Approximations to the Loglikelihood Function in the Nonlinear Mixed Effects Model

José C. Pinheiro and Douglas M. Bates

Department of Statistics

University of Wisconsin, Madison

Nonlinear mixed effects models have received a great deal of attention in the statistical literature in recent years because of the flexibility they offer in handling unbalanced repeated measures data that arise in different areas of investigation, such as pharmacokinetics and economics. Several different methods for estimating the parameters in nonlinear mixed effects model have been proposed. We concentrate here on two of them: maximum likelihood and restricted maximum likelihood. A rather complex numerical issue for (restricted) maximum likelihood estimation is the evaluation of the loglikelihood function of the data, since it involves the evaluation of a multiple integral that in most cases does not have a closed form expression. We consider here four different approximations to the loglikelihood, comparing their computational and statistical properties. We conclude that the alternating approximation suggested by Lindstrom and Bates (1990), the Laplacian approximation, and Gaussian quadrature centered at the conditional modes of the random effects are quite accurate and computationally efficient. Gaussian quadrature centered at the expected value of the random effects is quite inaccurate for a smaller number of abscissas and computationally inefficient for a larger number of abscissas. Importance sampling is accurate but quite inefficient computationally.

Keywords: Nonlinear mixed effects models, maximum likelihood estimation, Laplacian approximation, Gaussian quadrature, importance sampling.

1 Introduction.

Several different nonlinear mixed effects models and estimation methods for their parameters have been proposed in recent years (Sheiner and Beal, 1980; Mallet, Mentre, Steimer and Lokiek, 1988; Lindstrom and Bates, 1990; Vonesh and Carter, 1992; Davidian and Gallant, 1992; Wakefield, Smith, Racine-Poon and Gelfand, 1994). We consider here a slightly modified version of the model proposed in Lindstrom

and Bates (1990). This model can be viewed as a hierarchical model that in some ways generalizes both the linear mixed effects model of Laird and Ware (1982) and the usual nonlinear model for independent data (Bates and Watts, 1988). In the first stage the jth observation on the ith cluster is modeled as

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \boldsymbol{x}_{ij}) + \epsilon_{ij},\ i = 1, \ldots, M,\ j = 1, \ldots, n_i \tag{1}$$

where f is a nonlinear function of a cluster-specific parameter vector $\boldsymbol{\phi}_{ij}$ and the predictor vector $\boldsymbol{x}_{ij}$, ϵ_{ij} is a normally distributed noise term, M is the total number of clusters, and n_i is the number of observations in the ith cluster. In the second stage the cluster-specific parameter vector is modeled as

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{D}),$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of fixed population parameters, $\boldsymbol{b}_i$ is a q-dimensional random effects vector associated with the ith cluster (not varying with j), $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are design matrices for the fixed and random effects respectively, and $\sigma^2 \boldsymbol{D}$ is a (general) variance-covariance matrix. It is further assumed that observations made on different clusters are independent and that the ϵ_{ij} are *i.i.d.* $\mathcal{N}(0, \sigma^2)$ and independent of the $\boldsymbol{b}_i$.

We consider estimation of the model's parameters by either maximum likelihood, or restricted maximum likelihood, based on the marginal density of $\boldsymbol{y}$

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = \int p(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)\, p(\boldsymbol{b})\, d\boldsymbol{b} \tag{2}$$

In general this integral does not have a closed-form expression when the model function $\boldsymbol{f}$ is nonlinear in $\boldsymbol{b}$ so different approximations have been proposed for estimating it. Some of these methods consist of taking a first order Taylor expansion of the model function $\boldsymbol{f}$ around the expected value of the random effects (Sheiner and Beal, 1980; Vonesh and Carter, 1992), or around the conditional (on $\boldsymbol{D}$) modes of the random effects (Lindstrom and Bates, 1990). Others have proposed the use of Gaussian quadrature rules (Davidian and Gallant, 1992).

We consider here four different approximations to the loglikelihood (2): Lindstrom and Bates (1990)'s alternating method, a modified Laplacian approximation (Tierney and Kadane, 1986), importance sampling (Geweke, 1989), and Gaussian quadrature (Davidian and Gallant, 1992). We compare

them based on their computational and statistical properties, using both real data examples and simulation results. Section 2 contains a description of the different approximations to the loglikelihood as applied to the nonlinear mixed effects model (1). Section 3 presents a comparison of the different approximations based on real and simulated data. Our conclusions and suggestions for further investigation are given in section 4.

2 Approximations to the Loglikelihood

In this section we describe four different approximations to the loglikelihood of $\boldsymbol{y}$ in the nonlinear mixed effects model (1). We show that there exists a close relation between the Laplacian approximation, importance sampling and a Gaussian quadrature rule centered around the conditional modes of the random effects $\boldsymbol{b}$.

2.1 Alternating Approximation

Lindstrom and Bates (1990) propose an alternating algorithm for estimating the parameters in model (1). Based on the current estimates of $\boldsymbol{D}$ (the scaled variance-covariance matrix of the random effects), the conditional modes of the random effects $\boldsymbol{b}$ and the conditional estimates of the fixed effects $\boldsymbol{\beta}$ are obtained by minimizing a penalized nonlinear least squares (PNLS) objective function

$$\sum_{i=1}^{M} \left(\| \boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) \|^2 + \boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i \right) \tag{3}$$

where $[\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)]_j = f\left(\boldsymbol{\phi}_{ij}, \boldsymbol{x}_{ij}\right),\ i = 1, \ldots, M,\ j = 1, \ldots, n_i$.

To update the estimate of $\boldsymbol{D}$ at the wth iteration, Lindstrom and Bates use a first order Taylor expansion of the model function around the current estimates of $\boldsymbol{\beta}$ and the conditional modes of the random effects $\boldsymbol{b}$, which we will denote by $\widehat{\boldsymbol{\beta}}^{(w)}$ and $\widehat{\boldsymbol{b}}^{(w)}$ respectively. Letting

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

the approximate loglikelihood used for the estimation of $\boldsymbol{D}$ is

$$\begin{aligned}\ell_A\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{D} \mid \boldsymbol{y}\right) &= -\frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\sigma^2\left(\boldsymbol{I}+\widehat{\boldsymbol{Z}}_i^{(w)}\boldsymbol{D}\widehat{\boldsymbol{Z}}_i^{(w)^T}\right)\right|\right. \\ &\left. + \sigma^{-2}\left[\widehat{\boldsymbol{w}}_i^{(w)}-\widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]^T\left(\boldsymbol{I}+\widehat{\boldsymbol{Z}}_i^{(w)}\boldsymbol{D}\widehat{\boldsymbol{Z}}_i^{(w)^T}\right)^{-1}\left[\widehat{\boldsymbol{w}}_i^{(w)}-\widehat{\boldsymbol{X}}_i^{(w)}\boldsymbol{\beta}\right]\right\}\end{aligned} \tag{4}$$

This loglikelihood is identical to that of a linear mixed effects (LME) model (Laird and Ware, 1982) in which the response vector is given by $\widehat{\boldsymbol{w}}^{(w)}$ and the fixed and random effects design matrices are given by $\widehat{\boldsymbol{X}}^{(w)}$ and $\widehat{\boldsymbol{Z}}^{(w)}$. Using the results in Lindstrom and Bates (1988) one can express the optimal values of $\boldsymbol{\beta}$ and σ^2 as functions of $\boldsymbol{D}$ and work with the profile loglikelihood of $\boldsymbol{D}$, greatly simplifying the optimization problem. Lindstrom and Bates (1990) have also proposed an approximate restricted loglikelihood for the estimation of $\boldsymbol{D}$

$$\ell_A^R\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{D} \mid \boldsymbol{y}\right) = -\frac{1}{2}\sum_{i=1}^{M}\log\left|\sigma^2\widehat{\boldsymbol{X}}^{(w)^T}\left(\boldsymbol{I}+\widehat{\boldsymbol{Z}}_i^{(w)}\boldsymbol{D}\widehat{\boldsymbol{Z}}_i^{(w)^T}\right)\widehat{\boldsymbol{X}}^{(w)}\right| + \ell_A\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{D} \mid \boldsymbol{y}\right) \tag{5}$$

Their estimation algorithm alternates between the PNLS and LME steps until some convergence criterion is met. Such alternating algorithms tend to be more efficient when the estimates of the variance-covariance components ($\boldsymbol{D}$ and σ^2) are not highly correlated with the estimates of the fixed effects ($\boldsymbol{\beta}$). Pinheiro and Bates (1993) have demonstrated that, in the linear mixed effects model, the (restricted) maximum likelihood estimates of $\boldsymbol{D}$ and σ^2 are asymptotically independent of the (restricted) maximum likelihood estimates of $\boldsymbol{\beta}$. These results have not yet been extended to the nonlinear mixed effects model (1).

