Partitioned algorithms for maximum likelihood and other nonlinear estimation*

Gordon K. Smyth
Department of Mathematics
University of Queensland, Australia 4072.

15 April 1996

Abstract

There are a variety of methods in the literature which seek to make iterative estimation algorithms more manageable by breaking the iterations into a greater number of simpler or faster steps. Those algorithms which deal at each step with a proper subset of the parameters are called in this paper partitioned algorithms. Partitioned algorithms in effect replace the original estimation problem with a series of problems of lower dimension. The purpose of the paper is to characterize some of the circumstances under which this process of dimension reduction leads to significant benefits.

Four types of partitioned algorithms are distinguished: reduced objective function methods, nested (partial Gauss-Seidel) iterations, zigzag (full Gauss-Seidel) iterations, and leapfrog (non-simultaneous) iterations. Emphasis is given to Newton-type methods using analytic derivatives, but a nested EM algorithm is also given. Nested Newton methods are shown to be equivalent to applying to same Newton method to the reduced objective function, and are applied to separable regression and generalized linear models. Nesting is shown to general improve the convergence of Newton-type methods, both by improving the quadratic approximation to the log-likelihood and by improving the accuracy with which the observed information matrix can be approximated. Nesting is recommended whenever a subset of parameters is relatively easily estimated. The zigzag method is shown to produce a stable by generally slow iteration; it is fast and recommended when the parameter subsets have approximately uncorrelated estimates. The leapfrog iteration has less guaranteed properties in general, but is similar to nesting and zigzaging when the parameter subsets are orthogonal.

*Smyth, G. K. (1996). Partitioned algorithms for maximum likelihood and other nonlinear estimation. *Statistics and Computing*, **6**, 201–216.

Keywords: conditionally linear parameters, convergence, EM algorithm, Gauss-Seidel iteration, method of scoring, Newton method, nonlinear estimation, orthogonal parameters, separable regression.

1 Introduction

Statisticians have given increased attention to nonlinear models in recent years as computing resources have become ever more readily available. See for example the recent books by Ratkowsky (1983, 1989), Gallant (1987), Bates and Watts (1988), Seber and Wild (1989), Ross (1990) and Chambers and Hastie (1992). Usually the parameters of such models must be estimated iteratively. There are a variety of methods in the literature or in common usage which seek to make iterative estimation algorithms more manageable by breaking the iterations into a greater number of simpler or faster steps. Examples are separable least squares (Ross, 1990, Section 5.4), the Gauss-Seidel method (Ortega and Rheinboldt, 1970, Section 7.4; Thisted, 1988, Section 4.3.4), two-stage least squares (Seber, 1989, Section 6.2.3) and other methods are discussed in Ross (1970). In this paper those algorithms which deal at each step with a proper subset of the total parameter vector are called partitioned algorithms. Partitioned algorithms in effect replace the original estimation problem with a series of problems of lower dimension. The purpose of this paper is to characterize some of the circumstances under which this process of dimension reduction leads to significant benefits.

Let θ be the p-dimensional parameter vector to be estimated. In general it will be assumed that there is already available some unpartitioned algorithm, represented by the updating equation $\theta^{k+1} = F(\theta^k)$, which may or may not converge for the particular data set under consideration but which has the desired estimator $\hat{\theta}$ as a fixed point. When θ is partitioned into subvectors θ_1 and θ_2, the updating equation will be written formally as

$$\begin{aligned} \theta_1^{k+1} &= F_1(\theta_1^k, \theta_2^k) \\ \theta_2^{k+1} &= F_2(\theta_1^k, \theta_2^k) \end{aligned} \tag{1}$$

where F_1 and F_2 are just the corresponding partition of F.

One class of problems for which partitioning is useful are those in which some parameters are much easier to estimate than others. This situation arises in nonlinear regression when some parameters enter the fitted values linearly. Consider for example the exponential growth model $y = \theta_1 \exp(\theta_2 x)$ to be fitted to data pairs (x_i, y_i), $i = 1, \ldots, n$. An old technique is to note that the sum of squares is easily minimized with respect to θ_1 for any given value of θ_2 by the linear least squares estimator $\hat{\theta}_1(\theta_2) = \sum y_i \exp(\theta_2 x_i) / \sum \exp(2\theta_2 x_i)$. This leaves the problem of minimizing with respect to θ_2, which can be accomplished by substituting $\hat{\theta}_1(\theta_2)$ into the second part of (1), i.e.,

$$\theta_2^{k+1} = F_2(\hat{\theta}_1(\theta_2^k), \theta_2^k)$$

which is an example of what is called a nested iteration in this paper, since the estimation of θ_1 is nested within that of θ_2. Note that the nested iteration will

necessarily have the least squares estimate $\hat{\theta}_2$ as a fixed point, and has the advantage over the original iteration of lower dimension. This simple strategy is often very effective in practice. The nested iteration has the effect of restricting the working parameter estimates to the locus $\theta_1 = \hat{\theta}_1(\theta_2)$.

