simp. This shows, along with the plot, that the accuracy of simp does improve as the intensity increases. But there is still a significant difference between simp10 and the truth (sgq10).

Unlike simp for this model, claplace does an excellent job of approximating the log–likelihood. There is no significant difference between claplace and sgq10.
3.5.2 Analysis for Model 2

The mean values of relative error for Model 2 are given in Table 3.4 (for $\sigma = 0.15$) and Table 3.6 (for $\sigma = 0.3$) and plotted in Figures 3.6–3.13. The data for each level of the parameter $\sigma$ were analyzed separately.

<table>
<thead>
<tr>
<th>r</th>
<th>rep</th>
<th>n</th>
<th>Laplace’s Approx.</th>
<th>Importance Sampling</th>
<th>Gaussian Quadrature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ulaplace</td>
<td>claplace</td>
<td>gq</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-0.004</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>49</td>
<td>1.02 1.01</td>
<td>-0.004 -0.004</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
<td>0.85 0.85</td>
<td>0.002 0.002</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-0.003 0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9</td>
<td>0.91 0.90</td>
<td>-0.004 0.001</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>49</td>
<td>0.76 0.77</td>
<td>0.002 0.002</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100</td>
<td>0.83 0.83</td>
<td>0.002 0.002</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-0.006 0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9</td>
<td>0.92 0.91</td>
<td>-0.003 0.001</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>49</td>
<td>0.77 0.77</td>
<td>0.002 0.002</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100</td>
<td>0.83 0.83</td>
<td>0.002 0.002</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-0.006 0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>9</td>
<td>0.75 0.74</td>
<td>-0.003 0.001</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>49</td>
<td>0.61 0.61</td>
<td>0.003 0.003</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>100</td>
<td>0.67 0.67</td>
<td>0.002 0.002</td>
<td></td>
</tr>
</tbody>
</table>

Model 2 with $\sigma = 0.15$.

Figure 3.8 shows a plot of all three Gaussian quadrature methods. As before, $c_{gq}$ does not perform nearly as well as the other two but the scale of the plot is adequate to show a trend of how $a_{gq}$ actually performs worse as the intensity is increased. Figure 3.7 shows how $c_{imp}$ and $u_{imp}$ are far inferior to $s_{imp}$. The scale shows that these methods even performed less adequately than $gq$. In Figure 3.6 $u_{laplace}$ is shown to once again fare well against $c_{laplace}$. When the analysis of variance was performed on $u_{laplace}$, $c_{laplace}$, $s_{imp}$, and $s_{gq}$, a significant difference was still detected between
ulaplace and claplace. The p-value for this contrast was less than 0.0001.

Once again, the final analysis only considered claplace, simp, and sgq to concentrate only on the most viable methods (see Table 3.5). The radius by method interaction was not found to be significant, nor was the main effect of radius, while differences were found between the methods.

Both linear and quadratic effects were significant for sgq. This significance shows that an increase in intensity decreases the absolute relative error at least at a quadratic rate. Another comparison showed that sgq7 was not significantly different than sgq10, showing that 7 abscissas were sufficient for approximating the log-likelihood with sgq.

A linear effect was detected among the intensity levels of simp but not a quadratic effect. This shows that performance of simp increases linearly as the intensity is increased. Note that the plot appears to show a quadratic effect because the plots are of relative error while the analysis was performed on absolute relative error.

There were significant differences between claplace and sgq at all levels of intensity of sgq, showing that even sgq with only 3 abscissas is better than claplace for this model and choice of σ. In comparing claplace with simp it was found that claplace was better than simp3, not significantly different than simp7, and significantly worse than simp10.

**Model 2 with σ =0.3.**

When σ is increased to 0.3, the relative error actually decreased for gg and agq (shown in Figure 3.8), as well as for importance sampling (shown in Figure 3.7), while
Table 3.5: Analysis of Variance for Model 2 with \( \sigma = 0.15 \)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13</td>
<td>0.000577833</td>
<td>0.00004411</td>
<td>81.69</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>413</td>
<td>0.00022452</td>
<td>0.0000054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>426</td>
<td>0.00080185</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coef Var</th>
<th>Root MSE</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.720000</td>
<td>43.13905</td>
<td>0.000737</td>
<td>0.001709</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>6</td>
<td>0.000377112</td>
<td>0.00009519</td>
<td>175.10</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>radius</td>
<td>1</td>
<td>0.00000017</td>
<td>0.00000017</td>
<td>0.31</td>
<td>0.5866</td>
</tr>
<tr>
<td>radius*trt</td>
<td>6</td>
<td>0.000000174</td>
<td>0.00000029</td>
<td>0.53</td>
<td>0.7818</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>sqg linear</td>
<td>1</td>
<td>0.00011572</td>
<td>0.00011572</td>
<td>212.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sqg quadratic</td>
<td>1</td>
<td>0.00002406</td>
<td>0.00002406</td>
<td>44.26</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>simp linear</td>
<td>1</td>
<td>0.00006182</td>
<td>0.00006182</td>
<td>112.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>simp quadratic</td>
<td>1</td>
<td>0.00001115</td>
<td>0.00001115</td>
<td>2.12</td>
<td>0.1459</td>
</tr>
<tr>
<td>claplace vs simp10</td>
<td>1</td>
<td>0.00000814</td>
<td>0.00000814</td>
<td>14.97</td>
<td>0.0001</td>
</tr>
<tr>
<td>claplace vs simp7</td>
<td>1</td>
<td>0.00000023</td>
<td>0.00000023</td>
<td>0.42</td>
<td>0.5160</td>
</tr>
<tr>
<td>claplace vs simp3</td>
<td>1</td>
<td>0.00002365</td>
<td>0.00002365</td>
<td>43.50</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sqg10</td>
<td>1</td>
<td>0.00018162</td>
<td>0.00018162</td>
<td>334.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sqg7</td>
<td>1</td>
<td>0.00017735</td>
<td>0.00017735</td>
<td>326.24</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sqg3</td>
<td>1</td>
<td>0.00009988</td>
<td>0.00009988</td>
<td>18.35</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sqg7 vs simp7</td>
<td>1</td>
<td>0.00016481</td>
<td>0.00016481</td>
<td>306.17</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sqg7 vs simp10</td>
<td>1</td>
<td>0.00011286</td>
<td>0.00011286</td>
<td>207.61</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sqg3 vs simp3</td>
<td>1</td>
<td>0.00006435</td>
<td>0.00006435</td>
<td>118.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sqg7 vs sqg10</td>
<td>1</td>
<td>0.00000003</td>
<td>0.00000003</td>
<td>0.05</td>
<td>0.8291</td>
</tr>
</tbody>
</table>
it remained about the same for the Laplacian methods (see Figure 3.6). In spite of this change, \( gq \) and \( agq \) were still the worst performers followed by \( uimp \) and \( cimp \). After removing these methods from the analysis, it was seen that \( claplace \) performed significantly better than \( ulaplace \).

Table 3.6: Model 2 (\( \sigma = 0.3 \)), mean relative error.

<table>
<thead>
<tr>
<th>r</th>
<th>rep</th>
<th>n</th>
<th>Laplace’s Approx.</th>
<th>Importance Sampling</th>
<th>Gaussian Quadrature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( ulaplace )</td>
<td>( claplace )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-0.003</td>
<td>0.003</td>
<td>0.18</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>49</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
<td>-0.005</td>
<td>0.003</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-0.005</td>
<td>0.004</td>
<td>0.18</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>49</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100</td>
<td>-0.006</td>
<td>0.003</td>
<td>0.17</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>49</td>
<td></td>
<td></td>
<td>0.13</td>
</tr>
</tbody>
</table>

The final analysis of variance of the absolute relative error of \( claplace \), \( simp \), and \( sgq \) can be found in Table 3.9. As before, there was no significant interaction between radius and method, and the main effect of radius was not significant, while differences were found between the methods.

Linear and quadratic effects in \( simp \) and \( sgq \) were detected. \( sgq \) improves at least at a quadratic rate as the intensity is increased, and does a significantly better job than \( simp \) at each intensity level. \( sgq7 \) is not significantly different than \( sgq10 \) showing that 7 is a high enough intensity for sampling abscissas for the \( sgq \) method.

The variation in the \( simp \) method is apparently responsible for \( simp3 \) resulting in
Table 3.7: Analysis of Variance for Model 2 with $\sigma = 0.30$

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13</td>
<td>0.0080366</td>
<td>0.00000618</td>
<td>87.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>441</td>
<td>0.0031053</td>
<td>0.0000070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>454</td>
<td>0.00111419</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.721295

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Square</td>
<td>6</td>
<td>0.00077018</td>
<td>0.00012836</td>
<td>182.30</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Coef Var</td>
<td>1</td>
<td>0.00000000</td>
<td>0.00000000</td>
<td>0.00</td>
<td>0.9467</td>
</tr>
<tr>
<td>Root MSE</td>
<td>1</td>
<td>0.000839</td>
<td>0.000037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1</td>
<td>0.00000000</td>
<td>0.00000000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Contrast

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>sgq linear</td>
<td>1</td>
<td>0.00029180</td>
<td>0.00029180</td>
<td>414.40</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>sgq quadratic</td>
<td>1</td>
<td>0.00005317</td>
<td>0.00005317</td>
<td>75.51</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>simp linear</td>
<td>1</td>
<td>0.0000652</td>
<td>0.0000652</td>
<td>9.25</td>
<td>0.0025</td>
<td></td>
</tr>
<tr>
<td>simp quadratic</td>
<td>1</td>
<td>0.00005261</td>
<td>0.00005261</td>
<td>74.22</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>claplace vs simp10</td>
<td>1</td>
<td>0.00006427</td>
<td>0.00006427</td>
<td>91.27</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>claplace vs simp7</td>
<td>1</td>
<td>0.00000014</td>
<td>0.00000014</td>
<td>0.19</td>
<td>0.6613</td>
<td></td>
</tr>
<tr>
<td>claplace vs simp3</td>
<td>1</td>
<td>0.00002378</td>
<td>0.00002378</td>
<td>33.77</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>claplace vs sgq10</td>
<td>1</td>
<td>0.00035856</td>
<td>0.00035856</td>
<td>509.20</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>claplace vs sgq7</td>
<td>1</td>
<td>0.00033212</td>
<td>0.00033212</td>
<td>472.09</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>claplace vs sgq3</td>
<td>1</td>
<td>0.0000650</td>
<td>0.0000650</td>
<td>8.95</td>
<td>0.0029</td>
<td></td>
</tr>
<tr>
<td>sgq7 vs simp7</td>
<td>1</td>
<td>0.00031914</td>
<td>0.00031914</td>
<td>453.23</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>sgq10 vs simp10</td>
<td>1</td>
<td>0.00011922</td>
<td>0.00011922</td>
<td>160.32</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>sgq3 vs simp3</td>
<td>1</td>
<td>0.0000560</td>
<td>0.0000560</td>
<td>7.95</td>
<td>0.0050</td>
<td></td>
</tr>
<tr>
<td>sgq7 vs sgq10</td>
<td>1</td>
<td>0.0000049</td>
<td>0.0000049</td>
<td>0.70</td>
<td>0.4025</td>
<td></td>
</tr>
</tbody>
</table>
a smaller absolute relative error than \textit{simp7}. It is also responsible for how \textit{simp} performs significantly better than \textit{claplace} with 3 abscissas, worse with 7, and then again significantly better with 10. It appears that \textit{claplace} does not fare any worse by the increase in \( \sigma \) than either \textit{simp} or \textit{sgq}. \textit{claplace} is not significantly different than \textit{sgq3} but is not as good as \textit{sgq7}. \textit{simp} with 10 abscissas is still not good enough to rank with \textit{sgq7}.

