Biometrika (1993), **80**, 4, *pp.* 791–5
Printed in Great Britain

Laplace's approximation for nonlinear mixed models

BY RUSS WOLFINGER
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, U.S.A.

SUMMARY

An approximation to Laplace's method for integrals is applied to marginal distributions of data arising from models in which both fixed and random effects enter nonlinearly. The approach provides alternative derivations of some recent algorithms for fitting such models, and it has direct ties with Gaussian restricted maximum likelihood and the accompanying mixed model equations.

Some key words: First-order method; Generalized linear models; Modified profile likelihood; Random effects; Restricted maximum likelihood.

1. INTRODUCTION

The nonlinear mixed model considered here is

$$y = f(X\alpha + Z\beta) + \varepsilon, \tag{1}$$

where y is a vector of n observations, α is a vector of p unknown fixed-effects parameters with known $n \times p$ design matrix, X, and β is a vector of f unknown random-effects parameters with known $n \times g$ design matrix Z. The function $f(.)$ is a known vector-valued nonlinear function, and it may depend on known covariates. Assume that β and ε are uncorrelated with means 0 and variances D and R, respectively.

Sheiner & Beal (1985) use (1) in pharmacokinetics. They fit it using the 'first-order' method, which involves a linearization of f about an estimate of α and 0, the latter being the expectation of β. Vonesh & Carter (1992) and Gumpertz & Pantula (1992) use the same approximation in order to use estimated generalized least squares. Solomon & Cox (1992, § 4) extend Sheiner & Beal's first-order method by including four terms in the expansion about 0. For the exponential model they consider, even this approximation breaks down for relatively large values of the single variance component they model in D.

Lindstrom & Bates (1990) attempt to improve the first-order approximation by linearizing about $\hat{\beta}$, the non-zero estimate of β corresponding to the best linear unbiased predictor in the linear case. This approximation, along with a Gaussian posterior approximation used by Laird & Louis (1982) and Stiratelli, Laird & Ware (1984), forms the motivation for the Lindstrom & Bates algorithm.

Alternatively, the Lindstrom & Bates algorithm can be derived using Laplace's approximation,

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

where θ is a q-vector of interest, $\hat{\theta}$ maximizes $e^{nl(\theta)}$, and n is large. The accuracy is $O(1/n)$, and therefore, for reasons seen in § 2, the Lindstrom & Bates approach can appropriately be called an 'approximate second-order' method.

The expansion around $\hat{\beta}$ instead of 0 often improves the first-order approximation, a claim illustrated by Wong & Li (1992) and Davison (1992, § 6). One useful application is to generalized linear mixed models, and § 3 applies Laplace's method to extend the work of Schall (1991).

Sections 2 and 3 focus on the restricted likelihood version of the Lindstrom & Bates method. Restricted, or marginal, likelihood has well-established connections with the modified profile likelihood of Barndorff-Nielsen (1983) and the approximate conditional likelihood of Cox and Reid (1987). All of the methods are equivalent for the Gaussian linear mixed model (Bellhouse, 1990; Cox & Reid, 1992, Example 1), and § 4 discusses relationships in the nonlinear case.

2. THE LINDSTROM & BATES METHOD

The Lindstrom & Bates procedure assumes normality for both β and ε, and consists of iterating between two steps, pseudo-data and linear-mixed-effects. The pseudo-data step fixes D and R at their current estimates, $\hat{D}$ and $\hat{R}$, and performs a nonlinear least squares optimization for α and β with data, predicted values, and derivatives as follows:

$$\begin{bmatrix} \hat{R}^{-\frac{1}{2}} y \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \hat{R}^{-\frac{1}{2}} f \\ \hat{D}^{-\frac{1}{2}} \beta \end{bmatrix}, \quad \begin{bmatrix} \hat{R}^{-\frac{1}{2}} \tilde{X} & \hat{R}^{-\frac{1}{2}} \tilde{Z} \\ 0 & \hat{D}^{-\frac{1}{2}} \end{bmatrix},$$

where

$$\hat{X} \equiv \frac{\partial f}{\partial \alpha^{\mathrm{T}}}\bigg|_{\hat{\alpha},\hat{\beta}} = f'(X\hat{\alpha} + Z\hat{\beta})X, \quad \tilde{Z} \equiv \frac{\partial f}{\partial \beta^{\mathrm{T}}}\bigg|_{\hat{\alpha},\hat{\beta}} = f'(X\hat{\alpha} + Z\hat{\beta})Z$$

and $\hat{\alpha}$ and $\hat{\beta}$ are current estimates. In the above expression f' is an $n \times n$ diagonal matrix.