It can be shown that the maximum likelihood estimate of $\boldsymbol{\beta}$ and the conditional modes of the random effects $\boldsymbol{b}_i$ corresponding to the approximate loglikelihood (4) are the values obtained in the first iteration of the Gauss-Newton algorithm used to minimize the PNLS objective function (3). Therefore, at the converged value of $\widehat{\boldsymbol{D}}$, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ obtained from the LME and PNLS steps coincide. We will use ℓ_A when comparing the different approximations at the optimal values in section 3, but we do note that in Lindstrom and Bates (1990) approximation (4) is used only to update the estimates of $\boldsymbol{D}$ and not for estimating $\boldsymbol{\beta}$.

2.2 Laplacian Approximation

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

The integral that we want to estimate for the marginal distribution of $\boldsymbol{y}_i$ in model (1) can be written as

$$\begin{aligned} p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) &= \int (2\pi\sigma^2)^{-(n_i+q)/2} |\boldsymbol{D}|^{-1/2} \exp\left[-g(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i)/2\sigma^2\right] d\boldsymbol{b}_i, \quad \text{where} \\ g(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i) &= \|\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\|^2 + \boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i \end{aligned}$$

Let

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

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

$$g(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i) \simeq g\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}\right) + \frac{1}{2}\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}_i}\right]^T g''\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}\right)\left[\boldsymbol{b}_i - \widehat{\boldsymbol{b}_i}\right] \tag{6}$$

where the linear term of the approximation vanishes since $g'(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}) = 0$. The Laplacian approximation is defined as

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

where $N = \sum_{i=1}^{M} n_i$.

Now we consider an approximation to g'' similar to the one used in Gauss-Newton optimization.

We have

$$g''\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}\right) = \left.\frac{\partial^2 \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} \left[\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{\beta}, \widehat{\boldsymbol{b}_i})\right] + \left.\frac{\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^T}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} \left.\frac{\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} + \boldsymbol{D}^{-1}$$

At $\widehat{\boldsymbol{b}_i}$, the contribution of $\left.\partial^2 \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)/\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} \left[\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{\beta}, \widehat{\boldsymbol{b}_i})\right]$ is usually negligible compared to that of $\left.\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)/\partial \boldsymbol{b}_i^T\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} \left.\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)/\partial \boldsymbol{b}_i\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}}$ (Bates and Watts, 1980) so we use the approximation

$$g''\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}\right) \simeq \boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right) = \left.\frac{\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^T}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} \left.\frac{\partial \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}_i}} + \boldsymbol{D}^{-1}$$

This has the advantage of requiring only the first order partial derivatives of the model function with respect to the random effects, which are usually available from the estimation of $\widehat{\boldsymbol{b}_i}$. This estimation of $\widehat{\boldsymbol{b}_i}$ is a penalized least squares problem, for which standard and reliable code is available.

The modified Laplacian approximation to the loglikelihood of model (1) is then given by

$$\begin{aligned} &\ell_{LA}\left(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \mid \boldsymbol{y}\right) = \\ &\quad -\frac{1}{2}\left\{N \log\left(2\pi\sigma^2\right) + M \log|\boldsymbol{D}| + \sum_{i=1}^{M} \log\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right] + \sigma^{-2} \sum_{i=1}^{M} g\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}_i}\right)\right\} \end{aligned} \tag{7}$$

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

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

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

$$\ell_{LAp} = -\frac{1}{2}\left\{N\left[1 + \log\left(2\pi\right) + \log\left(\widehat{\sigma}^2\right)\right] + M \log|\boldsymbol{D}| + \sum_{i=1}^{M} \log\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right]\right\} \tag{8}$$

We note that if $\boldsymbol{f}$ is linear in $\boldsymbol{b}$ then the modified Laplacian approximation is exact because the second order Taylor expansion in (6) is exact when $\boldsymbol{f}\left(\boldsymbol{\beta}, \boldsymbol{b}\right) = \boldsymbol{f}\left(\boldsymbol{\beta}\right) + \boldsymbol{Z}\left(\boldsymbol{\beta}\right)\boldsymbol{b}$..

There does not yet seem to be a straightforward generalization of the concept of restricted maximum likelihood (Harville, 1974) to nonlinear mixed effects models. The difficulty is that restricted maximum likelihood depends heavily upon the linearity of the fixed effects in the model function, which does not occur in nonlinear models. Lindstrom and Bates (1990) circumvented that problem by using an

approximation to the model function $\boldsymbol{f}$ in which the fixed effects $\boldsymbol{\beta}$ occur linearly. This cannot be done for the Laplacian approximation, unless we consider yet another Taylor expansion of the model function, what would lead us back to something very similar to Lindstrom and Bates' approach. We will return to this topic later in section 4.

2.3 Importance Sampling

Importance sampling provides a simple and efficient way of performing Monte Carlo integration. The critical step for the success of this method is the choice of an importance distribution from which the sample is drawn and the importance weights calculated. Ideally this distribution corresponds to the density that we are trying to integrate, but in practice one uses an easily sampled approximation. For the nonlinear mixed effects model the function that we want to integrate is, up to a multiplicative constant, equal to $\exp\left[-g\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i\right)/2\sigma^2\right]$. As shown in subsection 2.2, by taking a second order Taylor expansion of $g(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i)$ around $\widehat{\boldsymbol{b}_i}$ the integrand is, up to a multiplicative constant, approximately equal to a $\mathcal{N}\left(\widehat{\boldsymbol{b}_i}, \sigma^2\left[\boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i)\right]^{-1}\right)$ density. This gives us a natural choice for the importance distribution.

Let N_{IS} denote the number of importance samples to be drawn. In practice one such sample can be generated by selecting a vector $\boldsymbol{z}^*$ with distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and calculating the sample of random effects as $\boldsymbol{b}_i^* = \widehat{\boldsymbol{b}}_i + \sigma\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right]^{-1/2}\boldsymbol{z}^*$, where $\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right]^{-1/2}$ denotes the inverse of the Cholesky factor of $\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)$. The importance sampling approximation to the loglikelihood of $\boldsymbol{y}$ is then defined as

$$\begin{aligned}\ell_{IS}\left(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \mid \boldsymbol{y}\right) &= -\frac{1}{2}\left[N\log\left(2\pi\sigma^2\right) + M\log|\boldsymbol{D}| + \sum_{i=1}^{M}\log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right|\right] \\ &\quad + \sum_{i=1}^{M}\log\left\{\sum_{j=1}^{N_{IS}}\exp\left[-g\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_{ij}^*\right)/2\sigma^2 + \left\|\boldsymbol{z}_j^*\right\|^2/2\right]/N_{IS}\right\}\end{aligned} \tag{9}$$

Note that we cannot in general obtain a closed form expression for the MLE of σ^2 for fixed $\boldsymbol{\beta}$ and $\boldsymbol{D}$, so that profiling on σ^2 is no longer reasonable.

As in the modified Laplacian approximation, importance sampling gives exact results when the

model function is linear in $\boldsymbol{b}$ because in this case

$$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)\, p(\boldsymbol{b}_i) = p\left(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2\right) \cdot \mathcal{N}\left(\widehat{\boldsymbol{b}}_i, \sigma^2 \left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right]^{-1}\right)$$

so that the importance weights are equal to $p\left(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2\right)$.