### 3.5.3 Analysis for Model 3

The mean values of relative error for each method, radii, and replicate for Model 3 are given in Table 3.2. The means are also shown in Figure 3.14 for the Laplacian methods, Figures 3.16 and 3.17 for the Gaussian quadrature methods, and Figure 3.15 for the importance sampling methods. A final plot is shown again for just the best methods, \textit{claplace}, \textit{simp}, and \textit{sgq}.

![Table 3.8: Model 3, mean relative error.](image)
Table 3.9: Analysis of Variance for Model 3

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13</td>
<td>4.332065E-6</td>
<td>3.332065E-7</td>
<td>237.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>252</td>
<td>3.539075E-7</td>
<td>1.404077E-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>265</td>
<td>4.687575E-6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coef Var</th>
<th>Root MSE</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.921500</td>
<td>35.21961</td>
<td>0.000037</td>
<td>0.000106</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>6</td>
<td>4.303899E-6</td>
<td>7.131663E-7</td>
<td>5.1076</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>radius</td>
<td>1</td>
<td>1.689379E-8</td>
<td>1.689379E-8</td>
<td>12.03</td>
<td>0.0006</td>
</tr>
<tr>
<td>radius*trt</td>
<td>6</td>
<td>1.287223E-8</td>
<td>2.145500E-9</td>
<td>1.5392</td>
<td>0.1605</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>sgsq linear</td>
<td>1</td>
<td>6.759145E-8</td>
<td>6.759145E-8</td>
<td>48.13</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgsq quadratic</td>
<td>1</td>
<td>1.501632E-8</td>
<td>1.501632E-8</td>
<td>10.69</td>
<td>0.0012</td>
</tr>
<tr>
<td>simp linear</td>
<td>1</td>
<td>1.430463E-6</td>
<td>1.430463E-6</td>
<td>10.85</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>simp quadratic</td>
<td>1</td>
<td>3.424984E-7</td>
<td>3.424984E-7</td>
<td>24.38</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp10</td>
<td>1</td>
<td>5.310332E-8</td>
<td>5.310332E-8</td>
<td>37.81</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp7</td>
<td>1</td>
<td>4.495686E-8</td>
<td>4.495686E-8</td>
<td>32.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs simp3</td>
<td>1</td>
<td>1.888723E-6</td>
<td>1.888723E-6</td>
<td>134.86</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sgsq10</td>
<td>1</td>
<td>5.985005E-8</td>
<td>5.985005E-8</td>
<td>42.62</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sgsq7</td>
<td>1</td>
<td>5.973098E-8</td>
<td>5.973098E-8</td>
<td>42.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>claplace vs sgsq3</td>
<td>1</td>
<td>1.925655E-11</td>
<td>1.925655E-11</td>
<td>0.01</td>
<td>0.9009</td>
</tr>
<tr>
<td>sgsq7 vs simp7</td>
<td>1</td>
<td>2.084115E-7</td>
<td>2.084115E-7</td>
<td>148.40</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgsq10 vs simp10</td>
<td>1</td>
<td>2.257050E-7</td>
<td>2.257050E-7</td>
<td>160.71</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgsq3 vs simp3</td>
<td>1</td>
<td>1.876680E-6</td>
<td>1.876680E-6</td>
<td>336.28</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sgsq7 vs sgsq10</td>
<td>1</td>
<td>5.927857E-14</td>
<td>5.927857E-14</td>
<td>0.00</td>
<td>0.9948</td>
</tr>
</tbody>
</table>

When ulaplace was still included in the analysis, a significant difference was detected between ulaplace and claplace. It was then discarded from the analysis as before. There were linear and quadratic trends detected for both the sgsq and simp methods, but sgsq7 was not different from sgsq10, showing that 7 abscissas was adequate once again for sgsq. sgsq once again was significantly better than simp at each intensity level, showing that even $10^2 = 100$ importance samples was still not good enough. claplace actually performed better than simp10, was equal in success to sgsq3, but not as good as sgsq7 or sgsq10.
3.5.4 Summary

The viable methods have been shown to be \textit{claplace}, \textit{simp}, and \textit{sgq}. For all models \textit{sgq} was never significantly different from the “truth.” The performance of \textit{claplace} varied somewhat but was very good for the case of constant variance and competed well with \textit{simp} even when \( \sigma \) was increased from 0.15 to 0.3. The clear choice would be to use \textit{sgq} with 7 abscissas. This method at this level of intensity did a good job of estimating the log–likelihood. If computing time is an issue, \textit{claplace} does a fairly good job compared to the other methods with 3 abscissas.

Since the number of viable methods have been reduced, a more stringent experiment could be performed on just these methods with more models and levels of the parameters. Specifically, another experiment could be performed to test more rigorously the effect of increasing \( \sigma \) has on the accuracy of \textit{claplace}. Other numerical methods thought to perform well such as the radial spherical method (see Monahan and Genz, 1997) could also be included.
Figure 3.1: Model 1, Means of Relative Error for the Laplacian Methods
Figure 3.2: Model 1. Means of Relative Error for the Importance Sampling Methods.
Figure 3.3: Model 1, Means of Relative Error for the Gaussian Quadrature Methods
Figure 3.4: Model 1, Means of Relative Error for the Gaussian Quadrature Methods without $gq$

-0.001 0.004 0.009 0.014 0.019 0.024

Intensity

Relative Error

sgq
agq
Figure 3.5: Model 1, Means of Relative Error for the Best Methods
Figure 3.6: Model 2, $\sigma = 0.15$, Means of Relative Error for the Laplacian Methods
Figure 3.7: Model 2, $\sigma = 0.15$, Means of Relative Error for the Importance Sampling Methods
Figure 3.8: Model 2, $\sigma = 0.15$, Means of Relative Error for the Gaussian Quadrature Methods.
Figure 3.9: Model 2, \( \sigma = 0.15 \), Means of Relative Error for the Best Methods
Figure 3.10: Model 2, $\sigma = 0.30$, Means of Relative Error for the Laplacian Methods
Figure 3.11: Model 2, $\sigma = 0.30$, Means of Relative Error for the Importance Sampling Methods.
Figure 3.12: Model 2, $\sigma = 0.30$, Means of Relative Error for the Gaussian Quadrature Methods
Figure 3.13: Model 2, $\sigma = 0.30$, Means of Relative Error for the Best Methods
Figure 3.14: Model 3, Means of Relative Error for the Laplacian Methods
Figure 3.15: Model 3. Means of Relative Error for the Importance Sampling Methods.
Figure 3.16: Model 3, Means of Relative Error for the Gaussian Quadrature Methods.
Figure 3.17: Model 3, Means of Relative Error for the Best Gaussian Quadrature Methods
Figure 3.18: Model 3, Means of Relative Error for the Best Methods
Chapter 4

Computational Framework for

Maximum Likelihood Estimation

4.1 Introduction

In Chapter 3, several approaches were taken to approximate the likelihood of NLMMs. To achieve accurate results, quadrature methods may be necessary with intensive sampling. These methods become extremely computationally demanding when linked with an optimizing algorithm to find the maximum likelihood estimates. As described in Chapter 1, optimizing algorithms have trouble with random input.

Due to increasing speed of computing and the expectation of faster computers in the future, methods should be developed that incorporate high intensity integration methods. Currently, it is not feasible to link the integration method directly with an
optimizer. Another option is to incorporate stochastic approximation, the method of iterating between sampling and estimation until the desired level of accuracy is reached.

In this chapter, a method is derived which maximizes the likelihood of a NLMM with high accuracy that involves a version of stochastic approximation and a simple, but accurate optimizer. The algorithm for this method was programmed in S–Plus with the following main steps (a flowchart diagram is given in Figure 4.1 to help visualize this process): read in the data and the mean function; find an estimate for the parameters and their standard errors (using a simple approach such as nonlinear least squares); construct a grid in the parameter space centered at the first estimate with spacings determined by their standard errors; sample the likelihood on the grid by using a numerical integration technique to integrate out the random effects; fit a quadratic surface to the likelihood and find the maximum at its vertex; construct a new grid in the parameter space by moving the center to the maximum of the quadratic surface and shrink the grid spacing; use numerical integration to evaluate the likelihood on the new grid; fit a new quadratic surface and find its maximum; and repeat until the maximum likelihood estimate converges. In the following sections the algorithm is described in more detail. First, the model is introduced in Section 4.2. Secondly, the numerical integration method is discussed in Section 4.3. Next, the algorithm is more fully developed with descriptions of how the initial nonlinear least squares estimate is found, construction of the grid, the quadratic fitting, and
the iterative process. Two simulations were performed to test the algorithm. The algorithm was first tested on linear data (see Section 4.6). The curvature of the likelihood for the linear data was found to be complex but the algorithm had no trouble converging to the maximum likelihood estimates. A second simulation was performed on nonlinear data. This study showed that the algorithm had difficulty finding a solution for nonlinear data and that finding proper starting values was consternating. This simulation is described in Section 4.7.
Read data and mean function.

