Biometrika (1996), **83**, 2, *pp.* 447–452
Printed in Great Britain

A note on the use of Laplace's approximation for nonlinear mixed-effects models

BY EDWARD F. VONESH

Applied Statistics Center, Baxter Healthcare Corporation, P.O. Box 490, Round Lake, Illinois 60073, U.S.A.

SUMMARY

The asymptotic properties of estimates obtained using Laplace's approximation for nonlinear mixed-effects models are investigated. Unlike the restricted maximum likelihood approach, e.g. Wolfinger (1993), here the Laplace approximation is applied only to the random effects of the integrated likelihood. This results in approximate maximum likelihood estimation. The resulting estimates are shown to be consistent with the rate of convergence depending on both the number of individuals and the number of observations per individual. Conditions under which the leading term Laplace approximation should be avoided are discussed.

Some key words: First-order method; Maximum likelihood; Nonlinear random effects.

1. INTRODUCTION

The model considered here is the Gaussian-based two-stage nonlinear mixed-effects model

$$y_i = f(X_i, \beta_i) + \varepsilon_i, \quad \beta_i = A_i\beta + B_i b_i, \tag{1}$$

where y_i is a p_i-vector of observations on individual i ($1 \leqslant i \leqslant n$), β_i is an r-dimensional vector of regression coefficients, X_i, A_i and B_i are $p_i \times t$, $r \times s$ and $r \times v$ known design matrices, β is an s-dimensional vector of fixed-effects, b_i is a v-dimensional vector of inter-individual random effects and ε_i is a p_i-dimensional vector of intra-individual random errors. Assume b_i is Gaussian with mean 0 and variance matrix Ψ, and ε_i is Gaussian with mean 0 and variance matrix $\sigma^2\Lambda_i(\gamma)$, independent of b_i. The intra-individual covariance structure depends on a g-dimensional vector γ and is dependent on i only through its dimension p_i.

This model has been considered in various forms by a number of authors including Sheiner & Beal (1980), Lindstrom & Bates (1990) and Vonesh & Carter (1992). It is useful in a number of applications involving the analysis of repeated measurements, particularly in the fields of pharmacokinetics, biological growth and epidemiology. The difficulty with this model is that the random effects may enter nonlinearly. Consequently, there is no closed form solution for the marginal distribution of y_i. While numerical integration and Monte Carlo methods have been employed, a more common approach has been to linearise the model with respect to the random effects. In particular, Sheiner & Beal (1980, 1985) and Vonesh & Carter (1992) use a first-order population-averaged approximation to the marginal distribution of y_i by expanding f about the average random effect, $b_i = 0$. Solomon & Cox (1992) also expand f about $b_i = 0$ but they include four terms in the expansion. These methods will provide a reasonable approximation to the likelihood whenever the inter-individual variability is small (Solomon & Cox, 1992). The first-order population-averaged approach is also appropriate when used primarily to approximate the unknown marginal covariance of y_i assuming the marginal means are correctly specified (Breslow & Clayton, 1993).

Lindstrom & Bates (1990) attempt to improve on the first-order population-averaged approximation by expanding f about $\hat{b}_i$, the posterior mode, and using the Gaussian posterior approxi-

mation of Stiratelli, Laird & Ware (1984) to estimate Ψ and $\sigma^2\Lambda_i(\gamma)$. This results in a first-order conditional or subject-specific approximation to the marginal distribution of y_i. They present a two-step algorithm for maximum likelihood and restricted maximum likelihood estimation. Wolfinger (1993) demonstrates how the restricted maximum likelihood version of Lindstrom & Bates' two-step algorithm can be derived using the Laplace approximation

$$\int e^{ml(\tau)}\,d\tau \doteq (2\pi/m)^{q/2}|-l''(\hat{\tau})|^{-\frac{1}{2}}e^{ml(\hat{\tau})}, \tag{2}$$

where τ is a q-dimensional parameter vector and $\hat{\tau}$ maximises $ml(\tau)$. For fixed $q < \infty$, the order of accuracy of the Laplace approximation is $O(1/m)$. It is shown here that, in the repeated measurements setting where the number of random effects coincides with the number of individuals, the order of accuracy depends both on the number of individuals and on the number of observations per individual. Section 2 shows that joint maximum likelihood estimation can be carried out using extended least squares and involves solving a set of second-order generalised estimating equations. Section 3 provides a somewhat heuristic argument showing that the resulting maximum likelihood estimates are consistent to order $O_p[\max\{n^{-\frac{1}{2}}, \min(p_i)^{-1}\}]$.