2.4 Gaussian quadrature

Gaussian quadrature is used to approximate integrals of functions with respect to a given kernel by a weighted average of the integrand evaluated at pre-determined abscissas. The weights and abscissas used in Gaussian quadrature rules for the most common kernels can be obtained from the tables of Abramowitz and Stegun (1964) or by using an algorithm proposed by Golub (1973) (see also Golub and Welsch (1969)). Gaussian quadrature rules for multiple integrals are known to be numerically complex (Davis and Rabinowitz, 1984), but using the structure of the integrand in the nonlinear mixed effects model we can transform the problem into successive applications of simple one dimensional Gaussian quadrature rules. Letting $z_j^*, w_j\ j = 1, \ldots, N_{GQ}$ denote respectively the abscissas and the weights for the (one dimensional) Gaussian quadrature rule with N_{GQ} points based on the $\mathcal{N}(0,1)$ kernel, we get

$$\begin{aligned}
&\int (2\pi\sigma^2)^{-q/2} \left|\boldsymbol{D}\right|^{-1/2} \exp\left[-\left\|\boldsymbol{y}_i - \boldsymbol{f}\left(\boldsymbol{\beta}, \boldsymbol{b}_i\right)\right\|^2 / 2\sigma^2\right] \exp\left(-\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b} / 2\sigma^2\right) d\boldsymbol{b}_i \qquad (10)\\
&= \int (2\pi)^{-q/2} \exp\left[-\left\|\boldsymbol{y}_i - \boldsymbol{f}\left(\boldsymbol{\beta}, \sigma \boldsymbol{D}^{T/2} \boldsymbol{z}^*\right)\right\|^2 / 2\sigma^2\right] \exp\left(-\left\|\boldsymbol{z}^*\right\|^2 / 2\right) d\boldsymbol{z}^*\\
&\simeq \sum_{j_1=1}^{N_{GQ}} \cdots \sum_{j_q=1}^{N_{GQ}} \exp\left[-\left\|\boldsymbol{y}_i - \boldsymbol{f}\left(\boldsymbol{\beta}, \sigma \boldsymbol{D}^{T/2} \boldsymbol{z}^*_{j_1,\ldots,j_q}\right)\right\|^2 / 2\sigma^2\right] \prod_{k=1}^{q} w_{j_k}
\end{aligned}$$

where $\boldsymbol{z}^*_{j_1,\ldots,j_q} = \left(z^*_{j_1}, \ldots, z^*_{j_q}\right)^T$. The corresponding approximation to the loglikelihood function is

$$\begin{aligned}
&\ell_{GQ}\left(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \mid \boldsymbol{y}\right) = \qquad (11)\\
&\quad - N \log(2\pi\sigma^2)/2 + \sum_{i=1}^{M} \log\left\{\sum_{\boldsymbol{j}}^{N_{GQ}} \exp\left[-\left\|\boldsymbol{y}_i - \boldsymbol{f}\left(\boldsymbol{\beta}, \sigma \boldsymbol{D}^{T/2} \boldsymbol{z}^*_{\boldsymbol{j}}\right)\right\|^2 / 2\sigma^2\right] \prod_{k=1}^{q} w_{j_k}\right\}
\end{aligned}$$

where $\boldsymbol{j} = (j_1, \ldots, j_q)^T$.

The Gaussian quadrature rule in this case can be viewed as a deterministic version of Monte Carlo integration in which random samples of $\boldsymbol{b}_i$ are generated from the $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{D})$ distribution. The samples

($\boldsymbol{z}_{\boldsymbol{j}}^{*}$) and the weights ($w_j$) are fixed beforehand, while in Monte Carlo integration they are left to random choice. Since importance sampling tends to be much more efficient than simple Monte Carlo integration, we also consider the equivalent of importance sampling in the Gaussian quadrature context, which we will denote by adaptive Gaussian quadrature. In this approach the grid of abscissas in the $\boldsymbol{b}_i$ scale is centered around the conditional modes $\widehat{\boldsymbol{b}}_i$ rather than $\mathbf{0}$, as in (10). Another modification is the use of $\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)$ instead of $\boldsymbol{D}$ in the scaling of the $\boldsymbol{z}^{*}$. The adaptive Gaussian quadrature is then given by

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

The corresponding approximation to the loglikelihood is then

$$
\begin{aligned}
\ell_{AGQ}\left(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \mid \boldsymbol{y}\right) = &- \left[N \log\left(2\pi\sigma^2\right) + M \log\left|\boldsymbol{D}\right| + \sum_{i=1}^{M} \log\left|\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right|\right] / 2 \\
&+ \sum_{i=1}^{M} \log\left[\sum_{\boldsymbol{j}}^{N_{GQ}} \exp\left(-g\left\{\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \widehat{\boldsymbol{b}}_i + \sigma\left[\boldsymbol{G}\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i\right)\right]^{-1/2} \boldsymbol{z}_{\boldsymbol{j}}^{*}\right\} / 2\sigma^2 + \left\|\boldsymbol{z}_{\boldsymbol{j}}^{*}\right\|^2 / 2\right) \prod_{k=1}^{q} w_{j_k}\right]
\end{aligned}
\tag{12}
$$

The adaptive Gaussian quadrature approximation very closely resembles that obtained for importance sampling. The basic difference is that the former uses fixed abscissas and weights, while the latter allows them to be determined by a pseudo-random mechanism. It is also interesting to note that the one point (i.e. $N_{GQ} = 1$) adaptive Gaussian quadrature approximation is simply the modified Laplacian approximation (8), since in this case $z_1^{*} = 0$ and $w_1 = 1$. The adaptive Gaussian quadrature also gives the exact loglikelihood when the model function is linear in $\boldsymbol{b}$, but that is not true in general for the Gaussian quadrature approximation (10). Like the importance sampling approximation, the Gaussian quadrature approximation cannot be profiled on σ^2 to reduce the dimensionality of the optimization problem.

3 Comparing the Approximations

In this section we present a comparison of the different approximations to the loglikelihood of model (1) described in section 2. Two real data examples, the orange trees and Theophylline data sets, and simulation results are used to compare the statistical and computational aspects of the various approximations.

3.1 Orange Trees

The data are presented on Figure 1 and consist of seven measurements of the trunk circumference (in millimeters) on each of five orange trees, taken over a period of 1600 days. These data were originally presented in Draper and Smith (1981, p. 524) and were described in Lindstrom and Bates (1990).

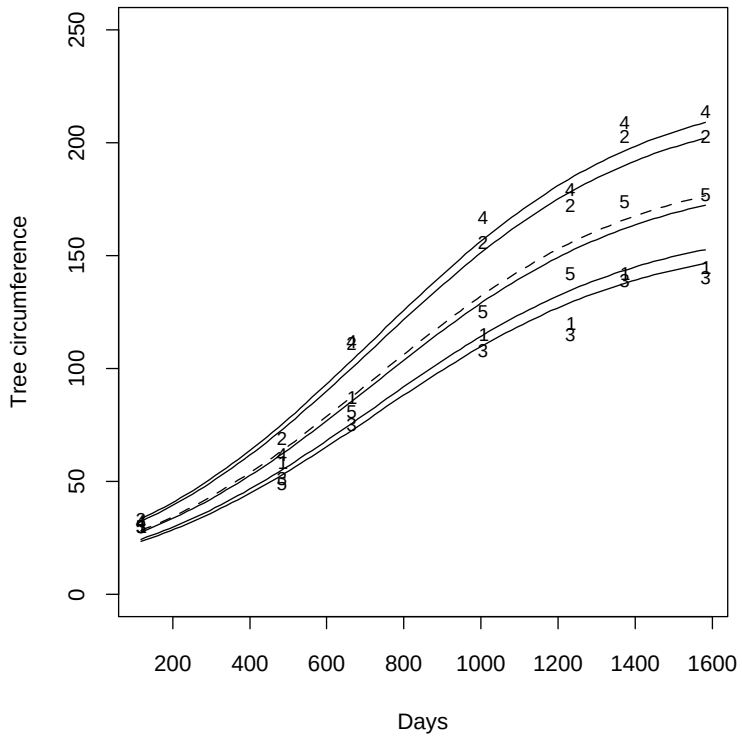

Figure 1: Trunk circumference (in millimeters) of five orange trees: Data and individual fitted curves from maximum likelihood estimation using the exact loglikelihood. The dashed line represents the mean curve.