Get nonlinear least squares estimate for each individual.

Construct grid in parameter space.

Use random numerical integration to estimate the likelihood on the grid.

Use regression to estimate a quadratic response surface of the likelihood.

Find the maximum of the quadratic surface.

Convergence criterion met?

Yes

No

Recenter grid.

Figure 4.1: Flowchart for Algorithm
4.2 The Model

Consider the following Hierarchical NLMM,

\[ y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \]  \hspace{1cm} (4.1)

\[ (e_i|\beta_i) \sim N(0, \sigma^2 I) \]

\[ \beta_i \sim N(\beta, D). \]

where \( y_{ij} \) is the \( j \)th response for the \( i \)th individual, \( i = 1, ..., M, j = 1, ..., n_i \); \( x_{ij} \) is the covariate for the \( i \)th individual for the \( j \)th response; \( \beta_i \) is a \( p \times 1 \) random effects vector for the \( i \)th individual with mean \( \beta \) \((p \times 1)\) and variance \( D \) \((p \times p)\); \( f(x_{ij}, \beta_i) \) is a nonlinear scalar mean function which will also be denoted as a vector mean function \( f(\beta_i) = [f_1(\beta_i), ..., f_{n_i}(\beta_i)]^T \); and the error for the \( j \)th response of the \( i \)th individual is \( e_{ij} \).

Let the marginal distribution of \( y_i \) given \( \beta_i \) be \( p(y_i|\beta_i) \) and the distribution of \( \beta_i \) be \( \pi(\beta_i) \). Because the distributions are assumed to be normal the distribution functions are

\[ \pi(\beta_i) = \frac{1}{(2\pi)^{p/2}} |D^{-1}|^{1/2} \exp \left[ -\frac{1}{2} (\beta_i - \beta)^T D^{-1} (\beta_i - \beta) \right] \]

and

\[ p(y_i|\beta_i) = \frac{1}{(2\pi)^{n_i/2}} \frac{1}{\sigma^{n_i}} \exp \left\{ -\frac{1}{2\sigma^2} [y_i - f_i(\beta_i)]^T [y_i - f_i(\beta_i)] \right\}. \]
Combining the unknown fixed parameters into one vector, let $\phi = [\beta^T, \xi^T, \text{vech} (D)^T]^T$.

Then the maximum likelihood estimates for $\phi$ can be found by maximizing in $\phi$

$$L(\phi) = \prod_{i=1}^{M} \int p(y_i|b_i) \pi(b_i) \, db_i$$

$$= \prod_{i=1}^{M} \int \frac{1}{(2\pi)^{n_i+p}|D^{-1}|^{1/2}} \exp \left\{ -\frac{1}{2} (\beta_i - \beta)^T D^{-1} (\beta_i - \beta) - \frac{1}{2\sigma^2} [y_i - f_i(\beta_i)]^T [y_i - f_i(\beta_i)] \right\} \, d\beta_i$$

$$= \int g_1(\beta_i) \, d\beta_i$$

where $M$ is the number of individuals.

### 4.3 Integration Method

To illustrate the numerical integration of (4.2), the method of importance sampling is outlined here. However, other methods could result in higher efficiency. The choice of the numerical integration technique should be a random method so that standard errors for the computation of the integral can be computed. (See Appendix A.2.)

To use importance sampling, a function $g_2(\beta)$ such that $g_2(\beta_i) \approx g_1(\beta_i)$ and a constant $A$ such that $Ag_2(\beta_i)$ is a p.d.f., are needed. First note that (by the quadratic Mean Value Theorem)

$$[y_i - f_i(\beta_i)]^T [y_i - f_i(\beta_i)] =$$

$$= \left[ y_i - f_i(\hat{\beta}_i) \right]^T [y_i - f_i(\hat{\beta}_i)] + 2 \left\{ -F(\beta)^T [y_i - f_i(\beta)] \bigg|_{\beta = \hat{\beta}_i} \right\} (\beta_i - \hat{\beta}_i)$$
$$+ \frac{1}{2} (\beta_i - t)^T \left( 2 \left( F(t)^T F(t) + \sum_{j=1}^{n} [y_{ij} - f_{ij}(t)] \nabla^2 f_j(t) \right) \right) (\beta_i - t)$$

for some $t$ in a neighborhood of $\beta_k$ where $\hat{\beta}_i$ is the Ordinary Least Squares (O.L.S.) estimate of $\beta$ and $F(\beta) = \nabla f_i(\beta)$. Since $\hat{\beta}_i$ is the O.L.S. estimate,

$$\nabla \left[ y_i - f_i(\hat{\beta}_i) \right]^T \left[ y_i - f_i(\hat{\beta}_i) \right] \bigg|_{\beta = \hat{\beta}_i} = 0.$$ 

Also, the Hessian at $t$ is $H(t) = F^T F + \sum_{j=1}^{n} [y_{ij} - f_{ij}(t)] \nabla^2 f_j(t)$, which implies that

$$[y_i - f_i(\hat{\beta}_i)]^T [y_i - f_i(\hat{\beta}_i)] = [y_i - f_i(\hat{\beta}_i)]^T [y_i - f_i(\hat{\beta}_i)] + (\beta_i - \hat{\beta}_i)^T H(t) (\beta_i - \hat{\beta}_i)$$

$$\approx [y_i - f_i(\hat{\beta}_i)]^T [y_i - f_i(\hat{\beta}_i)] + (\beta_i - \hat{\beta}_i)^T H(\hat{\beta}_i) (\beta_i - \hat{\beta}_i)$$

(4.3)

Let

$$g_2(\beta_i) = \frac{1}{(2\pi)^{n+1/2} \sigma^{n+2}} |D^{-1}|^{1/2} \exp \left\{ -\frac{1}{2} (\beta_i - \beta)^T D^{-1} (\beta_i - \beta) ight\}$$

$$- \frac{1}{2\pi} [y_i - f_i(\hat{\beta}_i)]^T [y_i - f_i(\hat{\beta}_i)] - \frac{1}{2\pi} (\beta_i - \hat{\beta}_i)^T H(\hat{\beta}_i) (\beta_i - \hat{\beta}_i) \}.$$ 

where $H$ refers to the Hessian at $\hat{\beta}_i$.

Then $g_1(\beta_i) \approx g_2(\beta_i)$ by (4.3). To find $A$ such that $Ag_2(\beta_i)$ is a p.d.f., let

$$\tilde{\beta}_i = \left( D^{-1} + \frac{H}{\sigma^2} \right)^{-1} \left( D^{-1} \beta + \frac{H}{\sigma^2} \hat{\beta}_i \right)$$

(4.4)

so that $(\beta_i - \beta)^T D^{-1} (\beta_i - \beta) + (\beta_i - \hat{\beta}_i)^T \frac{H}{\sigma^2} (\beta_i - \hat{\beta}_i) =$

$$= \beta_i^T D^{-1} \beta_i - 2\beta_i^T D^{-1} \beta_i + \beta_i^T D^{-1} \beta + \beta_i^T \frac{H}{\sigma^2} \beta_i - 2 \hat{\beta}_i \frac{H}{\sigma^2} \beta_i + \tilde{\beta}_i \frac{H}{\sigma^2} \tilde{\beta}_i$$

$$= \beta_i^T \left( D^{-1} + \frac{H}{\sigma^2} \right) \beta_i - 2 \left( \beta_i^T D^{-1} + \beta_i^T \frac{H}{\sigma^2} \right) \beta_i + \beta_i^T D^{-1} \beta + \tilde{\beta}_i \frac{H}{\sigma^2} \tilde{\beta}_i$$

$$= \beta_i^T \left( D^{-1} + \frac{H}{\sigma^2} \right) \beta_i$$
\[-2\tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \beta_i \]
\[+3\tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \tilde{\beta}_i - \tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \tilde{\beta}_i \]
\[+\beta^T \beta^{-1} \beta + \tilde{\beta}_i^T \frac{H}{\sigma^2} \tilde{\beta}_i \]

by \((4.4)\)

\[= (\beta_i - \tilde{\beta}_i)^T \left(D^{-1} + \frac{H}{\sigma^2}\right) (\beta_i - \tilde{\beta}_i) \]
\[-\tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \tilde{\beta}_i \]
\[+\beta^T \beta^{-1} \beta + \tilde{\beta}_i^T \frac{H}{\sigma^2} \tilde{\beta}_i \]

This results in

\[g_2(\beta_i) = \frac{1}{(2\pi)^{n_i+p/2}} \frac{1}{\sigma^{|n_i|/2}} |D^{-1}|^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} [y_i - f_i (\tilde{\beta}_i)]^T [y_i - f_i (\tilde{\beta})] \right\} \]
\[+\frac{1}{2} \tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \tilde{\beta}_i \]
\[-\frac{1}{2} \beta^T \beta^{-1} \beta \]
\[-\frac{1}{2} \tilde{\beta}_i^T \frac{H}{\sigma^2} \tilde{\beta}_i \]
\[-\frac{1}{2} \left(\beta_i - \tilde{\beta}_i\right)^T \left(D^{-1} + \frac{H}{\sigma^2}\right) (\beta_i - \tilde{\beta}_i) \}

\[= \frac{(2\pi)^{p/2}}{|D^{-1} + \frac{H}{\sigma^2}|^{1/2}} \frac{1}{(2\pi)^{n_i+p/2}} \frac{1}{\sigma^{|n_i|/2}} |D^{-1}|^{1/2} \]
\[\exp \left\{ -\frac{1}{2\sigma^2} [y_i - f_i (\tilde{\beta}_i)]^T [y_i - f_i (\tilde{\beta})] \right\} \]
\[+\frac{1}{2} \tilde{\beta}_i^T \left(D^{-1} + \frac{H}{\sigma^2}\right) \tilde{\beta}_i \]
\[-\frac{1}{2} \beta^T \beta^{-1} \beta \]
\[-\frac{1}{2} \tilde{\beta}_i^T \frac{H}{\sigma^2} \tilde{\beta}_i \} \]
\[\frac{1}{(2\pi)^{p/2}} |D^{-1} + \frac{H}{\sigma^2}|^{1/2} \exp \left\{ -\frac{1}{2} \left(\beta_i - \tilde{\beta}_i\right)^T \left(D^{-1} + \frac{H}{\sigma^2}\right) (\beta_i - \tilde{\beta}_i) \right\} \]
$$L_i = \int g_1 (\beta_i) d\beta_i = \int g_1 (\hat{\beta}_i) g_2 (\beta_i) d\beta_i = \int g_1 (\hat{\beta}_i) \frac{1}{A} g_3 (\beta_i) d\beta_i = \frac{1}{A} E \left( \frac{g_1 (\beta_i)}{g_2 (\beta_i)} \right)$$

where $\beta_i \sim N \left( \hat{\beta}_i, \left[ D^{-1} + \frac{H}{\sigma^2} \right]^{-1} \right)$ is the importance distribution. The numerical integration can now be performed as per Section 1.3.3. The numerical integration has now been developed for models with normal assumptions, but can be modified for models with other distributional assumptions.
4.4 Initial Estimation

An initial estimate is needed to center the grid for sampling the likelihood surface. Many points are sampled about this estimate and a quadratic approximation is made to the data of likelihood values. The grid needs to contain an adequate number of points to be able to fit a quadratic surface in \((p + 2)(p + 1)/2\) dimensions (the dimensionality of the parameter space).