2. MAXIMUM LIKELIHOOD ESTIMATION USING THE LAPLACE APPROXIMATION

Let $p(y_i|b_i)$ and $p(b_i)$ denote the respective normal density functions associated with the distributions of $y_i|b_i$ and b_i, respectively. Let $\theta^{\mathrm{T}} = (\mathrm{vech}(\Psi)^{\mathrm{T}}, \sigma^2, \gamma^{\mathrm{T}})$ denote the vector of distinct variance-covariance parameters associated with Ψ and $\sigma^2\Lambda_i(\gamma)$. For fixed β and θ, the marginal distribution of y_i is given by

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

Set $p_i l(b_i) = \log\{p(y_i|b_i)p(b_i)\}$ and let $\hat{b}_i$ maximise $p_i l(b_i)$ for fixed β and θ. By using the leading term Laplace approximation (2) along with arguments similar to those presented by Wolfinger (1993), it can be shown that the ith individual's contribution to minus twice the overall log-likelihood function is approximately

$$-2\log p(y_i) \doteq \log|V_i| + \{w_i - f(X_i, A_i\beta + B_i\hat{b}_i)\}^{\mathrm{T}}V_i^{-1}\{w_i - f(X_i, A_i\beta + B_i\hat{b}_i)\} + p_i\log(2\pi),$$

where

$$w_i = y_i + \tilde{Z}_i\hat{b}_i, \quad \tilde{Z}_i = (\partial f_i/\partial b_i^{\mathrm{T}})|_{\beta,\hat{b}_i}, \quad V_i = V_i(\beta,\theta) := \tilde{Z}_i\Psi\tilde{Z}_i^{\mathrm{T}} + \sigma^2\Lambda_i(\gamma).$$

This approximation is described by Beal & Sheiner (1992) in connection with a first-order conditional estimation approach, or conditional extended least squares, which they implement in their NONMEM software program.

Under this form of Laplace's approximation, the y_i will be approximately normally distributed with mean $\mu_i(\beta) = f(X_i, A_i\beta + B_i\hat{b}_i) - \tilde{Z}_i\hat{b}_i$ and variance matrix, $V_i(\beta,\theta)$. Two things about this version of Laplace's approximation are worth noting. First, the Hessian $l''(\hat{b}_i)$ is replaced by the usual first-order approximation $-E\{l''(\hat{b}_i)\}$: see, for example, Wolfinger (1993). Consequently, the order of accuracy associated with Laplace's formula is approximately $O(1/p_i)$. Secondly, since $V_i(\beta,\theta)$ depends on β through $\tilde{Z}_i$, joint maximum likelihood estimates of β and θ are obtained by jointly minimising the conditional extended least squares objective function:

$$Q(\beta,\theta|\hat{b}_1,\ldots,\hat{b}_n) := \sum_{i=1}^{n}[\{y_i - \mu_i(\beta)\}^{\mathrm{T}}V_i(\beta,\theta)^{-1}\{y_i - \mu_i(\beta)\} + \log|V_i(\beta,\theta)|]. \tag{3}$$