The linear-mixed-effects step consists of using maximum likelihood or restricted maximum likelihood to estimate D and R. This step fits a linear mixed model with data

$$w \equiv y - f(X\hat{\alpha} + Z\hat{\beta}) + \tilde{X}\hat{\alpha} + \tilde{Z}\hat{\beta},$$

fixed effects design matrix $\tilde{X}$, and random effects design matrix $\tilde{Z}$.

The approach is flexible in modelling variability because both D and R can have arbitrary covariance structures. In the terminology of Liang & Zeger (1986), variance modelling in D corresponds to a subject-specific approach, and modelling in R corresponds to a population-averaged approach. The method is relatively easy to implement with available software and it has exhibited reasonable convergence properties for a number of examples.

Lindstrom & Bates derive their approach by linearizing f about $\hat{\alpha}$ and $\hat{\beta}$ at each step and then using the Gaussian posterior approximation of Laird & Louis (1982) and Stiratelli et al. (1984). They approximate the restricted maximum likelihood estimates of D and R, which are computed by maximizing the modified marginal distribution of y,

$$p(y) = \int p(y|\alpha, \beta)p(\beta)d\alpha\, d\beta,$$

where $p(.)$ represents respective density functions and the modification consists of integration over a tacitly present flat prior for α.

To apply (2), fix D and R and minimize

$$-(2n)l(\alpha, \beta) = \log|R| + \{y - f(X\alpha + Z\beta)\}^{\mathrm{T}} R^{-1} \{y - f(X\alpha + Z\beta)\} + \log|D| + \beta^{\mathrm{T}} D^{-1} \beta + (n+g)\log(2\pi) \tag{3}$$

over α and β. This is precisely the pseudo-data step of Lindstrom & Bates, and can be carried out using standard nonlinear least squares on augmented data. Let the minima be $\hat{\alpha}$ and $\hat{\beta}$, and define $V \equiv \tilde{Z} D \tilde{Z}^{\mathrm{T}} + R$. Now, the direct computation of $-l''(\hat{\alpha}, \hat{\beta})$ can be tedious if not intractable, and so we use the usual nonlinear-least-squares approximation:

$$\begin{bmatrix} \tilde{X}^{\mathrm{T}} R^{-1} \tilde{X} & \tilde{X}^{\mathrm{T}} R^{-1} \tilde{Z} \\ \tilde{Z}^{\mathrm{T}} R^{-1} \tilde{X} & \tilde{Z}^{\mathrm{T}} R^{-1} \tilde{Z} + D^{-1} \end{bmatrix}. \tag{4}$$

This can also be treated as $-E\{l''(\hat{\alpha}, \hat{\beta})\}$, and this Fisher score is what makes the final result approximately second-order instead of precisely so. Partitioned determinant formulae then yield

$$-2\log p(y) \doteq \log|V| + \{y - f(X\hat{\alpha} + Z\hat{\beta})\}^{\mathrm{T}} R^{-1} \{y - f(X\hat{\alpha} + Z\hat{\beta})\} \log|\tilde{X}^{\mathrm{T}} V^{-1} \tilde{X}| + \hat{\beta}^{\mathrm{T}} D^{-1} \hat{\beta} + (n-p)\log(2\pi) + (p+g)\log n.$$

Assuming the estimates $\hat{\alpha}$ and $\hat{\beta}$ are interior, they satisfy the first-order conditions

$$\tilde{X}^{\mathrm{T}} R^{-1} \{y - f(X\hat{\alpha} + Z\hat{\beta})\} = 0, \quad \tilde{Z}^{\mathrm{T}} R^{-1} \{y - f(X\hat{\alpha} + Z\hat{\beta})\} = D^{-1}\hat{\beta}, \tag{5}$$

which can be rewritten as

$$\begin{bmatrix} \tilde{X}^{\mathrm{T}} R^{-1} \tilde{X} & \tilde{X}^{\mathrm{T}} R^{-1} \tilde{Z} \\ \tilde{Z}^{\mathrm{T}} R^{-1} \tilde{X} & \tilde{Z}^{\mathrm{T}} R^{-1} \tilde{Z} + D^{-1} \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \tilde{X}^{\mathrm{T}} R^{-1} w \\ \tilde{Z}^{\mathrm{T}} R^{-1} w \end{bmatrix}. \tag{6}$$