The logistic model $y = \phi_1 / \{1 + \exp[-(t - \phi_2)/\phi_3]\}$ seems to fit the data well. Lindstrom and Bates (1990) concluded in their analysis that only the asymptotic circumference ϕ_1 needs a random

effect to account for tree to tree variation and suggested the following nonlinear mixed effects model

$$y_{ij} = \frac{\beta_1 + b_{i1}}{1 + \exp\left[-\left(t_{ij} - \beta_2\right)/\beta_3\right]} + \varepsilon_{ij} \tag{13}$$

where y_{ij} represents the jth circumference measurement on the ith tree, t_{ij} represents the day corresponding to the jth measurement on the ith tree, $b_{i1},\ i = 1, \ldots, 5$ are *i.i.d.* $\mathcal{N}(0, \sigma^2 D)$, and $\varepsilon_{ij},\ i = 1, \ldots, 5,\ j = 1, \ldots, 7$ are *i.i.d.* $\mathcal{N}(0, \sigma^2)$ and independent of the b_{i1}. Note that the single random effect occurs linearly in (13) and therefore the modified Laplacian (8), the importance sampling (9), and the adaptive Gaussian quadrature (12) approximations are all exact.

Table 1 presents the results of estimation using the alternating approximation, Gaussian quadrature with 10 and 200 abscissas, and the exact loglikelihood. Since only the alternating approximation provides a version of restricted maximum loglikelihood, we will just consider maximum likelihood estimation in this and the next subsection. The subscript on *Gaussian* refers to the number of abscissas used in the approximation and the scalar L is $\sqrt{D}$, the square root of the scaled variance of the random effects. In general this is a matrix but there is only one random effect here.

Table 1: Estimation Results – Orange Trees Data

Approximation	$\log(L)$	β_1	β_2	β_3	$\log(\sigma^2)$	ℓ
Alternating	1.389	191.049	722.556	344.164	4.120	-131.585
Gaussian_{10}	1.123	194.325	727.490	348.065	4.102	-130.497
Gaussian_{200}	1.396	192.293	727.074	348.074	4.119	-131.571
Exact	1.395	192.053	727.906	348.073	4.119	-131.572

The estimation results in Table 1 indicate that the different approximations produce similar fits. The Gaussian approximation with only 10 abscissas gives the worst approximation, in terms of the value of the loglikelihood, but even that is not far from the exact value. The Gaussian quadrature with 200 abscissas is almost identical to the exact loglikelihood. The alternating approximation is also very close to the exact value.

Another important issue regarding the different approximations is how well they behave in a neighborhood of the optimal value, since this behavior is often used to assess the variability of maximum likelihood estimates. Figure 2 displays the profile traces and contours (Bates and Watts, 1988) for the exact loglikelihood and the alternating approximation. This plot could not be obtained for the Gaussian approximation because the objective function presented several local optima during the profiling algo-

rithm. We believe that this is related to the fact that the Gaussian approximation is centered at $\boldsymbol{b}_i = 0$ and not at the conditional modes of the random effects, where the integrand in (2) takes its highest values.

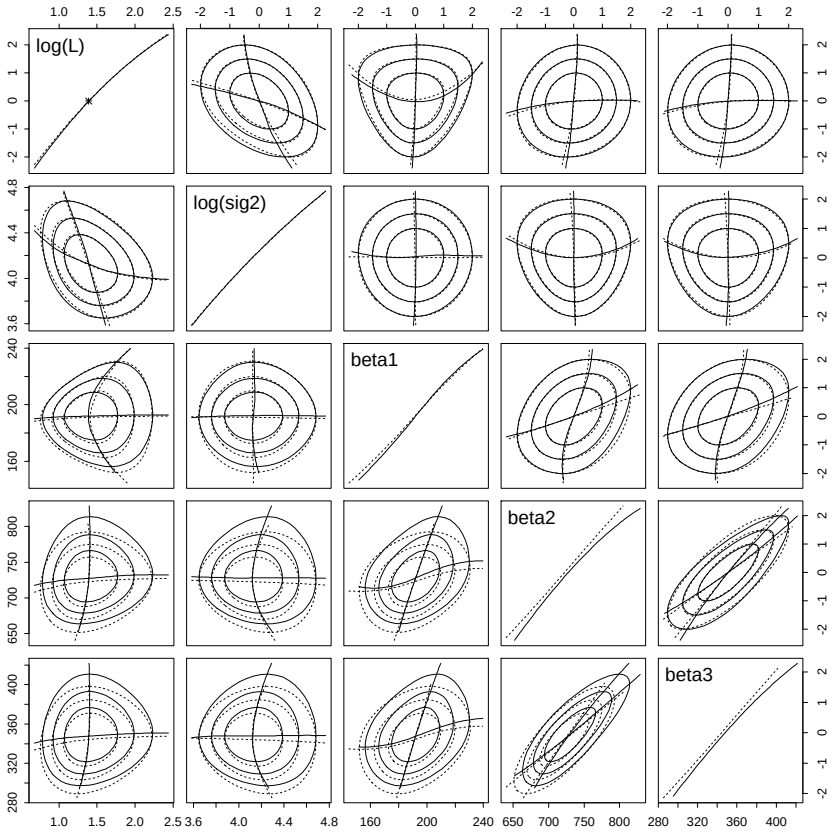

Figure 2: Profile traces and profile contour plots for the orange trees data based on the exact loglikelihood (solid line) and the alternating approximation (dashed line). Plots below the diagonal are in the original scale and plots above the diagonal are in the zeta scale (Bates and Watts, 1988). Interpolated contours correspond approximately to joint confidence levels of 68%, 87%, and 95%.

It can be seen from Figure 2 that the alternating method gives a good approximation to the loglikelihood in a neighborhood of the optimal values. It is interesting to note that the profile traces for the variance-covariance components (D and σ^2) and the fixed effects ($\boldsymbol{\beta}$) meet almost perpendicularly. This indicates a local uncorrelation between the variance-covariance components and the fixed effects, which explains why the alternating method was so successful in approximating the loglikelihood. The same pattern was observed in several other data sets that we have analyzed, leading us to conjecture that the asymptotic uncorrelation between the estimators of the variance-covariance components and the fixed

effects verified in the linear mixed effects model also holds, at least approximately, for the nonlinear mixed effects model.

To compare the computational efficiency of the different approximations we consider the number of function evaluations needed until convergence. For the alternating approximation there are two different functions being evaluated during the iterations: the objective function (3) within the PNLS step and the approximate loglikelihood ℓ_A (4) within the LME step. We will use here the total number of evaluations of either (3) or ℓ_A, multiplied by the number of clusters. For the other approximations we will use the total number of calls to $g(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}_i, \boldsymbol{b}_i)$. Even though the number of function evaluations used for the alternating approximation is not directly comparable to the number of function evaluations of the remaining approximations, it gives a good idea of the relative computational efficiency of this algorithm.

Table 2 presents the number of function evaluations for the different approximations in the orange trees example. The Gaussian quadrature approximations are considerably less efficient than either the alternating approximation or the exact loglikelihood. As expected the alternating approximation is the most computationally efficient.

Table 2: Number of Function Evaluations to Convergence – Orange Trees Data

Approximation	Function Evaluations
Alternating	200
Exact	420
Gaussian_{10}	8,150
Gaussian_{200}	101,000

3.2 Theophylline Kinetics

The data considered here are courtesy of Dr. Robert A. Upton of the University of California, San Francisco. Theophylline was administered orally to 12 subjects whose serum concentrations were measured at 11 times over the next 25 hours. This is an example of a laboratory pharmacokinetic study characterized by many observations on a moderate number of individuals (clusters). Figure 3 displays the data and the individual fits obtained through maximum likelihood using the adaptive Gaussian approximation with 10 abscissas.