A second class of problems for which partitioning is useful are those for which the parameters fall into two or more groups with approximately uncorrelated estimates. For example consider the following heteroscedastic regression model (Aitkin, 1987; Smyth, 1989). Suppose observations y_i are normal and independent with means $\mu_i = \beta_1 + \beta_2 x_i$ and variances $\sigma_i^2 = \exp(\gamma_1 + \gamma_2 z_i)$ where the x_i and z_i are observed covariates. Given the variances, the β_i are estimated by linear regression of the y_i on the x_i with weights $1/\sigma_i^2$. Given the means, the γ_i can be estimated by gamma regression of the squared residuals $(y_i - \mu_i)^2$ on the z_i. Cycling between these two regressions leads, if convergence is achieved, to maximum likelihood estimates for all parameters. The iteration here is

$$\begin{array}{rcl} \theta_1^{k+1} & = & \hat{\theta}_1(\theta_2^k) \\ \theta_2^{k+1} & = & \hat{\theta}_2(\theta_1^{k+1}), \end{array} \tag{2}$$

where $\theta_1^T = (\beta_1, \beta_2)^T$ and $\theta_2^T = (\gamma_1, \gamma_2)^T$ and $\hat{\theta}_1(\theta_2)$ and $\hat{\theta}_2(\theta_1)$ maximize the likelihood with respect to θ_1 and θ_2 respectively while keeping the other fixed. The iteration zigzags between the two surfaces $\theta_1 = \hat{\theta}_1(\theta_2)$ and $\theta_2 = \hat{\theta}_2(\theta_1)$. This type of iteration is used a great deal by statisticians, so it is important to understand its properties.

In this paper, four types of partitioned algorithms are distinguished.

1. *Reduced objective function methods.* Given the objective function ℓ to be maximized with respect to parameter vectors θ_1 and θ_2, the partial estimate $\hat{\theta}_2(\theta_1)$ is substituted into ℓ to obtain $r(\theta_2) = \ell(\hat{\theta}_1(\theta_2), \theta_2)$. The reduced objective function $r(\theta_2)$ may then be maximized by any algorithm, but is often passed to a derivative free method.

2. *Nested (partial Gauss-Seidel) iterations.* Given the iteration function (1), discard θ_1^{k+1} at each iteration and replace it with $\hat{\theta}_1(\theta_2^{k+1})$. Alternatively, replace it with $\tilde{\theta}_1(\theta_2)$, where $\tilde{\theta}_1(\theta_2)$ is the value of θ_1 which solves $\theta_1 = F_1(\theta_1, \theta_2^{k+1})$.

3. *Zigzag (full Gauss-Seidel) iterations.* To solve the simultaneous normal equations $\dot{\ell}_1(\theta_1, \theta_2) = \dot{\ell}_2(\theta_1, \theta_2) = 0$, alternate between solving $\dot{\ell}_1 = 0$ with respect to θ_1 and $\dot{\ell}_2 = 0$ with respect to θ_2. Here $\dot{\ell}_1$ and $\dot{\ell}_2$ are the derivatives of ℓ with respect to θ_1 and θ_2 respectively. This gives the zigzag iteration (2). Alternatively we could apply the Gauss-Seidel method to the iteration functions and cycle between solving $\theta_1 = F_1(\theta_1, \theta_2)$ with respect to θ_1 and $\theta_2 = F_2(\theta_1, \theta_2)$ with respect to θ_2.

4. *Leapfrog (non-simultaneous) iterations.* Given iteration functions for θ_1 and θ_2, use the already updated value of θ_1 to update θ_2, i.e.,

$$\begin{array}{rcl} \theta_1^{k+1} & = & F_1(\theta_1^k, \theta_2^k) \\ \theta_2^{k+1} & = & F_2(\theta_1^{k+1}, \theta_2^k), \end{array}$$

In the heteroscedastic regression example above, performing only one cycle of the gamma regression at each iteration would result in a leapfrog iteration.