A third formulation is

$$\hat{\alpha} = (X^{\mathrm{T}} V^{-1} X)^{-} X^{\mathrm{T}} V^{-1} w, \quad \hat{\beta} = M \tilde{Z}^{\mathrm{T}} R^{-1} (w - \tilde{X}\hat{\alpha}),$$

where $M \equiv (\tilde{Z}^{\mathrm{T}} R^{-1} \tilde{Z} + D^{-1})^{-1}$. This result combined with $V^{-1} = R^{-1} - R^{-1} \tilde{Z} M \tilde{Z}^{\mathrm{T}} R^{-1}$ yields

$$-2\log p(y) \doteq \log|V| + (w - \tilde{X}\hat{\alpha})^{\mathrm{T}} V^{-1} (w - \tilde{X}\hat{\alpha}) + \log|\tilde{X}^{\mathrm{T}} V^{-1} \tilde{X}| + (n-p)\log(2\pi) + (p+g)\log n, \tag{7}$$

which is the Gaussian restricted log likelihood of a linear mixed model. The minimization of (7) over the parameters of D and R found in V and the subsequent update of $\hat{\alpha}$ yields the restricted maximum likelihood version of Lindstrom & Bates linear-mixed-effects step. A minor modification of the algorithm can be achieved by also updating $\hat{\beta}$ before performing another pseudo-data step.

The mixed-model equations (6) are a nice summary of the whole approach. They can be viewed as a Gauss–Markov extension of the normal equations in the usual linear model (Harville, 1976), and since they also describe the fixed-point relation from the pseudo-data step, the final $\hat{\alpha}$ and $\hat{\beta}$ from both steps agree. The coefficient matrix in (6) is the same as (4), so the estimated variance-covariance matrices at convergence agree as well. In fact, solving (6) is the first Gauss–Newton iteration of the pseudo-data step. Therefore, one could potentially bypass the nonlinear least squares optimization altogether and employ an iterative restricted maximum likelihood algorithm, thereby extending iteratively reweighted least squares (McCullagh & Nelder, 1989, § 2.5).

3. GENERALIZED LINEAR MIXED MODELS

Building on the development of the previous section, define $\mu = f(X\alpha + Z\beta)$, and assume that the elements of f are evaluations of g^{-1}, where $g(.)$ is a link function. Assume further that $E(y|\mu) = \mu$ and

$$\text{cov}\,(y|\mu) = \Lambda_\mu \equiv R_\mu^{\frac{1}{2}} R R_\mu^{\frac{1}{2}}. \tag{8}$$

Here R_μ is a diagonal matrix with components equal to evaluations of the appropriate variance function at each component of μ; R contains the dispersion parameter, if one exists, and it can be extended to be a 'working' correlation matrix following Liang & Zeger (1986) or a parameterized covariance matrix that models extra dispersion. The random effects β are still assumed to be $N(0, D)$. Compared to the Gaussian nonlinear mixed model discussed in the previous section, this model has a more general error distribution but a more restrictive nonlinear function.

Rather than writing down a form for $p(y|\alpha, \beta)$, we take a quasi-likelihood approach (McCullagh & Nelder, 1989, §9.3) and assume that its derivatives, along with the normal ones for $p(\beta)$, produce the score equations

$$\tilde{X}^{\mathrm{T}} \Lambda_\mu^{-1} (y - \mu) = 0, \quad \tilde{Z}^{\mathrm{T}} \Lambda_\mu^{-1} (y - \mu) = D^{-1} \beta.$$

These are the analogues of (5), and can be viewed as an extension of the generalized estimating equations of Liang & Zeger (1986). They can be rewritten, as in (6), as

$$\begin{bmatrix} \tilde{X}^{\mathrm{T}} \Lambda_\mu^{-1} \tilde{X} & \tilde{X}^{\mathrm{T}} \Lambda_\mu^{-1} \tilde{Z} \\ \tilde{Z}^{\mathrm{T}} \Lambda_\mu^{-1} \tilde{X} & \tilde{Z}^{\mathrm{T}} \Lambda_\mu^{-1} \tilde{Z} + D^{-1} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \tilde{X}^{\mathrm{T}} \Lambda_\mu^{-1} v \\ \tilde{Z}^{\mathrm{T}} \Lambda_\mu^{-1} v \end{bmatrix}, \tag{9}$$

in which $v \equiv g(\mu) + g'(\mu)(y - \mu)$ is a Taylor series approximation to the linked response $g(y)$.