A common model for such data is a first order compartment model with absorption in a peripheral

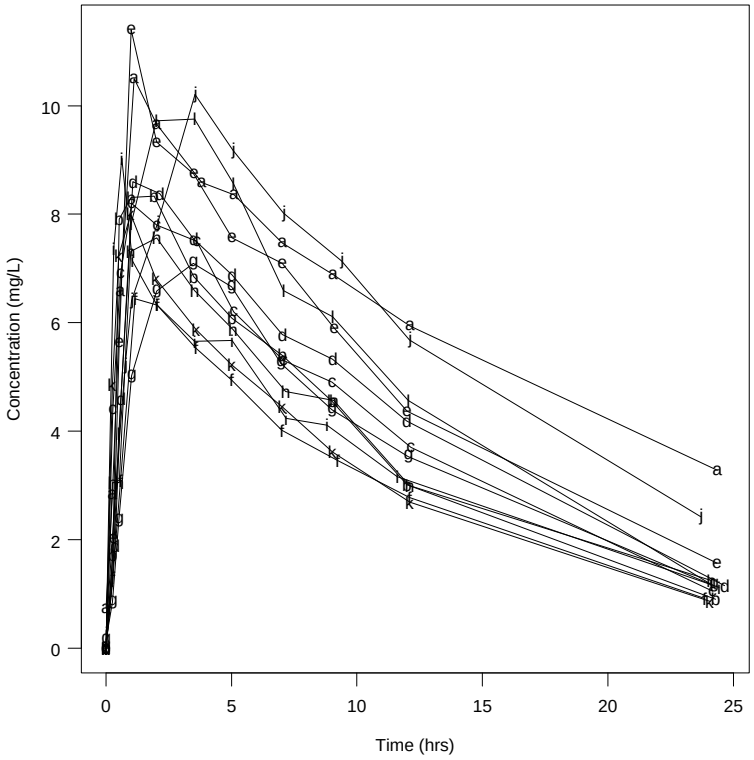

Figure 3: Theophylline concentrations (in mg/L) of twelve patients: Data and individual fitted curves from maximum likelihood estimation using the adaptive Gaussian approximation.

compartment

$$C_t = \frac{DKk_a}{\mathrm{Cl}(k_a - K)} \left[\exp\left(-Kt\right) - \exp\left(-k_a t\right)\right] \tag{14}$$

where C_t is the observed concentration at time t (mg/L), t is the time (hr), D is the dose (mg/kg), Cl is the clearance (L/kg), K is the elimination rate constant (1/hr), and k_a is the absorption rate constant (1/hr). In order to ensure positivity of the rate constants and the clearance, the logarithms of these quantities were used in the fit. Analysis of the Theophylline data using model (14) indicated that only $\log(\mathrm{Cl})$ and $\log(\mathrm{k}_a)$ needed random effects to account for the patient-to-patient variability. The nonlinear mixed effects model used for the Theophylline data is

$$C_t = \frac{D \exp\left[-\left(\beta_1 + b_{i1}\right) + \left(\beta_2 + b_{i2}\right) + \beta_3\right]}{\exp\left(\beta_2 + b_{i2}\right) - \exp\left(\beta_3\right)} \left\{\exp\left[-\exp\left(\beta_3\right) t\right] - \exp\left[-\exp\left(\beta_2 + b_{i2}\right) t\right]\right\} \tag{15}$$

Table 3 presents the estimation results from the various approximations to the loglikelihood. Only maximum likelihood estimation is considered. The subscripts on *Gaussian* and on *Adap*. *Gaussian*

refer to the number of abscissas used in the Gaussian and adaptive Gaussian approximations, while the subscript on *Imp. Sampling* refers to the number of importance samples used in this approximation. $\boldsymbol{L}$ denotes the vector with elements given by the upper triangular half of the Cholesky decomposition of $\boldsymbol{D}$, stacked by columns.

Table 3: Estimation Results – Theophylline Data

Approximation	$\log(L_1)$	L_2	$\log(L_3)$	β_1	β_2	β_3	$\log(\sigma^2)$	ℓ
Alternating	-1.44661	0.00271	-0.09992	-3.22719	0.46548	-2.45464	-0.68660	-177.0237
Laplacian	-1.44376	0.00271	-0.09973	-3.22946	0.46876	-2.46432	-0.68658	-176.9995
Imp. Sampling$_{1000}$	-1.44380	0.00271	-0.09877	-3.22682	0.47614	-2.45851	-0.68747	-177.7689
Gaussian$_5$	-1.55539	0.00241	-0.39687	-3.30411	0.50046	-2.48743	-0.48395	-182.4680
Gaussian$_{10}$	-1.56422	0.00232	-0.20432	-3.23814	0.59525	-2.46872	-0.70276	-176.1008
Gaussian$_{100}$	-1.44572	0.00271	-0.09820	-3.22684	0.47947	-2.45893	-0.68539	-177.7293
Adap. Gaussian$_5$	-1.44600	0.00271	-0.09905	-3.22503	0.47566	-2.45788	-0.68677	-177.7499
Adap. Gaussian$_{10}$	-1.44750	0.00271	-0.09937	-3.22705	0.47377	-2.45942	-0.68533	-177.7473

We can see from Table 3 that the alternating approximation, the Laplacian approximation, the importance sampling approximation, and the adaptive Gaussian approximation all give similar estimation results. The Gaussian approximation only approaches the other approximations when the number of abscissas is increased considerably. Note that the actual number of points used in the grid that defines the Gaussian approximation for this example is the square of the number of abscissas. The adaptive Gaussian approximations for 1 (Laplacian), 5, and 10 abscissas give similar results, indicating that just a few points are needed for this approximation to be accurate. The importance sampling approximation caused some numerical difficulties for the optimization algorithm (the `ms()` function in *S* (Chambers and Hastie, 1992)) used to obtain the maximum likelihood estimates, since the stochastic variability associated with different importance samples overwhelmed the numerical variability of the loglikelihood for small changes in the parameter values (used to calculate numerical derivatives). We ended up having to keep the random number generator seed fixed during the optimization process, thus using the same importance samples throughout the calculations. Since the results obtained using importance sampling were very similar to those of the adaptive Gaussian approximation, we concluded that the latter is to be preferred for its greater simplicity and computational efficiency.

Table 4 gives the number of function evaluations until convergence for the different approximations. The alternating approximation is the most efficient, followed by the Laplacian and adaptive Gaussian approximations. Gaussian quadrature with 5 abscissas is efficient compared to the adaptive Gaussian, but is quite inaccurate. The more reliable Gaussian approximation with 100 abscissas takes about 100

times more function evaluations than the adaptive Gaussian with 10 abscissas. The importance sampling approximation had the worst performance in terms of function evaluations.

Table 4: Number of Function Evaluations to Convergence – Theophylline Data

Approximation	Function Evaluations
Alternating	1,512
Laplacian	7,683
Adap. Gaussian$_5$	30,020
Adap. Gaussian$_{10}$	96,784
Gaussian$_5$	47,700
Gaussian$_{10}$	318,000
Gaussian$_{100}$	10,200,000
Imp. Sampling$_{1000}$	11,211,284

Next we consider the approximations in a neighborhood of the optimal value. We will restrict ourselves here to the alternating, the Laplacian, and the adaptive Gaussian approximation, as the Gaussian approximation for a moderate number of abscissas is not reliable, and both the Gaussian approximation with a larger number of abscissas and the importance sampling approximation are very inefficient computationally and give results quite similar to the adaptive Gaussian approximation. We used five abscissas for the adaptive Gaussian quadrature, as this gives roughly the same precision as the ten-abscissa quadrature rule.

The alternating approximation gives results very similar to the adaptive Gaussian quadrature. As in the orange trees example, the profile traces of the variance-covariance components and the fixed effects meet almost perpendicularly, indicating a local uncorrelation between these estimates. The Laplacian and the adaptive Gaussian approximations give virtually identical plots (not included here). This suggests there is little to be gained by increasing the number of abscissas past one in the quadrature rule. The major gain in precision is obtained by centering the grid at the conditional modes and scaling it using the approximate Hessian.