Actual estimation is carried out by solving the set of joint estimating equations:

$$U(\beta,\theta) := \sum_{i=1}^{n}\begin{pmatrix} D_i(\beta) & 0 \\ E_i(\beta) & E_i(\theta)\end{pmatrix}^{\mathrm{T}}\begin{pmatrix} V_i(\beta,\theta) & 0 \\ 0 & W_i(\beta,\theta)\end{pmatrix}^{-1}\begin{pmatrix} y_i - \mu_i(\beta) \\ s_i(\beta) - v_i(\beta,\theta)\end{pmatrix} = 0, \tag{4}$$

where

$$D_i(\beta)=\frac{\partial \mu_i(\beta)}{\partial \beta^{\mathrm{T}}}, \quad E_i(\beta)=\frac{\partial v_i(\beta,\theta)}{\partial \beta^{\mathrm{T}}}, \quad E_i(\theta)=\frac{\partial v_i(\beta,\theta)}{\partial \theta^{\mathrm{T}}}, \quad v_i(\beta,\theta)=\mathrm{vec}\{V_i(\beta,\theta)\},$$

$$W_i(\beta,\theta)=2V_i(\beta,\theta)\otimes V_i(\beta,\theta), \quad s_i(\beta)=\mathrm{vec}[\{y_i-\mu_i(\beta)\}\{y_i-\mu_i(\beta)\}^{\mathrm{T}}].$$

These correspond to the second-order generalised estimating equations presented by Prentice & Zhao (1991) for normally distributed data. Given new estimates $\hat{\beta}$ and $\hat{\theta}$, the $\hat{b}_i$ can be updated by maximising $\log\{p(y_i|b_i)p(b_i)\}$ with respect to b_i. As suggested by a referee, this can be done jointly across all i using a modified version of the pseudo-data step of Lindstrom & Bates' algorithm in which β is held fixed at its current estimate. The fitting algorithm thus consists of iterating back and forth between (i) updating the $\hat{b}_i$ for fixed $\hat{\beta}$ and $\hat{\theta}$ using a modified pseudo-data step, and (ii) solving the second-order generalised estimating equations (4) holding the $\hat{b}_i$ fixed at their current values. This is apparently the strategy behind the first-order conditional estimation method described by Beal & Sheiner (1992).

The maximum likelihood estimates obtained here using the Laplace approximation differ from those obtained using the maximum likelihood version of Lindstrom & Bates' two-step algorithm. In particular, it can be shown that, under the linear-mixed-effects step of their algorithm, the resulting maximum likelihood estimates correspond to solving the joint generalised estimating equations (4) but with $\partial \tilde{Z}_i \hat{b}_i/\partial \beta^{\mathrm{T}}$ and $\partial\,\mathrm{vec}\{V_i(\beta,\theta)\}/\partial\beta^{\mathrm{T}}$ both set equal to 0. Thus the estimates computed using the maximum likelihood version of Lindstrom & Bates' two-step algorithm correspond to approximate conditional maximum likelihood estimates with β estimated via iteratively reweighted least squares and θ estimated via conditional maximum likelihood.

In the preceding arguments, the application of the Laplace approximation assumes the intrasubject covariance matrix is of the form $\sigma^2\Lambda_i(\gamma)$. If, in fact, Λ_i has the more general structure $\Lambda_i(\beta_i,\gamma)$ as is often found in pharmacokinetics or applications involving generalised linear mixed models, then the application of the Laplace approximation would require minimising, with respect to b_i, the more complicated extended least squares objective function

$$[\sigma^{-2}\{y_i-f_i(\beta_i)\}^{\mathrm{T}}\Lambda_i(\beta_i,\gamma)^{-1}\{y_i-f(\beta_i)\}+b_i^{\mathrm{T}}\Psi^{-1}b_i+\log|\sigma^2\Lambda_i(\beta_i,\gamma)|].$$

Computational difficulties arise from the fact that both β_i and Λ_i depend on b_i. One way to avoid this problem is to assume Λ_i to be a function of β and γ only. Alternatively, Breslow & Clayton (1992) and Wolfinger (1993) avoid this problem by employing a pseudo-likelihood approach in which they effectively set $\partial\,\mathrm{vec}\{\Lambda_i(\beta_i,\gamma)\}/\partial b_i^{\mathrm{T}}=0$.