The SAS procedure `FACTEX` can be used to design a grid which allows for a resolution 5 factorial experiment. (Resolution 5 requires that all main effects and first order interactions to not be aliased with any other main effect or first order interaction.) An appropriate choice would be to set the options in the SAS procedure for a grid with 5 levels per factor with \((p + 2)(p + 1)/2\) factors. The levels of each factor can be chosen so that the middle level falls on the values of the initial estimate and the spacings between the levels are small enough so that the surface fitted to the sample from the likelihood is quadratic (see Section 4.5) but large enough so that the maximum of the surface is inside the grid. This is perhaps the toughest obstacle faced using this method.

The choice for the initial estimate could be as simple as the nonlinear least squares estimate if there are enough observations on each subject (more than 2 observations per subject). This estimate will be easily available for any NLMM that has constant variance. The estimate for \(\beta_i\) for each subject is needed. The \(i\)th nonlinear least
squares estimate (i.e. the O.L.S. estimate), \( \hat{\beta}_i \), solves the \( p \) estimating equations
\[
\sum_{j=1}^{n_i} [y_{ij} - f(x_{ij}, \beta_i)] f_{\beta_i}(x_{ij}, \beta_i) = 0
\]
where \( f_{\beta_i}(x_{ij}, \beta_i) = \partial / \partial \beta_i f(x_{ij}, \beta_i) \).

The sample mean, \( \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_i \), can be used for the initial estimate for \( \beta \); a sample covariance matrix for the \( \hat{\beta}_i \) can be used for an initial estimate for \( D \); and the sample mean of the \( \sigma_i \) from each fitting can be used to estimate \( \sigma \). To choose the spacing between the intervals, the standard errors of \( \hat{\beta} \) can be taken as the square root of the diagonal elements of the matrix \( D \). Assuming a \( \chi^2 \) distribution for the elements of \( D \) and for \( \sigma \), the standard errors for these elements of \( \phi \) can be generated as \( \sqrt{2/(n-1)} \) times the element, e.g. the standard error for \( \hat{D}_{11} \) could be approximated as \( \sqrt{2/(n-1)} \hat{D}_{11} \). The spacings between the levels for each factor can now be based on the standard errors or confidence intervals. It may be necessary to choose a small confidence in each parameter to be sure that the likelihood surface is quadratic in the region sampled.

It is also necessary to reparameterize \( \phi \) when constructing the grid. It will be necessary for any sampled point in the parameter space to generate a matrix \( D \) that is positive definite. If \( D \) is not positive definite, the formulas involved in evaluating the likelihood will generate infinite results and cause many numerical errors. Since it is not realistic to have a negative definite variance in any event, this parameterization will restrict the sampling to real situations.

One possible parameterization is the following which is written for the case where
$D$ is a $(2 \times 2)$ matrix but can be generalized for other dimensionalities. Let

\[
\begin{align*}
    t_1 & = \frac{1}{2} \ln D_{11} \\
    t_2 & = \frac{1}{2} \ln D_{22} \\
    t_3 & = \sqrt{\frac{D_{12}^2}{D_{11}D_{22} - D_{12}^2}} \\
    t_4 & = \ln \sigma
\end{align*}
\]

Notice that $\sigma$ was also reparameterized. To find standard errors for the new parameterization one could use the “$\Delta$”-method described in Appendix A.1.

Another option for an initial estimate could be to use one of the common approximate methods such as Laplace’s method if convergence can be achieved. The SAS procedure \texttt{NLMIXED} estimates both the fixed parameters and their standard errors, including standard errors for the elements of $D$ and $\sigma$. This procedure also requires a model with constant error variance. Particularly helpful is that this procedure reports the sample covariance matrix for all the estimates. This matrix can be used in a multivariate form of the “$\Delta$”-method to find a suitable grid for the reparameterization. This is also explained in Appendix A.1.

Another approach needs to be developed if there are too few observations to use one of the existing methods for an initial estimation.
4.5 Optimization via a Heuristic Variation on Stochastic Approximation

Once a grid is sampled from the parameter space, the numerical integration method can be used to estimate the likelihood at each point on the grid. Because the likelihood surface is approximately quadratic in a neighborhood about the MLE (Monahan, 2000), \( L(\phi) \) is further approximated with a quadratic function,

\[
\hat{L}_q = a + b^T \phi + \phi^T C \phi.
\] (4.5)

If the Hessian of the quadratic surface is negative-definite, then the maximum of the quadratic surface is at its vertex. Once the coefficients for the quadratic surface are estimated, finding the vertex is elementary and the two procedures combined are much less computationally intense than most maximization algorithms.

In practice, it is not likely that the MLE will be found at the end of the first iteration of this method. The vertex may be outside the grid or the grid may be too large. In either situation a quadratic approximation to the likelihood is not appropriate. An algorithm is needed to deal with these issues.

If the vertex is outside the grid, a new grid must be constructed at the vertex or at a dampened step towards the vertex. The process of numerical integration at each point in the new grid, followed by a quadratic approximation to the likelihood, and finally a maximization by finding the vertex, all follow in due course.

If the Hessian is not negative-definite, then a new, smaller grid must be constructed
and the process repeated until a negative-definite Hessian is attained.

Once the vertex is contained in the grid and the Hessian is negative-definite, the algorithm should still be iterated by moving the center of the grid to the new vertex and shrinking the grid size until convergence of the parameters is achieved. In this manner, the method of stochastic approximation is included in the algorithm; new points are added for approximating the likelihood surface until the accuracy of the quadratic fit is sufficient. Unlike historical uses of stochastic approximation (see Liddle and Monahan, 1988), instead of adding new points to existing data to increase knowledge of the surface, the new data must replace the old. The old data that are not found within a sufficiently small neighborhood about the MLE will be influential in the efforts to fit a quadratic surface.

By checking the value of $R^2$, the appropriateness of the quadratic fit can be evaluated. If the $R^2$ is low, then the grid may need to be reconstructed on a smaller scale about the initial estimate. A high $R^2$ could mean that the maximum of the surface may be found within the grid or that the space where the grid was sampled only represents a side of the quadratic surface, similar to a plane. In the latter case, the maximum would lie somewhere outside the grid. These two situations can be differentiated by simply checking to see that the vertex lies within the grid. The formula for finding the vertex of this quadratic function (4.5) is $\phi = -C^{-1}b$.

The eigenvalues of the matrix $C$ also need to be checked to be sure that the quadratic surface is concave down, i.e. the extreme value corresponds to a maximum
instead of a minimum. This situation occurs when the eigenvalues are all negative. If there is mix of positive and negative eigenvalues, the vertex corresponds to a saddle point requiring the grid to be revised until negative eigenvalues are achieved. In this situation, if the eigenvalues are not all negative, a few options are available: the grid can be reconstructed on a smaller scale about the same previous center, the center can be relocated, or both. It may be very difficult to construct a grid resulting in negative eigenvalues for $C$.

### 4.6 Simulation for a Linear Model

To investigate the performance of this new method, it was appropriate to first check its application to a linear mixed model where the likelihood is well known. The complexity of the likelihood surface for a linear model represents the best possible scenario; a NLMM certainly has a more complex likelihood surface.

#### 4.6.1 Model

Data were simulated from the following linear model.

\[
y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \\
= \beta_1 + \beta_2 t_{ij} + e_{ij} \\
(e_i|\beta_i) \sim N(0, \sigma^2 I_m).
\]
Further, the subject to subject variation is modeled

\[ \beta_i = [\beta_{1i}, \beta_{2i}]^T \]

\[ \beta_i = \beta + b_i, \]

\[ b_i \sim N(0, D), \]

\[ \beta = [\beta_1, \beta_2]^T \text{ and } b_i = [b_{1i}, b_{2i}]^T. \]

Data for 36 subjects were simulated using this model. For each subject, \(t_{ij}\) was set at the values 30, 60, 90, 115, 160, 200, 240, 245, 250, 260, 275, 295, 320, and 360. The fixed parameters were set at \(\beta = [-5.5, -1.9]^T, D = \begin{bmatrix} 0.14 & 0.01 \\ 0.01 & 0.01 \end{bmatrix} \), and \(\sigma = 0.1.\)

### 4.6.2 Results

The algorithm outlined above was programmed in S-Plus and implemented on a Sun Ultra 2 workstation with two 170 MHz CPUs and 128 MB of memory. The general form of the grid was chosen using Proc Factex. To achieve a grid for a resolution 5 factorial experiment with 6 factors, 625 points were needed. The initial choice of center was found using the nls function to first find estimates for the \(\beta_i\) for each individual and then averaging over individuals to estimate \(\beta\). The procedure followed the outline given in Section 4.4. The algorithm was operated manually; the choice for updating the grid from one iteration to the next was not contained within the program but instead influenced by the limits of human patience.
Table 4.1: Results for a Linear Model, $R^2$ and Eigenvalues

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$R^2$</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.990306</td>
<td>-31.8 -79.1 -102.8 -268.8 -915.7 -6742.1</td>
</tr>
<tr>
<td>2</td>
<td>0.999380</td>
<td>-23.8 -68.2 -71.7 -247.3 -855.9 -5846.5</td>
</tr>
<tr>
<td>3</td>
<td>0.999950</td>
<td>-26.2 -76.0 -82.5 -270.7 -863.3 -6677.5</td>
</tr>
<tr>
<td>4</td>
<td>0.999986</td>
<td>-26.7 -76.1 -83.4 -273.0 -864.1 -6704.0</td>
</tr>
</tbody>
</table>

The results for the iterations can be found in Tables 4.1 and 4.2. The $R^2$ values are all very high showing that the quadratic fit is good and gets better as the grid spacings are made smaller from iteration to iteration. The eigenvalues are negative at each iteration showing that the vertex represents a maximum. As outlined in Section 4.5, the vertex from one iteration was chosen to be the center for the grid of the next iteration. These choices were somewhat arbitrary, but yet successful.