3.3 Simulation Results

In this section we include a comparison of the approximations to the loglikelihood in model (1) using simulation. We restrict ourselves to the alternating, the Laplacian, and the (five-abscissa) adaptive Gaussian approximations as these seem to be more accurate and/or more efficient than the Gaussian and the importance sampling approximations. Two models were used in the simulation analysis: a

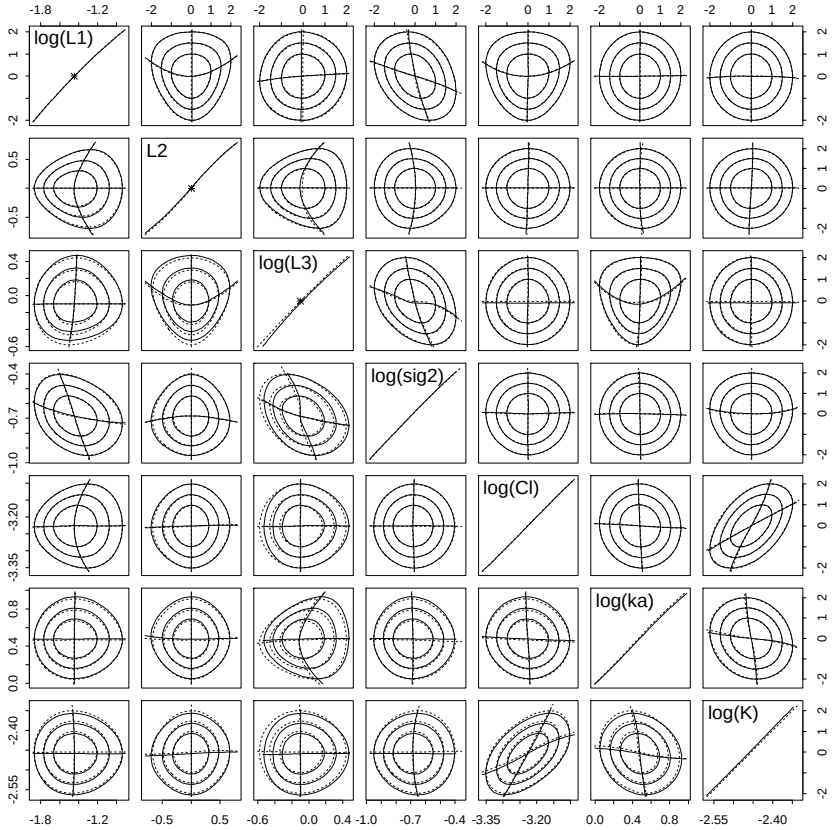

Figure 4: Profile traces and profile contour plots for the Theophylline data based on the adaptive Gaussian approximation with 5 abscissas (solid line) and the alternating approximation (dashed line). Plots below the diagonal are in the original scale and plots above the diagonal are in the zeta scale (Bates and Watts, 1988). Interpolated contours correspond approximately to joint confidence levels of 68%, 87%, and 95%.

logistic model similar to the one used for the orange trees data and a first order open compartment model similar to the one used for the Theophylline example. For both models 1000 samples were generated and maximum likelihood (ML) estimates based on the different approximations obtained. For the alternating approximation, restricted maximum likelihood (RML) estimates were also obtained.

3.3.1 Logistic Model

A logistic model similar to (13), but with two random effects instead of one, was used to generate the data. The model is given by

$$y_{ij} = \frac{\beta_1 + b_{i1}}{1 + \exp\{-[t_{ij} - (\beta_2 + b_{i2})]/\beta_3\}} + \varepsilon_{ij},\ i = 1, \ldots, M,\ j = 1, \ldots, n_i \qquad (16)$$

where the $\boldsymbol{b}_i$ are *i.i.d.* $\mathcal{N}(0, \sigma^2 D)$, and the ε_{ij} are *i.i.d.* $\mathcal{N}(0, \sigma^2)$ and independent of the $\boldsymbol{b}_i$. We used $M = 15, n_i = 10,\ i, \ldots, 15, \sigma^2 = 25, \boldsymbol{\beta} = (200, 700, 350)^T$, and $\boldsymbol{D} = \begin{bmatrix} 4 & -2 \\ -2 & 25 \end{bmatrix}$.

Table 5 summarizes the simulation results for the variance-covariance components (*MSE* denotes the mean square error of the estimators). The different approximations to the loglikelihood give similar simulation results for all the parameters involved. The cluster specific variance (σ^2) is estimated with more relative precision than the elements of the scaled variance-covariance matrix of the random effects ($\boldsymbol{D}$). This is probably because the precision of the estimate of σ^2 (as well as the estimates of $\boldsymbol{\beta}$) is determined by the total number of observations, while the precision of the estimates of $\boldsymbol{D}$ is determined by the number of clusters. We can also see a tendency for the restricted maximum likelihood to give positively biased estimates of D_{11} and D_{22}, while the other approximations give negatively biased estimates. The rationale for restricted maximum likelihood is to reduce bias in estimating variance components. It does not seem to do so in this case; it just changes its direction.

Table 5: Simulation results for the variance-covariance components in the logistic model

	D_{11}			D_{12}		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	4.2000	0.2000	3.9161	-1.9460	0.0540	18.4208
Alternating – ML	3.9218	-0.0782	3.4370	-1.9947	0.0053	16.1845
Laplacian	3.9349	-0.0651	3.3748	-1.9781	0.0219	15.7242
Adap. Gaussian	3.9408	-0.0592	3.4081	-1.9651	0.0349	15.7540
	D_{22}			σ^2		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	26.0890	1.0890	360.9847	24.8849	-0.1151	9.7557
Alternating – ML	23.3216	-1.6784	314.5025	24.6511	-0.3489	9.6473
Laplacian	23.8638	-1.1362	310.0535	24.6252	-0.3748	9.5700
Adap. Gaussian	23.9337	-1.0662	312.4221	24.6168	-0.3832	9.5671

Figure 5 presents the scatter plots of the variance-covariance component (σ^2 and $\boldsymbol{D}$) estimates for

the alternating RML, the alternating ML, and the Laplacian approximations versus the adaptive Gaussian approximation. We see that, except for the alternating RML approximation, all methods lead to very similar estimates. In general the alternating RML approximation gives larger values for the estimates of the variance components (especially D_{11} and D_{22}) than the other methods. The higher mean square error for D_{12} from the alternating ML and RML methods is visible in the plot, as each of the panels comparing these estimates to those from the adaptive Gaussian method has a vertical clump of points at the true value.

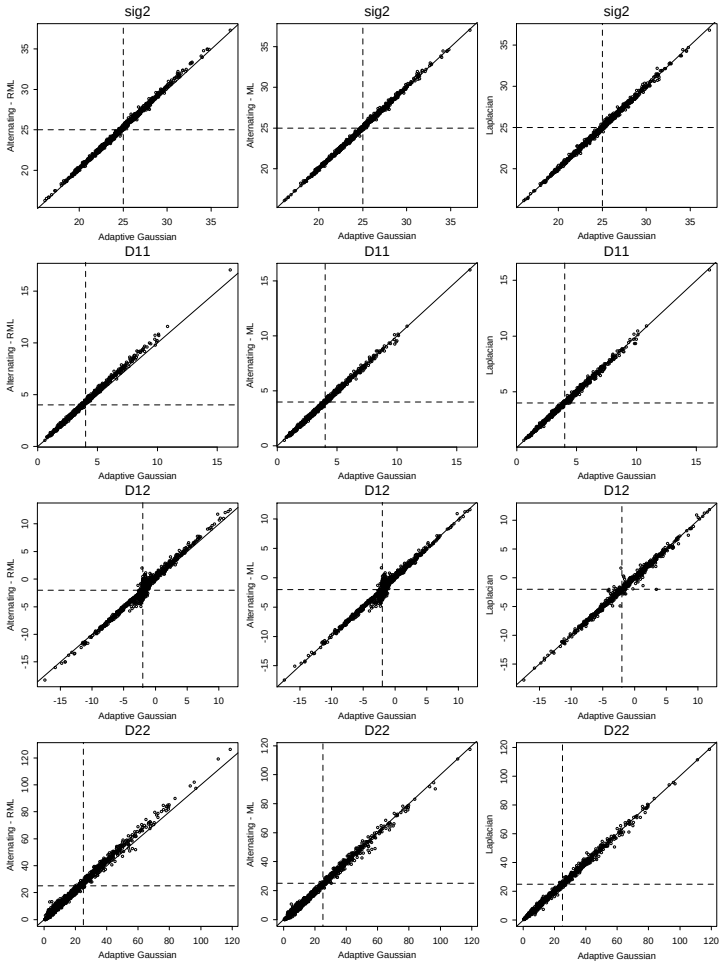

Figure 5: Scatter plots of variance-covariance components estimates for the alternating (RML and ML), Laplacian, and adaptive Gaussian approximations in the logistic model (16). The dashed lines indicate the true values of the parameters.