3. Asymptotic properties

Approximating the marginal distribution by expanding about the current random effects rather than the average should improve the overall estimation of β and θ. Indeed such improvement was observed in a Monte Carlo study reported by Vonesh (1992). However, both n and p_i need to be sufficiently large if this approximation is to work. To see this, consider the following heuristic argument; details of the proof are given in the Appendix. Let $\theta=(\psi,\sigma^2,\gamma)$ be known and suppose the ith individual's function $f(X_i,\beta_i)$ depends on a single parameter $\beta_i=\beta+b_i$ such that $s=v=1$. Let $\hat{b}_i$ minimise

$$\sigma^{-2}\{y_i-f_i(\beta_i)\}^{\mathrm{T}}\Lambda_i(\gamma)^{-1}\{y_i-f_i(\beta_i)\}+b_i^2/\psi$$

for fixed β and θ. Then up to a constant, the unspecified marginal log-likelihood, $l(\beta)$, may be approximated as

$$l(\beta)\simeq l^*(\beta)+O\{n\min(p_i)^{-1}\}, \tag{5}$$

where $l^*(\beta)=-\frac{1}{2}Q(\beta,\theta|\hat{b}_1,\ldots,\hat{b}_n)$. Let $\hat{\beta}$ be the Laplace-based maximum likelihood estimate obtained by minimising $-2l^*(\beta)$. Under suitable regularity conditions on $l(\beta)$ and $f(X_i,\beta_i)$, it is

shown in the Appendix that

$$(\hat{\beta} - \beta) = O_p[\max\{n^{-\frac{1}{2}}, \min(p_i)^{-1}\}]. \quad (6)$$

Thus, the approximate maximum likelihood estimate $\hat{\beta}$ will be consistent only as both n and $\min(p_i) \to \infty$. Intuitively, the $n^{-\frac{1}{2}}$ term comes from standard asymptotic theory while the $\min(p_i)^{-1}$ term comes from the Laplace approximation. This result easily extends to the more general case where both β and θ are unknown: see the Appendix for details. The requirement that $\min(p_i) \to \infty$ can be relaxed provided the random effects enter the model in a strictly linear fashion. This is because the Laplace approximation will be exact when the random effects are strictly linear.

It is interesting to note that the accuracy of the leading term Laplace approximation to the log-likelihood function is $O\{n/\min(p_i)\}$ or, equivalently, $o(1)$ provided $\min(p_i)$ grows faster than n. In this case, $(\hat{\beta} - \beta) = O_p(n^{-\frac{1}{2}})$ with $\hat{\beta}$ being asymptotically equivalent to the unconditional maximum likelihood estimate: see the Appendix. This reflects the fact that, as the accuracy of the Laplace approximation to the marginal log-likelihood increases, $\hat{\beta}$ will behave more and more like the actual maximum likelihood estimate. When the minimum number of observations on individuals, $\min(p_i)$, grows at a rate slower than the number of individuals, n, then the Laplace approximation will grow worse but we still attain consistency. In particular, as $\min(p_i)$ grows at a rate greater than $n^{\frac{1}{2}}$ but less than or equal to n, the rate of consistency will still be $O_p(n^{-\frac{1}{2}})$ but the resulting estimate will no longer be asymptotically equivalent to the maximum likelihood estimate since $l^*(\beta)$ will no longer converge to the true log-likelihood, $l(\beta)$. It is only in the case where $\min(p_i)$ grows at a rate slower than $n^{\frac{1}{2}}$ that the rate of consistency will be adversely affected.