Table 4.2: Results for a Linear Model, Estimates

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$D_{11}$</th>
<th>$D_{12}$</th>
<th>$D_{22}$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.4584</td>
<td>-1.89875</td>
<td>0.141</td>
<td>0.0112</td>
<td>0.00707</td>
<td>0.0971</td>
</tr>
<tr>
<td>2</td>
<td>-5.4583</td>
<td>-1.89874</td>
<td>0.128</td>
<td>0.0115</td>
<td>0.00643</td>
<td>0.0966</td>
</tr>
<tr>
<td>3</td>
<td>-5.4583</td>
<td>-1.89874</td>
<td>0.127</td>
<td>0.0113</td>
<td>0.00634</td>
<td>0.0966</td>
</tr>
<tr>
<td>4</td>
<td>-5.4583</td>
<td>-1.89874</td>
<td>0.127</td>
<td>0.0113</td>
<td>0.00634</td>
<td>0.0966</td>
</tr>
</tbody>
</table>

Table 4.3: Results for a Linear Model, Grid Spacings

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0362</td>
<td>0.0362</td>
<td>0.0575</td>
<td>0.0753</td>
</tr>
<tr>
<td>2</td>
<td>0.0090</td>
<td>0.0090</td>
<td>0.0144</td>
<td>0.0188</td>
</tr>
<tr>
<td>3</td>
<td>0.0023</td>
<td>0.0023</td>
<td>0.0036</td>
<td>0.0047</td>
</tr>
<tr>
<td>4</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0018</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Convergence to 3 significant figures in each parameter was reached in only 4
iterations. The results match those using software appropriate for finding maximum likelihood estimates for linear mixed models such as Proc Mixed in SAS. The ease the stochastic method had in converging for this linear model is somewhat misleading. Finding the proper grid spacings for the first iteration was a grueling task. They had to be small enough so that the surface could be approximated well by a quadratic surface and yet still capture the resulting vertex. Several initial guesses were made before finally settling on the successful choice above. With more experience, a proper choice may be found more easily.

4.7 Simulation for a NLMM

After achieving success with a linear mixed model, the method was applied to a nonlinear mixed model.

4.7.1 Model

The following model is based loosely on the dose escalation study of argatroban (see Section 3.3). Let

$$y_{ij} = f(x_{ij}, \beta_i) + e_{ij}$$

$$= \ln \left\{ \frac{D_i}{\text{Cl}_i} \left[ \exp \left( -\frac{\text{Cl}_i t^*_i}{\text{V}_i} \right) - \exp \left( -\frac{\text{Cl}_i t_{ij}}{\text{V}_i} \right) \right] \right\} + e_{ij}$$ (4.6)

$$t^* = \begin{cases} 0 & \text{for } t \leq t_{inf} \\ t - t_{inf} & \text{for } t > t_{inf} \end{cases}$$
Here, $y_{ij}$ is the $j$th concentration measurement taken at dose $D_i$ and time $t_{ij}$ for the $i$th subject, $j = 1, \ldots, n_i$. $Cl_i$ is the clearance and $V_i$ is the volume for the $i$th subject.

$$x_{ij} = [D_i, t_{ij}]^T,$$

$$\xi = \sigma.$$

Further, the subject to subject variation is modeled

$$\beta_i = [\beta_{1i}, \beta_{2i}]^T = [\ln(Cl_i), \ln(V_i)]^T$$

$$\beta_i = \beta + b_i,$$

$$b_i \sim N(0, D),$$

$$\beta = [\beta_1, \beta_2]^T \text{ and } b_i = [b_{1i}, b_{2i}]^T.$$

Data for 36 subjects were simulated using the above model. The model is the same as that used for the argatroban data (see Section 3.3), except that the mean function used here is the log of the mean function used to model the argatroban data. Infusion rates, $D_i$, ranged from 1 to 5 in increments of 0.5. Four subjects were dosed at each of the 9 levels. Each subject had serial blood samples taken at 30, 60, 90, 115, 160, 200, 240, 245, 250, 260, 275, 295, 320, and 360 minutes after infusion started.

The following values were chosen to represent data similar to the argatroban data:

$$t_{inf} = 240, \beta = [-5.5, -1.9]^T, D = \begin{bmatrix} 0.14 & 0.01 \\ 0.01 & 0.1 \end{bmatrix}, \text{ and } \sigma = 0.1. \text{ The } b_i \text{ were sampled from } N(0, D) \text{ and the } e_i \text{ from } N(0, \sigma^2 I). \text{ The } y_{ij} \text{ were then computed from (4.6).}$$
4.7.2 Results

S–Plus was once again used on a Sun Ultra 2 workstation with two 170 MHz CPUs and 128 MB of memory.

At this time, several unfruitful efforts have been made to construct a grid with proper grid spacings to achieve a negative definite matrix $C$. Initial estimates were first found using the \texttt{nls} function in S–Plus and standard errors calculated as suggested in Section 4.4. Another effort was made using estimates from Laplace’s approximation using SAS’s \texttt{Proc NLMIXED}. In both efforts, a negative definite matrix for $C$ was unattainable.

4.8 Discussion

The method works well for a linear mixed model. For these models, starting values can be obtained and the grid spacings reduced until the results are favorable. Of course, this is of no practical use since the likelihood for linear mixed models do not need to be approximated. The integration is exact, resulting in easily used estimating equations.

In dealing with NLMMs, this method has not yet shown success. However, with more effort and experience, an understanding for the choice of grid spacings may be learned. Until that time, this method is not viable.

The computing time involved in using this method to evaluate 625 points of the
likelihood surface is somewhat unwieldy with current machines. When using a shared network server, a Sun Ultra 2, it can take as long as a week for 5 iterations. This method may become more useful as the speed of computers increase, but may still be a better choice than linking the numerical integration with a standard optimizer.

An additional concern is that this method is not feasible for cases when the number of observations on an individual are sparse. Efforts could be made to incorporate prior information to make this a viable method in this situation.
Chapter 5

Conclusions and Future Work

Three simulation studies were conducted to investigate different methods of approximating the likelihood of NLMMs. The first study (described in Chapter 2) was conducted to compare the consequences of misspecification of the random effects when using First Order linearization and Laplace’s approximation. The comparison was made on their ability to estimate the parameters in a NLMM and to detect correctly a covariate effect (both size and power of the test were checked). Several distributions were chosen for simulating the random effects: a normal distribution for the case where the assumptions are satisfied; a $t$ distribution to represent a heavier tailed distribution; contaminated normal distributions, at both mild and moderate contamination levels; an asymmetric distribution; and a bimodal distribution. The composition of the data was changed from intense sampling to more sparse data. Finally, the ability of the models to detect the covariate was also compared when the
secondary structure, the inter–individual model $d(a_i, \beta, b_i)$, was incorrectly specified.

Although Laplace’s approximation showed difficulty in achieving convergence, it was found to be fairly robust to mild deviations from normality of the random effects distribution. The convergence of Laplace’s approximation was also sensitive to misspecification of the secondary structure while the first order method was less sensitive. When both methods converged, Laplace’s approximation seemed to achieve more precise parameter estimation. Also when both methods converged, both Wald and Likelihood Ratio Tests for covariate effects were reliable although liberal if random effects were not normal. Individual estimation results for combinations of method and distribution are reported in Appendix A.3.

Other models should be investigated as well as other methods. In particular, a study designed to compare what causes the non–convergence of the methods is needed. A study similar to that in Chapter 3 could address these issues.

The ability of the methods described in Section 1.3 to find approximate maximum likelihood estimates depends on how well the methods approximate the likelihood at points near the maximum. The second simulation study, described in Chapter 3, compared the ability of Laplace’s approximation, importance sampling, and Gaussian quadrature, to approximate the likelihood surface at several points in the parameter space. Methods developed by Pinheiro and Bates (1995) were enhanced to accommodate models with nonconstant error variance. Also, new versions of importance sampling and Gaussian quadrature were developed which incorporate additional ‘scal-
ing” and “centering” ideas. These newer versions were shown to be more accurate than previously used methods. In particular, the “scaled” and “centered” version of Gaussian quadrature resulted in the most accurate approximations to the likelihood surface while demanding relatively little additional computer time. The most surprising result was that Laplace’s approximation was found to be more accurate than expected, even for a larger intra-individual variance.

The most viable methods were seen to be Laplace’s approximation and the “scaled” and “centered” versions of importance sampling and Gaussian quadrature. This result suggests another study could be performed to more intensely examine just these methods and another method believed to be a possible contender, the spherical-radial method (see Monahan and Genz, 1997). Data from other models could be simulated to ensure the results are not specific to the models chosen in this work. Because Laplace’s method is known to work well for small intra-individual error, larger values for $\sigma$ could be used, increasing $\sigma$ until Laplace’s approximation fails to perform well. The intensity of the numerical integration methods could also be further studied; larger values could be used to see when importance sampling becomes more reliable.