Table 6 presents the simulation results for the fixed effects estimates. The results are very similar for all approximations considered. We also note that the relative variability of the fixed effects estimates

is much smaller than those of the estimates of the elements of $\boldsymbol{D}$. There is very little, if any, bias in the fixed effects estimates.

Table 6: Simulation results for the fixed effects in the logistic model

	β_1			β_2			β_3		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	199.6097	-0.3903	10.1830	698.4286	-1.5714	138.2153	348.8091	-1.1909	57.1686
Alternating – ML	199.6081	-0.3919	10.1836	698.4292	-1.5707	138.2237	348.8224	-1.1776	57.1297
Laplacian	199.9275	-0.0725	10.2012	700.0317	0.0317	138.3760	350.2019	0.2019	56.9361
Adap. Gaussian	199.9229	-0.0771	10.1561	699.9082	-0.0918	138.4409	350.0640	0.0640	57.0550

Figure 6 presents the scatter plots of the fixed effects estimates for the alternating RML, alternating ML, and Laplacian approximations versus the adaptive Gaussian approximation. Again we observe a strong agreement in the estimates obtained through the various approximations. The alternating approximations tend to give estimates slightly smaller than the Laplacian and adaptive Gaussian, but the differences are minor.

3.3.2 First Order Compartment Model

The model used in the simulation is identical to (15). As in the Theophylline example we set $M = 12$ and $n_i = 11,\ i = 1, \ldots, 12$. The parameter values used were $\sigma^2 = 0.25$, $\boldsymbol{\beta} = (-3.0, 0.5, -2.5)^T$, and $\boldsymbol{D} = \begin{bmatrix} 0.2 & 0 \\ 0 & 1 \end{bmatrix}$.

Table 7 summarizes the simulation results for the variance-covariance components estimates. As in the logistic model analysis, we observe that the elements of $\boldsymbol{D}$ are estimated with less relative precision than σ^2. The alternating ML, Laplacian, and adaptive Gaussian approximations seem to lead to slightly downward biased estimates of D_{11} and D_{22}, while the alternating RML approximation appears to give unbiased estimates (thus achieving its main purpose). Note however that the unbiasedeness of the RML estimates does not translate into smaller mean square error – all four estimation methods lead to similar MSE, for all parameters.

Figure 7 presents the scatter plots of the variance-covariance estimates for the alternating RML, alternating ML, and Laplacian approximations versus the adaptive Gaussian approximation. The alternating RML approximation tends to give larger values for D_{11} and D_{22}, and larger absolute values for D_{12}, while the remaining approximations lead to very similar estimates. There was one sample for which the alternating approximations apparently converged to a different solution than the Laplacian and

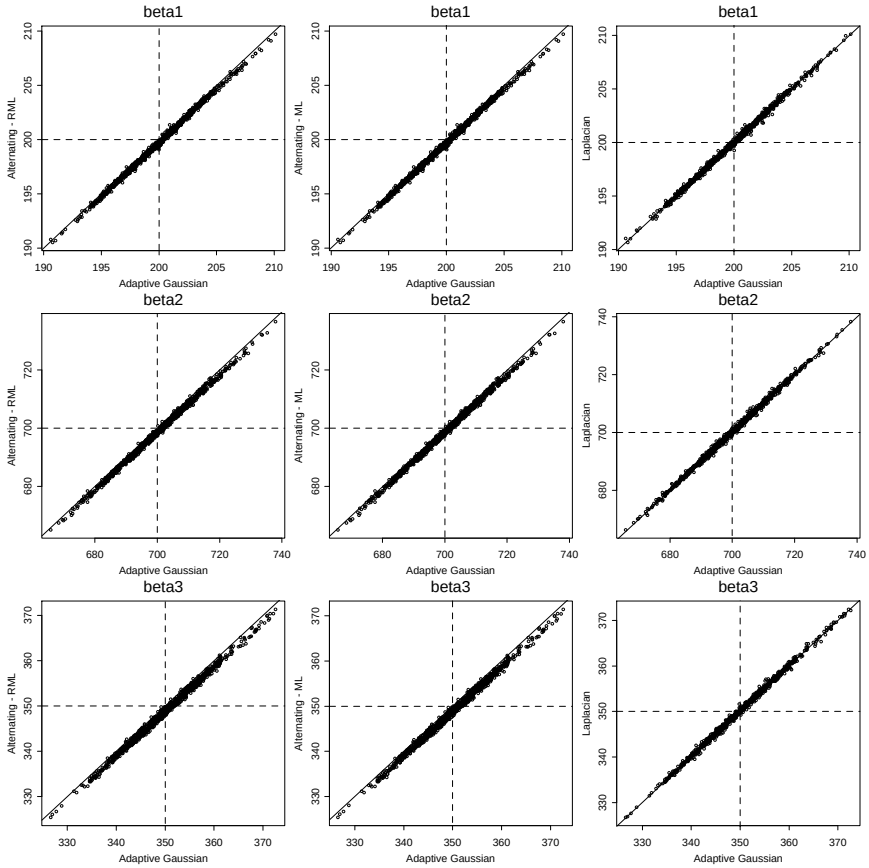

Figure 6: Scatter plots of fixed effects estimates for the alternating (RML and ML), Laplacian, and adaptive Gaussian approximations in the logistic model (16). The dashed lines indicate the true values of the parameters.

adaptive Gaussian. Overall there were no major differences between the approximations in estimating the variance-covariance components.

Table 6 gives the simulation results for the fixed effects estimates. All four approximations give virtually identical results for the estimation of the fixed effects. They all show very little bias and smaller relative variability when compared to the estimates of the variance-covariance components.

The scatter plots of the fixed effects estimates, not included here, show practically identical results for the alternating RML and ML, the Laplacian , and the adaptive Gaussian approximations.

Table 7: Simulation results for the variance-covariance components in the first order compartment model

	D_{11}			D_{12}		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	0.1996	-0.0004	0.0089	-0.0013	-0.0013	0.0210
Alternating – ML	0.1840	-0.0160	0.0078	-0.0023	-0.0023	0.0179
Laplacian	0.1862	-0.0138	0.0078	-0.0011	-0.0011	0.0178
Adap. Gaussian	0.1860	-0.0140	0.0077	0.0002	0.0002	0.0180
	D_{22}			σ^2		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	1.0095	0.0095	0.2565	0.2508	0.0008	0.0012
Alternating – ML	0.9249	-0.0751	0.2240	0.2486	-0.0014	0.0011
Laplacian	0.9388	-0.0612	0.2276	0.2480	-0.0020	0.0011
Adap. Gaussian	0.9476	-0.0524	0.2332	0.2481	-0.0019	0.0011

Table 8: Simulation results for the fixed effects in the first order compartment model

	β_1			β_2			β_3		
Approximation	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
Alternating – RML	-2.9989	0.0011	0.0053	0.4876	-0.0124	0.0244	-2.4965	0.0035	0.0020
Alternating – ML	-2.9992	0.0008	0.0053	0.4869	-0.0131	0.0244	-2.4965	0.0035	0.0020
Laplacian	-3.0009	-0.0009	0.0053	0.4983	-0.0017	0.0242	-2.5045	-0.0045	0.0020
Adap. Gaussian	-2.9987	0.0013	0.0053	0.4984	-0.0016	0.0246	-2.5008	-0.0008	0.0020

4 Conclusions

The results of section 3 indicate that the alternating approximation (4) to the loglikelihood function in the nonlinear mixed effects model (1) proposed by Lindstrom and Bates (1990) gives accurate and reliable estimation results. The main advantages of this approximation are its computational efficiency (allowing the use of linear mixed effects techniques to estimate the scaled variance-covariance matrix of the random effects $\boldsymbol{D}$) and the availability of a restricted likelihood version of it, which is not yet defined for other approximations/estimation methods. With regard to the restricted maximum likelihood estimation though, the results of section 3 suggest that the bias correction ability of this method depends on the nonlinear model that is being considered: RML estimation achieved its purpose for the first order compartment model (15), but it *increased* the bias in the logistic model (16). More research is needed in this area. Since it is simpler computationally the alternating approximation should be used to provide starting values for the more accurate approximations (e.g. Laplacian and adaptive Gaussian) if they are preferred.