4. DISCUSSION

A closer examination of the first-order subject-specific expansion supports the preceding results. Let $\min(p_i) \to \infty$ so that $\hat{b}_i \to b_i$ for all i. Then the first-order subject-specific model, $y_i|b_i = \mu_i(\beta) + \hat{Z}_i b_i + \varepsilon_i$, will converge to the true conditional model given in the first stage of (1). In that case, for known b_i, we can obtain $n^{\frac{1}{2}}$-consistent estimates of β, σ^2 and γ based directly on the first-stage model. Thus the first-order subject-specific Laplace approximation is conceptually very similar to the various two-stage methods described in the literature (Racine-Poon, 1985; Davidian & Giltinan, 1993). The requirement that $\min(p_i) \to \infty$ is consistent with the fact that we are trying to approximate the marginal distribution of each of the y_i's. In fact, it is only for large p_i that we would expect improved estimation under the Laplace approximation when compared to the first-order population-averaged approximation of Sheiner & Beal (1980, 1985). The reason, as indicated by Beal & Sheiner (1992), is that the posterior mode, $\hat{b}_i$, will effectively shrink towards 0 for subjects with sparse data. Thus, as the number of data per subject decreases, the difference in estimates obtained using the Laplace approximation versus the first-order population-averaged approximation should decrease. Simulation studies along these lines are needed.

The approach taken here differs from the restricted likelihood approach described by Wolfinger (1993) in that the approximate likelihood does not resemble a modified profile likelihood (Barndorff-Nielsen, 1983) or approximate conditional likelihood (Cox & Reid, 1987) wherein β is regarded as a nuisance parameter. Rather, the approach used here is to apply the Laplace approximation holding both β and θ fixed. This results in an approximate joint log-likelihood function whereby the marginal mean and variance matrix of y_i are no longer orthogonal to one another. While this entails greater computational difficulties in terms of estimation, it may yield a more efficient estimate of β as a result of incorporating information about β from the covariance structure. However, care should be taken when using this approach as it can lead to inconsistent estimates under model misspecification (Vonesh, 1992).

Finally, improved accuracy of the integrated likelihood is possible by taking the first three terms of the Laplace approximation. This may prove particularly useful in applications with moderately sparse data.

Appendix

Consistency of Laplace-based maximum likelihood estimates

Consistency of the Laplace-based maximum likelihood estimate is investigated assuming model (1). For simplicity, we prove the rate of consistency (6) obtains under the following conditions:

(i) $\theta = (\psi, \sigma^2, \gamma)$ is known;
(ii) $f(X_i, \beta_i)$ depends on a single parameter $\beta_i = \beta + b_i$ such that $s = v = 1$;
(iii) $b_i \sim N(0, \psi)$, $\varepsilon_i \sim N\{0, \sigma^2 \Lambda_i(\gamma)\}$ and b_i and ε_i are independent of one another;
(iv) $\Lambda_i(\gamma)$ is independent of β_i;
(v) if we let $l(\beta)$ be the true but unspecified marginal log-likelihood function, then $l(\beta)$ and $f(X_i, \beta)$ satisfy the usual regularity conditions.

Let $\hat{b}_i$ minimise

$$\{y_i - f_i(\beta_i)\}^{\mathrm{T}} \Lambda_i(\gamma)^{-1} \{y_i - f_i(\beta_i)\} + b_i^2/\psi$$

for fixed β and θ. Then up to a constant, the ith individual's contribution to the overall log-likelihood may be approximated as

$$\log p(y_i) \doteq -\tfrac{1}{2} Q_i(\beta, \theta | \hat{b}_i) + O(1/p_i),$$

where $Q_i(\beta, \theta | \hat{b}_i)$ is the ith summand in the extended least squares objective function (3). Hence, up to a constant, the log-likelihood with respect to β can be written as

$$l(\beta) \doteq l^*(\beta) + O\{n \min(p_i)^{-1}\},$$

where $l^*(\beta) = -\tfrac{1}{2} Q(\beta, \theta | \hat{b}_1, \ldots, \hat{b}_n)$. Let $U^*(\beta) = \partial l^*(\beta)/\partial \beta$ and let $\hat{\beta}$ be the Laplace-based maximum likelihood estimate satisfying $U^*(\hat{\beta}) = 0$. Under suitable regularity conditions on $l(\beta)$ and assuming $\hat{\beta}$ is an interior point in a neighbourhood containing β, a Taylor series expansion about β yields

$$n^{-1} U(\hat{\beta}) = n^{-1} U(\beta) + n^{-1} H(\beta)(\hat{\beta} - \beta) + O_p(1)(\hat{\beta} - \beta)^2,$$

where $U(\beta) = \partial l(\beta)/\partial \beta$ and $H(\beta) = \partial^2 l(\beta)/\partial \beta^2$ are the first and second order derivatives of the true but unknown marginal log-likelihood $l(\beta)$. Now, given the above assumptions, it can be shown that

$$n^{-1} H(\beta)(\hat{\beta} - \beta) + O_p(1)(\hat{\beta} - \beta)^2 = O_p(1)(\hat{\beta} - \beta).$$