The final study in Chapter 4 attempted to find an alternative method for finding maximum likelihood estimates. The method used importance sampling to integrate out the random effects and evaluated the likelihood on a grid around an initial estimate of its maximum. A quadratic approximation in the parameter vector $\phi$ was fitted to the likelihood surface and maximized. This maximum was believed to be
a good approximation of the maximum likelihood since the likelihood surface is approximately quadratic in an appropriately small enough neighborhood. To continue to hone the estimate, the method incorporated the idea of stochastic approximation by continuing to sample the likelihood surface on grids covering increasingly smaller regions about the current estimate. This process of sampling on a grid, approximating the likelihood with a quadratic surface, finding the maximum of the approximate surface, shrinking the grid–size, and sampling again, was iterated until a convergence criterion was met. This method was shown to work for a linear mixed model but was unsuccessful for data simulated from a NLMM. It was difficult to construct an appropriate grid at the onset. The curvature of the likelihood surface was complex enough that the grid–spacing was either too large to be approximated well by a quadratic surface or the initial estimate too far away from the truth. Due to the latter, the quadratic approximation resulted in a flat surface that mistakenly pushed future “Newton” steps for finding the maximum far away from the region of interest. This study has shown that the likelihood surface deserves future study. The extent of the curvature of the likelihood is unclear. Work should also be continued in finding appropriate starting values/initial estimates for the parameters of the NLMM. If a grid cannot be constructed using existing methods then perhaps it is because current methods are not accurate enough.
Chapter 6

Bibliography


Appendix A

Appendices

A.1 “Δ”–Method for Finding Approximate Variances

The “Δ”–method is used to find approximate variances. If $X$ is a random variable known to have mean $\mu$ and variance $\sigma^2$ and $f(X)$ is a differentiable function then

$$f(X) \approx f(\mu) + f'(\mu)(X - \mu)$$

by a first order Taylor series expansion about $\mu$.

This leads to the result

$$\text{Var}[f(X)] \approx \text{Var}[f(\mu) + f'(\mu)(X - \mu)]$$

$$\approx [f'(\mu)]^2 \sigma^2$$

If $f(X)$ is a $(p \times 1)$ vector function of a random $(q \times 1)$ vector $X$ with mean vector
\( \mu, \) then
\[
f(X) \approx f(\mu) + \nabla f(\mu)(X - \mu)
\]
and
\[
\text{Var}[f(X)] \approx \nabla f(\mu)\text{Var}(X)\nabla f(\mu)^T.
\]

In Section 4.4 approximate standards were found for the parameters \( \beta_1, \beta_2, D_{11}, D_{12}, D_{22}, \) and \( \sigma^2. \) A new parameterization was needed to keep the variance matrix \( D \) non-negative definite. Let \( \hat{\Sigma} \) be the sample variance matrix of the parameter estimates. The new parameterization is
\[
f(\beta_1, \beta_2, D_{11}, D_{12}, D_{22}, \sigma^2) =
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
t_1 \\
t_2 \\
t_3 \\
t_4
\end{pmatrix}
= \begin{pmatrix}
\beta_1 \\
\beta_2 \\
\frac{1}{2} \ln D_{11} \\
\frac{1}{2} \ln D_{22} \\
\sqrt{\frac{D_{12}^2}{D_{11}D_{22} - D_{12}^2}} \\
\ln \sigma
\end{pmatrix}.
\]

The first partial derivative of \( f \) with respect to the parameters \( \beta_1, \beta_2, D_{11}, D_{12}, D_{22}, \) and \( \sigma^2 \) is
This results in $\text{Var}[f(\beta_1, \beta_2, D_{11}, D_{12}, D_{22}, \sigma^2)] \approx \hat{\Omega} F^T$. The standard errors used in the construction of the grid for sampling from the parameter space can be taken as the square roots of the diagonal elements of this variance matrix.
A.2 Standard Errors for Approximations using Importance Sampling

One of the advantages of using a random numerical integration technique such as importance sampling is that standard errors are accessible for estimates of the integral. If \( \int h(x) \, dx \) is the integral to be approximated by importance sampling, i.e.

\[
\int h(x) \, dx = \int f(x) g(x) \, dx = E[f(x)]
\]

where \( g(x) \) is the importance distribution, then

\[
\text{Var}[f(x)] = E\{[f(x)]^2\} - \{E[f(x)]\}^2 \\
\approx \frac{1}{N} \sum_{i=1}^{N} [f(x_i)]^2 - \left[ \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right]^2
\]

where \( x_1, x_2, \ldots, x_N \) is a random sample from the importance distribution.

The standard error of the numerical integration is then

\[
\sqrt{\frac{1}{N} \sum_{i=1}^{N} [f(x_i)]^2 - \left[ \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right]^2}
\]

By coupling this result with the “\( \Delta \)”–method, the standard error for the approximation to the log–likelihood can be approximated. The log–likelihood of interest is

\[
\ln \prod_{i=1}^{M} \int p(y_i | b_i) p(b_i) \, db_i = \sum_{i=1}^{M} \ln \int p(y_i | b_i) p(b_i) \, db_i
\]

which has variance

\[
\sum_{i=1}^{M} \text{Var}[\ln \int p(y_i | b_i) p(b_i) \, db_i] \approx \sum_{i=1}^{M} \frac{1}{[\int p(y_i | b_i) p(b_i) \, db_i]^2} \text{Var}\left[ \int p(y_i | b_i) p(b_i) \right]
\]
where $M$ is the number of subjects and $x_{i1}, x_{i2}, \ldots, x_{iN}$ is the importance sample for the $i$th individual.
### A.3 Additional Tables for Chapter 2

Table A.1: Simulation results for intensive Normal data for testing size.

#### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1758</td>
<td>-21.2356</td>
<td>0.1303</td>
<td>0.1294</td>
<td>1.9232</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0033</td>
<td>.</td>
<td>0.1783</td>
<td>0.1764</td>
<td>3.1813</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.2819</td>
<td>-0.8980</td>
<td>0.0420</td>
<td>0.0424</td>
<td>0.2189</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0837</td>
<td>-2.6410</td>
<td>0.0198</td>
<td>.</td>
<td>0.0399</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0533</td>
<td>2.5890</td>
<td>0.0171</td>
<td>.</td>
<td>0.0295</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0854</td>
<td>-0.6781</td>
<td>0.0208</td>
<td>.</td>
<td>0.0432</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1639</td>
<td>9.2505</td>
<td>0.0087</td>
<td>.</td>
<td>0.0268</td>
</tr>
</tbody>
</table>

The iterative method converged 999 times out of 1000.

#### Laplace's Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2240</td>
<td>0.4048</td>
<td>0.1309</td>
<td>0.1307</td>
<td>1.7142</td>
<td>0.894</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0053</td>
<td>.</td>
<td>0.1784</td>
<td>0.1786</td>
<td>3.1840</td>
<td>1.0008</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.2989</td>
<td>-0.1594</td>
<td>0.0411</td>
<td>0.0419</td>
<td>0.1699</td>
<td>0.7761</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0827</td>
<td>-3.8333</td>
<td>0.0172</td>
<td>.</td>
<td>0.0308</td>
<td>0.7729</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0511</td>
<td>-1.6538</td>
<td>0.0146</td>
<td>.</td>
<td>0.0215</td>
<td>0.7273</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0839</td>
<td>-2.4164</td>
<td>0.0174</td>
<td>.</td>
<td>0.0308</td>
<td>0.7135</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1510</td>
<td>0.6447</td>
<td>0.0063</td>
<td>.</td>
<td>0.0041</td>
<td>0.1525</td>
</tr>
</tbody>
</table>

The iterative method converged 999 times out of 1000.
Table A.2: Simulation results for intensive Normal data for testing power.

### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1936</td>
<td>-13.2390</td>
<td>0.1352</td>
<td>0.1285</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5583</td>
<td>-6.9479</td>
<td>0.1830</td>
<td>0.1751</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3107</td>
<td>0.3529</td>
<td>0.0414</td>
<td>0.0430</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0818</td>
<td>-4.8788</td>
<td>0.0208</td>
<td>0.0215</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0515</td>
<td>-0.9241</td>
<td>0.0177</td>
<td>0.0184</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0862</td>
<td>0.2517</td>
<td>0.0211</td>
<td>0.0215</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1845</td>
<td>22.9762</td>
<td>0.0165</td>
<td>0.0170</td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

### Laplace’s Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2279</td>
<td>2.1263</td>
<td>0.1350</td>
<td>0.1289</td>
<td>1.8254</td>
<td>0.9526</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.5959</td>
<td>-0.6867</td>
<td>0.1820</td>
<td>0.1760</td>
<td>3.3156</td>
<td>0.9410</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3034</td>
<td>0.0342</td>
<td>0.0409</td>
<td>0.0420</td>
<td>0.1677</td>
<td>0.9439</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0819</td>
<td>-4.7382</td>
<td>0.0176</td>
<td>0.0184</td>
<td>0.0328</td>
<td>0.7304</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0509</td>
<td>-2.0669</td>
<td>0.0146</td>
<td>0.0154</td>
<td>0.0213</td>
<td>0.6801</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0843</td>
<td>-1.9639</td>
<td>0.0181</td>
<td>0.0190</td>
<td>0.0330</td>
<td>0.7398</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1504</td>
<td>0.2837</td>
<td>0.0061</td>
<td>0.0061</td>
<td>0.0038</td>
<td>0.0260</td>
</tr>
</tbody>
</table>

The iterative method converged 989 times out of 1000.
Table A.3: Simulation results for intensive t data for testing size.

### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1875</td>
<td>-15.9791</td>
<td>0.1349</td>
<td>0.1309</td>
<td>1.9462</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>-0.0104</td>
<td>.</td>
<td>0.1811</td>
<td>0.1774</td>
<td>3.2889</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.2841</td>
<td>-0.8027</td>
<td>0.0465</td>
<td>0.0433</td>
<td>0.2501</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1031</td>
<td>19.8596</td>
<td>0.1927</td>
<td></td>
<td>3.7435</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0652</td>
<td>25.3089</td>
<td>0.1026</td>
<td></td>
<td>1.0691</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0959</td>
<td>11.4741</td>
<td>0.1144</td>
<td></td>
<td>1.3175</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1686</td>
<td>12.4301</td>
<td>0.0241</td>
<td></td>
<td>0.0929</td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

### Laplace’s Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2390</td>
<td>7.1161</td>
<td>0.1325</td>
<td>0.1279</td>
<td>1.7808</td>
<td>0.9150</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>-0.0114</td>
<td>.</td>
<td>0.1787</td>
<td>0.1747</td>
<td>3.2052</td>
<td>0.9746</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3028</td>
<td>0.0108</td>
<td>0.0430</td>
<td>0.0412</td>
<td>0.1847</td>
<td>0.7387</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0815</td>
<td>-5.2200</td>
<td>0.0280</td>
<td></td>
<td>0.0803</td>
<td>0.0215</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0502</td>
<td>-3.4088</td>
<td>0.0216</td>
<td></td>
<td>0.0468</td>
<td>0.0438</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0820</td>
<td>-4.6040</td>
<td>0.0273</td>
<td></td>
<td>0.0763</td>
<td>0.0579</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1510</td>
<td>0.6843</td>
<td>0.0063</td>
<td></td>
<td>0.0040</td>
<td>0.0436</td>
</tr>
</tbody>
</table>

The iterative method converged 973 times out of 1000.
Table A.4: Simulation results for intensive t data for testing power.