Letting $V_\mu \equiv \Lambda_\mu + \tilde{Z} D \tilde{Z}^{\mathrm{T}}$, Laplace's method results in

$$\begin{aligned} -2 \log p(y) \doteq & \log |V_{\hat{\mu}}| - \log |\Lambda_{\hat{\mu}}| - 2 \log p(y|\alpha, \beta)|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}} \\ & + \log |\tilde{X}^{\mathrm{T}} V_{\hat{\mu}}^{-1} \tilde{X}| - p \log (2\pi) + (p + g) \log n. \end{aligned}$$

At this stage one must produce some form for $p(y|\alpha, \beta)$, presumably involving R_μ and R, and then optimize the resulting expression over the parameters in D and R to obtain the analogue of Lindstrom & Bates's linear-mixed-effects step. One simple alternative is to approximate $p(y|\alpha, \beta)$ by a Gaussian distribution with mean μ and variance (8). In fact, Lindstrom & Bates use such an approximation to derive their algorithm.

Given such a Gaussian approximation, one is led to the linear mixed model $v = X\alpha + Z\beta + \varepsilon$, in which β is $N(0, D)$ and ε is $N(0, \Lambda_\mu)$. An algorithm suggested from this is to specify an initial value for μ, and then to iteratively fit the Gaussian linear mixed model, updating μ from the mixed-model equations at each iteration. Such an approach turns out to be equivalent to the method proposed by Schall (1991), who considers the case where R is the identity matrix and D is diagonal containing variance components.

4. DISCUSSION

In (7), $\log |\tilde{X}^{\mathrm{T}} V^{-1} \tilde{X}|$ corresponds to the restricted maximum likelihood adjustment to maximum likelihood (Harville, 1974). It also is the analogue of the observed information component of formula (3·2) given by Barndorff-Nielsen (1983), that is, the approximation is akin to a modified profile likelihood. But as in the approximate conditional adjustment

of Cox & Reid (1987, § 4.1), Barndorff-Nielsen's Jacobian term, $\log|\partial\hat{\alpha}/\partial\hat{\alpha}_{D,R}|$ is ignored. Here $\hat{\alpha}$ is the maximum likelihood estimate and $\hat{\alpha}_{D,R}$ is the estimate for fixed values of D and R.

In the linear case $\log|\partial\hat{\alpha}/\partial\hat{\alpha}_{D,R}| = 0$, and in general Cox & Reid show it is $O_p(1/n)$ by assuming parameter orthogonality in the Fisher information metric. The orthogonality of α and (D, R) is evident from (3) and (5), a phenomenon already noted for the mean and variance parameters of the multivariate normal distribution and for mean/canonical mixtures in exponential families (Cox & Reid, 1987, § 3.2–3). This orthogonality also provides a justification for alternating between the two steps in the Lindstrom & Bates approach.

Acknowledgements

The author thanks the referees for helpful criticisms.

References

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

Bellhouse, D. R. (1990). On the equivalence of marginal and approximated conditional likelihoods for correlation parameters under a normal model. *Biometrika* **77**, 743–6.

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

Cox, D. R. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–11.

Davison, A. C. (1992). Treatment effect heterogeneity in paired data. *Biometrika* **79**, 463–74.

Gumpertz, M. L. & Pantula, S. G. (1992). Nonlinear regression with variance components. *J. Am. Statist. Assoc.* **87**, 201–9.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–5.

Harville, D. A. (1976). Extension of the Gauss–Markov theorem to include the estimation of random effects. *Ann. Statist.* **4**, 384–95.

Laird, N. M. & Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. *J. R. Statist. Soc.* B **44**, 190–200.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–27.

Sheiner, L. B. & Beal, S. L. (1985). Pharmacokinetic parameter estimates from several least squares procedures: Superiority of extended least squares. *J. Pharmacokin. Biopharm.* **13**, 185–201.

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

Stiratelli, R., Laird, N. & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrika* **40**, 961–71.

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

Wong, W. H. & Li, B. (1992). Laplace expansion for posterior densities of nonlinear functions of parameters. *Biometrika* **79**, 393–8.

[*Received August* 1992. *Revised March* 1993]