Furthermore, given sufficient regularity conditions on $l(\beta)$, we know that $n^{-1} U(\beta) = O_p(n^{-\frac{1}{2}})$. Also, given suitable regularity conditions on $f(X_i, \beta_i)$, for example that fifth order derivatives exist and are continuous in an open neighbourhood about β, we have

$$n^{-1} U(\hat{\beta}) = n^{-1} U^*(\hat{\beta}) + O\{\min(p_i)^{-1}\}.$$

Combining these results, it follows that

$$\begin{aligned}(\hat{\beta} - \beta) &= \{n^{-1} U(\hat{\beta}) - n^{-1} U(\beta)\}/O_p(1) = n^{-1} U^*(\hat{\beta}) + O_p\{\min(p_i)^{-1}\} + O_p(n^{-\frac{1}{2}}) \\ &= O_p[\max\{n^{-\frac{1}{2}}, \min(p_i)^{-1}\}].\end{aligned}$$

This result extends to the more general case with β and θ unknown by replacing the scalar β with the vector $\tau = (\beta^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$ and using the multivariate version of Taylor's theorem in the expansion of $n^{-1} U(\hat{\tau})$ about τ.

Finally, let $\hat{\beta}_{\mathrm{ML}}$ denote the unconditional maximum likelihood estimate with $U(\hat{\beta}_{\mathrm{ML}}) = 0$. Let $\min(p_i) = O(n^{\alpha})$ for $\alpha > 1$ so that the accuracy of the Laplace approximation to the marginal log-likelihood is approximately $O(n^{1-\alpha}) = o(1)$. Then, under the same assumptions as before, we have

$$U(\hat{\beta}) = U^*(\hat{\beta}) + o_p(1) = 0 + o_p(1).$$

Thus $U(\hat{\beta}) - U(\hat{\beta}_{\mathrm{ML}}) = o_p(1)$ and hence $\hat{\beta}$ is asymptotically equivalent to the unconditional maximum likelihood estimate, $\hat{\beta}_{\mathrm{ML}}$.

REFERENCES

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

BEAL, S. L. & SHEINER, L. B. (Ed.) (1992). *NONMEM User's Guide*. University of California, San Francisco: NONMEM Project Group.

BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc.* B **49**, 1–39.

DAVIDIAN, M. & GILTINAN, D. M. (1993). Some simple methods for estimating intra-individual variability in nonlinear mixed effects models. *Biometrics* **49**, 59–73.

LINDSTROM, M. J. & BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–87.

PRENTICE, R. L. & ZHAO, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–39.

RACINE-POON, A. (1985). Bayesian approach to nonlinear random effects models. *Biometrics* **41**, 1015–23.

SHEINER, L. B. & BEAL, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data. *J. Pharmacokin. Biopharmac.* **8**, 553–71.

SHEINER, L. B. & BEAL, S. L. (1985). Pharmacokinetic parameter estimates from several least squares procedures: superiority of extended least squares. *J. Pharmacokin. Biopharmac.* **13**, 185–201.

SOLOMON, P. J. & COX, D. R. (1992). Nonlinear component of variance models. *Biometrika* **79**, 1–11.

STIRATELLI, R., LAIRD, N. M. & WARE, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–71.

VONESH, E. F. (1992). Nonlinear models for the analysis of longitudinal data. *Statist. Med.* **11**, 1929–54.

VONESH, E. F. & CARTER, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48**, 1–17.

WOLFINGER, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791–5.

[*Received October* 1994. *Revised July* 1995]