The Gaussian quadrature approximation (11) only seems to give accurate results for large number of

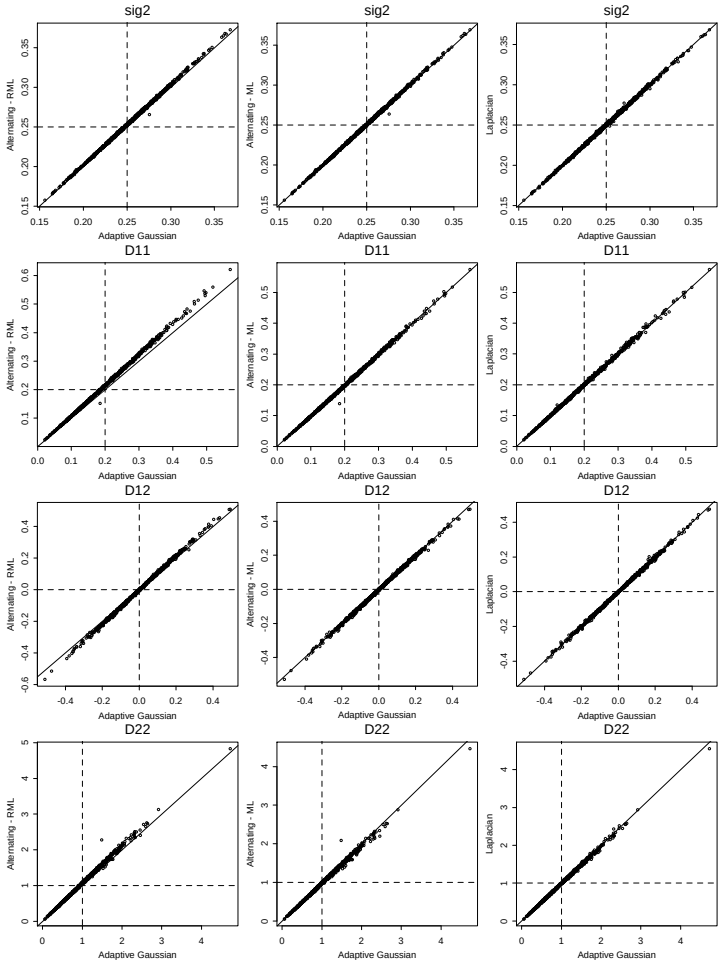

Figure 7: Scatter plots of variance-covariance components estimates for the alternating (RML and ML), Laplacian, and adaptive Gaussian approximations in the first order compartment model (15). The dashed lines indicate the true values of the parameters.

abscissa (> 100), what makes it very inefficient computationally. The basic problem is that it centers the grid of abscissas at $\mathbf{0}$ (the expected value of the random effects) and scales it according to $\boldsymbol{D}$, while the highest values of the integrand in (2) are concentrated around the posterior modes of the random effects ($\widehat{\boldsymbol{b}}$) and scaled according to $g''\left(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{y}, \widehat{\boldsymbol{b}}\right)$. The advantages of this approximation are that it does not require the estimation of the posterior modes of the random effects at each iteration and it admits closed form partial derivatives with respect to the parameters of interest ($\boldsymbol{\beta}$, $\boldsymbol{D}$, and σ^2), provided these are available for the model function $\boldsymbol{f}$ (Davidian and Gallant, 1992). We feel that these advantages do not compensate for the inaccuracy or computational inefficiency of the Gaussian approximation.

The importance sampling approximation (9) gives reliable estimation results, comparable to those of

the adaptive Gaussian and Laplacian approximations, but is considerably less efficient computationally than these approximations. Also, the stochastic variability associated with the different importance samples may overwhelm the numerical variability of the loglikelihood for small changes in the parameter values, making it difficult to calculate numerical derivatives. The main advantage of the importance sampling approximation is its versatility in handling distributions other than the normal, for both the random effects and the cluster-specific error term (ϵ). For example it would be rather straightforward to adapt the importance sampling integration to handle a multivariate t distribution for the random effects, but that would not be a trivial task for either the alternating, the Laplacian, or the adaptive Gaussian approximations. Wakefield et al. (1994) use the similar property of Gibbs sampler methods to check for outliers in nonlinear mixed effects models. If one is willing to stick with the normal distribution for $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ in the nonlinear mixed effects model (1) then the importance sampling approximation is not the most efficient choice.

Of all approximations considered here, the Laplacian and adaptive Gaussian approximations probably give the best mix of efficiency and accuracy. The former can be regarded as a particular case of the latter, where just one abscissa is used. Both approximations (and the importance sampling approximation as well) give the exact loglikelihood when the model function $\boldsymbol{f}$ in (1) is a linear function of the random effects. In the examples that we analyzed not much was gained by going from a one-point adaptive Gaussian quadrature (Laplacian) approximation to approximations with a larger number of abscissas. It appears that the major gain in adaptive Gaussian approximations is related to the centering and scaling of the abscissas. Increasing the number of points in the evaluation grid only gives marginal improvement. The Laplacian approximation has the additional advantage over the adaptive Gaussian approximation with more than one abscissa of allowing profiling of the loglikelihood over σ^2, thus reducing the dimensionality of the optimization problem.

For statistical analysis purpose we would recommend using a hybrid scheme in which the alternating algorithm would be used to get *good* initial values for the more refined Laplacian approximation to the loglikelihood of model (1). This way the computational efficiency of the alternating algorithm would be combined with the greater accuracy of the Laplacian approximation.

Acknowledgment

This research was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil and NSF Grant DMS-9309101.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and mathematical Tables*, Dover, New York.

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity, *Journal of the Royal Statistical Society, Ser. B* **42**: 1–25.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

Chambers, J. M. and Hastie, T. J. (eds) (1992). *Statistical Models in S*, Wadsworth, Belmont, CA.

Davidian, M. and Gallant, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine, *Journal of Pharmacokinetics and Biopharmaceutics* **20**: 529–556.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration*, second edn, Academic Press, New York.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edn, Wiley, New York.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration, *Econometrica* **57**: 1317–1339.

Golub, G. H. (1973). Some modified matrix eigenvalue problems, *SIAM Review* **15**: 318–334.

Golub, G. H. and Welsch, J. H. (1969). Calculation of Gaussian quadrature rules, *Math. Comp.* **23**: 221–230.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**: 383–385.

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

Leonard, T., Hsu, J. S. J. and Tsui, K. W. (1989). Bayesian marginal inference, *Journal of the American Statistical Association* **84**: 1051–1058.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association* **83**: 1014–1022.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**: 673–687.

Mallet, A., Mentre, F., Steimer, J.-L. and Lokiek, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with applications to Cyclosporine, *J. Pharmacokin. Biopharm.* **16**: 311–327.

Pinheiro, J. C. and Bates, D. M. (1993). Asymptotic properties of maximum lieklihood estimates in the general linear mixed effects model. Submitted to *Annals of Statistics*.

Sheiner, L. B. and Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. michaelis-menten model: Routine clinical pharmacokinetic data, *Journal of Pharmacokinetics and Biopharmaceutics* **8**(6): 553–571.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures, *Biometrics* **48**: 1–18.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler, *Applied Statistics*. Accepted for publication.