<table>
<thead>
<tr>
<th>Linear Approximation</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2000</td>
<td>-10.3562</td>
<td>0.1457</td>
<td>0.1331</td>
<td>2.1762</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5526</td>
<td>-7.8923</td>
<td>0.1984</td>
<td>0.1812</td>
<td>4.1605</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3061</td>
<td>0.1531</td>
<td>0.0444</td>
<td>0.0439</td>
<td>0.1988</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0928</td>
<td>7.9371</td>
<td>0.0705</td>
<td></td>
<td>0.5015</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0540</td>
<td>3.8682</td>
<td>0.0575</td>
<td></td>
<td>0.3312</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0938</td>
<td>9.0331</td>
<td>0.0749</td>
<td></td>
<td>0.5668</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1986</td>
<td>32.3961</td>
<td>0.0462</td>
<td></td>
<td>0.4494</td>
</tr>
</tbody>
</table>

The iterative method converged 999 times out of 1000.

<table>
<thead>
<tr>
<th>Laplace’s Approximation</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2340</td>
<td>4.8432</td>
<td>0.1336</td>
<td>0.1242</td>
<td>1.7971</td>
<td>0.8258</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5912</td>
<td>-1.4691</td>
<td>0.1801</td>
<td>0.1694</td>
<td>3.2508</td>
<td>0.7814</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3004</td>
<td>-0.0962</td>
<td>0.0421</td>
<td>0.0414</td>
<td>0.1775</td>
<td>0.8930</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0773</td>
<td>-10.1453</td>
<td>0.0238</td>
<td></td>
<td>0.0644</td>
<td>0.1285</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0487</td>
<td>-6.2625</td>
<td>0.0197</td>
<td></td>
<td>0.0397</td>
<td>0.1200</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0827</td>
<td>-3.7864</td>
<td>0.0286</td>
<td></td>
<td>0.0829</td>
<td>0.1462</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1511</td>
<td>0.7094</td>
<td>0.0064</td>
<td></td>
<td>0.0042</td>
<td>0.0094</td>
</tr>
</tbody>
</table>

The iterative method converged 920 times out of 1000.
Table A.5: Simulation results for intensive Contaminated Normal data (with $\alpha = .05$) for testing size.

**Linear Approximation**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1701</td>
<td>-23.7626</td>
<td>0.1441</td>
<td>0.1333</td>
<td>2.3569</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0096</td>
<td>.</td>
<td>0.1926</td>
<td>0.1817</td>
<td>3.7181</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.2826</td>
<td>-0.8689</td>
<td>0.0447</td>
<td>0.0442</td>
<td>0.2396</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0914</td>
<td>6.2828</td>
<td>0.0280</td>
<td></td>
<td>0.0812</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0591</td>
<td>13.6659</td>
<td>0.0215</td>
<td></td>
<td>0.0511</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0934</td>
<td>8.6222</td>
<td>0.0262</td>
<td></td>
<td>0.0742</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1656</td>
<td>10.4326</td>
<td>0.0103</td>
<td></td>
<td>0.0350</td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

**Laplace’s Approximation**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2255</td>
<td>1.0373</td>
<td>0.1459</td>
<td>0.1347</td>
<td>2.1301</td>
<td>0.9038</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0063</td>
<td>.</td>
<td>0.1963</td>
<td>0.1840</td>
<td>3.8585</td>
<td>1.0378</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3020</td>
<td>-0.0256</td>
<td>0.0446</td>
<td>0.0437</td>
<td>0.1985</td>
<td>0.8286</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.0894</td>
<td>3.9726</td>
<td>0.0207</td>
<td></td>
<td>0.0439</td>
<td>0.5413</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0561</td>
<td>7.8916</td>
<td>0.0171</td>
<td></td>
<td>0.0310</td>
<td>0.6072</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0915</td>
<td>6.4344</td>
<td>0.0208</td>
<td></td>
<td>0.0464</td>
<td>0.6257</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1512</td>
<td>0.8181</td>
<td>0.0064</td>
<td></td>
<td>0.0043</td>
<td>0.1223</td>
</tr>
</tbody>
</table>

The iterative method converged 997 times out of 1000.
Table A.6: Simulation results for intensive Contaminated Normal data (with $\alpha = .05$) for testing power.

<table>
<thead>
<tr>
<th>Linear Approximation</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$D_{11}$</th>
<th>$D_{12}$</th>
<th>$D_{22}$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.2231</td>
<td>0.5528</td>
<td>2.3026</td>
<td>0.0860</td>
<td>0.0520</td>
<td>0.0860</td>
<td>0.1500</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>0.6000</td>
<td>2.3091</td>
<td>0.2834</td>
<td>8.2026</td>
<td>9.4405</td>
<td>10.2239</td>
<td>27.7958</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>-13.8666</td>
<td>0.2834</td>
<td>-0.0194</td>
<td>5.3291</td>
<td>7.8023</td>
<td>6.6471</td>
<td>0.8144</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.1480</td>
<td>0.0411</td>
<td>0.0431</td>
<td>0.0211</td>
<td>0.0179</td>
<td>0.0280</td>
<td>0.0204</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.1368</td>
<td>0.0449</td>
<td>0.0437</td>
<td>0.0437</td>
<td>0.0252</td>
<td>0.0280</td>
<td>0.0204</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>2.2853</td>
<td>4.4689</td>
<td>0.1988</td>
<td>0.0960</td>
<td>0.0661</td>
<td>0.0861</td>
<td>0.2154</td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

<table>
<thead>
<tr>
<th>Laplace’s Approximation</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$D_{11}$</th>
<th>$D_{12}$</th>
<th>$D_{22}$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.2318</td>
<td>0.5918</td>
<td>2.3021</td>
<td>0.0906</td>
<td>0.0561</td>
<td>0.0917</td>
<td>0.1512</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>3.8828</td>
<td>-1.3727</td>
<td>-0.0194</td>
<td>5.3291</td>
<td>7.8023</td>
<td>6.6471</td>
<td>0.8144</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>0.1475</td>
<td>0.2045</td>
<td>0.0431</td>
<td>0.0208</td>
<td>0.0179</td>
<td>0.0211</td>
<td>0.0064</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.1350</td>
<td>0.1842</td>
<td>0.0437</td>
<td>0.0452</td>
<td>0.0336</td>
<td>0.0476</td>
<td>0.0043</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>2.1829</td>
<td>4.1869</td>
<td>0.1860</td>
<td>0.4703</td>
<td>0.5079</td>
<td>0.5529</td>
<td>0.0199</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.9552</td>
<td>0.9369</td>
<td>0.9352</td>
<td>0.4703</td>
<td>0.5079</td>
<td>0.5529</td>
<td></td>
</tr>
</tbody>
</table>

The iterative method converged 971 times out of 1000.
Table A.7: Simulation results for intensive Contaminated Normal data (with $\alpha = .10$) for testing size.

### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>$MSE \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1551</td>
<td>-30.5011</td>
<td>0.1713</td>
<td>0.1557</td>
<td>3.3988</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0092</td>
<td>0.2289</td>
<td>0.2109</td>
<td>5.2483</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.2741</td>
<td>-1.2353</td>
<td>0.0562</td>
<td>0.0529</td>
<td>0.3971</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1433</td>
<td>66.6510</td>
<td>0.1230</td>
<td>1.8420</td>
<td></td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0942</td>
<td>81.2012</td>
<td>0.0803</td>
<td>0.8229</td>
<td></td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.1418</td>
<td>64.8618</td>
<td>0.0931</td>
<td>1.1784</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1795</td>
<td>19.6977</td>
<td>0.0250</td>
<td>0.1497</td>
<td></td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

### Laplace’s Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>$MSE \times 100$</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2261</td>
<td>1.3184</td>
<td>0.1604</td>
<td>0.1526</td>
<td>2.5726</td>
<td>0.7569</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0103</td>
<td>0.2202</td>
<td>0.2083</td>
<td>4.8584</td>
<td>0.9257</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3051</td>
<td>0.1087</td>
<td>0.0522</td>
<td>0.0498</td>
<td>0.2726</td>
<td>0.6865</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1169</td>
<td>35.8883</td>
<td>0.0372</td>
<td>0.2333</td>
<td>0.1267</td>
<td></td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0734</td>
<td>41.1832</td>
<td>0.0298</td>
<td>0.1348</td>
<td>0.1638</td>
<td></td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.1218</td>
<td>41.5717</td>
<td>0.0364</td>
<td>0.2605</td>
<td>0.2211</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1510</td>
<td>0.6612</td>
<td>0.0065</td>
<td>0.0044</td>
<td>0.0291</td>
<td></td>
</tr>
</tbody>
</table>

The iterative method converged 918 times out of 1000.
Table A.8: Simulation results for intensive Contaminated Normal data (with $\alpha = .10$) for testing power.

**Linear Approximation**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1676</td>
<td>-24.8911</td>
<td>0.1742</td>
<td>0.1660</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5582</td>
<td>-6.9637</td>
<td>0.2383</td>
<td>0.2256</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3057</td>
<td>0.1359</td>
<td>0.0550</td>
<td>0.0537</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1506</td>
<td>75.0978</td>
<td>0.1064</td>
<td>1.5490</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0874</td>
<td>68.0430</td>
<td>0.0763</td>
<td>0.7078</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.1415</td>
<td>64.5825</td>
<td>0.0861</td>
<td>1.0501</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.2217</td>
<td>47.8033</td>
<td>0.0488</td>
<td>0.7522</td>
</tr>
</tbody>
</table>

The iterative method converged 998 times out of 1000.

**Laplace’s Approximation**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2254</td>
<td>1.0227</td>
<td>0.1614</td>
<td>0.1515</td>
<td>2.6067</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.6059</td>
<td>0.9801</td>
<td>0.2229</td>
<td>0.2065</td>
<td>4.9740</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3039</td>
<td>0.0565</td>
<td>0.0517</td>
<td>0.0499</td>
<td>0.2673</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1186</td>
<td>37.9125</td>
<td>0.0357</td>
<td>0.2335</td>
<td>0.1507</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0749</td>
<td>44.0790</td>
<td>0.0287</td>
<td>0.1351</td>
<td>0.1909</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.1219</td>
<td>41.7935</td>
<td>0.0370</td>
<td>0.2659</td>
<td>0.2532</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1508</td>
<td>0.5144</td>
<td>0.0063</td>
<td>0.0040</td>
<td>0.0053</td>
</tr>
</tbody>
</table>

The iterative method converged 827 times out of 1000.
Table A.9: Simulation results for intensive Asymmetric data for testing size.

### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE \times 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2862</td>
<td>28.2386</td>
<td>0.2694</td>
<td>0.2604</td>
<td>7.6552</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>-0.0101</td>
<td>0.3757</td>
<td>0.3629</td>
<td>14.1245</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3574</td>
<td>2.3787</td>
<td>0.0444</td>
<td>0.0461</td>
<td>0.4972</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.2135</td>
<td>148.2064</td>
<td>0.0358</td>
<td>1.7530</td>
<td></td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>-0.0033</td>
<td>-106.3034</td>
<td>0.0231</td>
<td>0.3591</td>
<td></td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0954</td>
<td>10.8887</td>
<td>0.0227</td>
<td>0.0605</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.2367</td>
<td>57.8035</td>
<td>0.0269</td>
<td>0.8239</td>
<td></td>
</tr>
</tbody>
</table>

The iterative method converged 1000 times out of 1000.

### Laplace’s Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE \times 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.3965</td>
<td>77.6701</td>
<td>0.2595</td>
<td>0.2590</td>
<td>9.7373</td>
<td>1.2720</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.0000</td>
<td>0.0032</td>
<td>.03622</td>
<td>0.3605</td>
<td>13.1218</td>
<td>0.9290</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3018</td>
<td>-0.0336</td>
<td>0.0424</td>
<td>0.0419</td>
<td>0.1795</td>
<td>0.3610</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.2223</td>
<td>158.5018</td>
<td>0.0393</td>
<td>2.0128</td>
<td>1.1482</td>
<td></td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0234</td>
<td>-54.9221</td>
<td>0.0214</td>
<td>0.1272</td>
<td>0.3543</td>
<td></td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0836</td>
<td>-2.8039</td>
<td>0.0176</td>
<td>0.0317</td>
<td>0.5243</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1512</td>
<td>0.8286</td>
<td>0.0061</td>
<td>0.0039</td>
<td>0.0047</td>
<td></td>
</tr>
</tbody>
</table>

The iterative method converged 950 times out of 1000.
Table A.10: Simulation results for intensive Asymmetric data for testing power.

### Linear Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2440</td>
<td>9.3254</td>
<td>0.3109</td>
<td>0.2707</td>
<td>9.7117</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5571</td>
<td>-7.1461</td>
<td>0.4411</td>
<td>0.3770</td>
<td>19.6379</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.4437</td>
<td>6.1304</td>
<td>0.0464</td>
<td>0.0529</td>
<td>2.2076</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.2259</td>
<td>162.6352</td>
<td>0.0401</td>
<td></td>
<td>2.1171</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>-0.0416</td>
<td>-180.0137</td>
<td>0.0423</td>
<td></td>
<td>1.0556</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.1121</td>
<td>30.4065</td>
<td>0.0332</td>
<td></td>
<td>0.1783</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.3711</td>
<td>147.3977</td>
<td>0.0731</td>
<td></td>
<td>5.4226</td>
</tr>
</tbody>
</table>

The iterative method converged 988 times out of 1000.

### Laplace’s Approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE × 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.3697</td>
<td>65.6608</td>
<td>0.2906</td>
<td>0.2710</td>
<td>10.5901</td>
<td>1.0904</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5721</td>
<td>-4.6451</td>
<td>0.4086</td>
<td>0.3773</td>
<td>16.7707</td>
<td>0.8540</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.3037</td>
<td>0.0484</td>
<td>0.0425</td>
<td>0.0414</td>
<td>0.1807</td>
<td>0.0818</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.2410</td>
<td>180.2453</td>
<td>0.0427</td>
<td></td>
<td>2.5848</td>
<td>1.2209</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0140</td>
<td>-73.1066</td>
<td>0.0243</td>
<td></td>
<td>0.2036</td>
<td>0.1929</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0820</td>
<td>-4.6381</td>
<td>0.0183</td>
<td></td>
<td>0.0351</td>
<td>0.1970</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1505</td>
<td>0.3479</td>
<td>0.0059</td>
<td></td>
<td>0.0035</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

The iterative method converged 330 times out of 1000.
Table A.11: Simulation results for intensive Bimodal data for testing size.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>0.2231</td>
<td>0.1868</td>
<td>-16.2733</td>
<td>0.1794</td>
<td>0.1738</td>
<td>3.3508</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.0000</td>
<td>-0.0021</td>
<td>.</td>
<td>0.2506</td>
<td>0.2421</td>
<td>6.2799</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>2.3026</td>
<td>2.4334</td>
<td>5.6812</td>
<td>0.0419</td>
<td>0.0423</td>
<td>1.8868</td>
</tr>
<tr>
<td>( D_{11} )</td>
<td>0.0860</td>
<td>0.0942</td>
<td>9.5500</td>
<td>0.0191</td>
<td>.</td>
<td>0.0432</td>
</tr>
<tr>
<td>( D_{12} )</td>
<td>0.0520</td>
<td>0.0000</td>
<td>-99.9748</td>
<td>0.0152</td>
<td>.</td>
<td>0.2935</td>
</tr>
<tr>
<td>( D_{22} )</td>
<td>0.0860</td>
<td>0.0832</td>
<td>-3.2189</td>
<td>0.0188</td>
<td>.</td>
<td>0.0362</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.1500</td>
<td>0.1821</td>
<td>21.4321</td>
<td>0.0133</td>
<td>.</td>
<td>0.1210</td>
</tr>
</tbody>
</table>

The iterative method converged 998 times out of 1000.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>0.2231</td>
<td>0.2345</td>
<td>5.1085</td>
<td>0.1971</td>
<td>0.1895</td>
<td>3.8985</td>
<td>1.1635</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.0000</td>
<td>-0.0018</td>
<td>.</td>
<td>0.2737</td>
<td>0.2640</td>
<td>7.4894</td>
<td>1.1926</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>2.3026</td>
<td>2.4236</td>
<td>5.2538</td>
<td>0.0424</td>
<td>0.0417</td>
<td>1.6433</td>
<td>0.8709</td>
</tr>
<tr>
<td>( D_{11} )</td>
<td>0.0860</td>
<td>0.1141</td>
<td>32.6457</td>
<td>0.0235</td>
<td>.</td>
<td>0.1342</td>
<td>3.1086</td>
</tr>
<tr>
<td>( D_{12} )</td>
<td>0.0520</td>
<td>0.0031</td>
<td>-94.0975</td>
<td>0.0178</td>
<td>.</td>
<td>0.2712</td>
<td>0.9241</td>
</tr>
<tr>
<td>( D_{22} )</td>
<td>0.0860</td>
<td>0.0829</td>
<td>-3.6448</td>
<td>0.0179</td>
<td>.</td>
<td>0.0331</td>
<td>0.9146</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.1500</td>
<td>0.1512</td>
<td>0.7690</td>
<td>0.0066</td>
<td>.</td>
<td>0.0044</td>
<td>0.0367</td>
</tr>
</tbody>
</table>

The iterative method converged 997 times out of 1000.
Table A.12: Simulation results for intensive Bimodal data for testing power.

<table>
<thead>
<tr>
<th>Linear Approximation</th>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.1662</td>
<td>-25.5361</td>
<td>0.2029</td>
<td>0.1818</td>
<td>4.4402</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.5670</td>
<td>-5.4933</td>
<td>0.2835</td>
<td>0.2533</td>
<td>8.1444</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.4800</td>
<td>7.7045</td>
<td>0.0415</td>
<td>0.0447</td>
<td>3.3192</td>
</tr>
<tr>
<td></td>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1015</td>
<td>18.0148</td>
<td>0.0216</td>
<td></td>
<td>0.0706</td>
</tr>
<tr>
<td></td>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>-0.0133</td>
<td>-125.6718</td>
<td>0.0207</td>
<td></td>
<td>0.4699</td>
</tr>
<tr>
<td></td>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0879</td>
<td>2.2417</td>
<td>0.0217</td>
<td></td>
<td>0.0473</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.2453</td>
<td>63.5086</td>
<td>0.0356</td>
<td></td>
<td>1.0345</td>
</tr>
</tbody>
</table>

The iterative method converged 999 times out of 1000.

<table>
<thead>
<tr>
<th>Laplace’s Approximation</th>
<th>Parameter</th>
<th>True</th>
<th>Mean</th>
<th>RelBias</th>
<th>SD</th>
<th>SE</th>
<th>MSE x 100</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.2231</td>
<td>0.2225</td>
<td>-0.2895</td>
<td>0.2027</td>
<td>0.1904</td>
<td>4.1106</td>
<td>0.9258</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>0.6000</td>
<td>0.6087</td>
<td>1.4441</td>
<td>0.2809</td>
<td>0.2651</td>
<td>7.9000</td>
<td>0.9700</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>2.3026</td>
<td>2.4215</td>
<td>5.1656</td>
<td>0.0414</td>
<td>0.0416</td>
<td>1.5861</td>
<td>0.4778</td>
</tr>
<tr>
<td></td>
<td>$D_{11}$</td>
<td>0.0860</td>
<td>0.1165</td>
<td>35.5128</td>
<td>0.0237</td>
<td></td>
<td>0.1494</td>
<td>2.1161</td>
</tr>
<tr>
<td></td>
<td>$D_{12}$</td>
<td>0.0520</td>
<td>0.0020</td>
<td>-96.1692</td>
<td>0.0188</td>
<td></td>
<td>0.2855</td>
<td>0.6075</td>
</tr>
<tr>
<td></td>
<td>$D_{22}$</td>
<td>0.0860</td>
<td>0.0826</td>
<td>-3.9972</td>
<td>0.0173</td>
<td></td>
<td>0.0312</td>
<td>0.6610</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.1500</td>
<td>0.1509</td>
<td>0.6222</td>
<td>0.0068</td>
<td></td>
<td>0.0047</td>
<td>0.0045</td>
</tr>
</tbody>
</table>

The iterative method converged 958 times out of 1000.