It has been known for some time that reduced objective function methods are useful in conjunction with optimization methods that do not use derivatives (Ross, 1970; Lawton and Sylvestre, 1970). The calculation of analytic derivatives is however very desirable when there are many parameters, when the parameters are highly correlated or when high accuracy is otherwise required. See for example Seber and Wild (1989, p. 611). Special attention is given in this paper to Newton-type methods based on first and second derivatives. In the next section it is shown that nested Newton methods are equivalent to applying the same Newton method to the reduced objective function. In Sections 5 and 6 it is shown that nesting generally improves convergence of Newton-type methods, both by improving the quadratic approximation to the log-likelihood and by improving the accuracy with which the observed information matrix can be approximated.

Section 5 also gives a nested EM algorithm, and shows that to be effective nesting should be applied to those parameters about which there is relatively least information in the incomplete data.

Nested algorithms are never worse than full algorithms, so nesting can be recommended generally whenever it is easily implemented. The zigzag method produces a stable but generally slow iteration; it is fast and recommended when θ_1 and θ_2 have approximately uncorrelated estimates. The leapfrog iteration has less guaranteed properties in general, but has similar properties to nesting and zigzaging when θ_1 and θ_2 are orthogonal.

In practice, Newton methods are often implemented with modifications designed to ensure convergence. This is especially true of the Gauss-Newton algorithm for nonlinear least squares, which is now almost always implemented with Levenberg-Marquardt damping or a line search method. Such methods can be extended to the general likelihood context (Jørgensen, 1984; Osborne, 1987; Thisted, 1988), although they are still seldom used with exponential families or generalized linear models. In this paper, the algorithms are examined unmodified so as to not obscure their structure and in the belief that it is easier to modify an algorithm that is already well behaved. It is recommended that in practice damping or line search methods be applied to the appropriate partitioned algorithm. Some results proved here apply directly to modified algorithms. In particular, the limiting convergence rate results of Section 5 apply directly to Levenberg-Marquardt modified iterations provided that it is arranged that the damping parameter converges to zero as the iteration converges.

Nested and reduced objective function methods arise naturally from an inferential point of view when θ_1 is to be treated as a nuisance parameter, since $r(\theta_2)$ is the profile likelihood for θ_2. On the other hand, the zigzag iteration is natural when θ_1 and θ_2 correspond to interpretable submodels as in the heteroscedastic example.

Section 2 of this paper discusses nested Newton-type methods, Section 3 discusses the zigzag iteration and Section 4 discusses their application to nonlinear regression. Convergence of the algorithms is discussed in Sections 5 and 6. Section 5 deals with local convergence rates close to the stationary value, while Section 6 deals with global convergence and the quadratic approximation to the log-likelihood function.

2 Nested Newton-type Algorithms

Let $\ell(\theta)$ be the log-likelihood function of a p-dimensional parameter vector θ. The Newton-Raphson algorithm attempts to maximize ℓ by maximizing instead its quadratic Taylor series expansion at the current working value. The iteration function is

$$\theta^{k+1} = \theta^k - \ddot{\ell}^{-1}(\theta^k)\dot{\ell}(\theta^k)$$

where $\dot{\ell}$ is the gradient or score vector and $-\ddot{\ell}$ is the observed information matrix. More generally Newton-type algorithms have iteration functions of the form

$$\theta^{k+1} = \theta^k + A^{-1}(\theta^k)\dot{\ell}(\theta^k)$$

where A is a suitably chosen approximation to $-\ddot{\ell}$. In statistics the most common choice for A is the expected information $\mathcal{I}$, because $\mathcal{I}$ has a particularly elegant form when ℓ is an exponential or curved exponential family. The resulting iteration is known as Fisher's method of scoring (Rao, 1973; Osborne, 1992).

Now suppose that θ is partitioned into θ_1 and θ_2 and that θ_1 is to be treated as a nuisance parameter. In an often overlooked paper, Richards (1961) derived the Newton-Raphson iteration for maximizing the profile likelihood $r(\theta_2)$. Letting

$$\ddot{\ell} = \begin{pmatrix} \ddot{\ell}_{11} & \ddot{\ell}_{12} \\ \ddot{\ell}_{21} & \ddot{\ell}_{22} \end{pmatrix}$$

be the obvious partition of $\ddot{\ell}$, Richards shows that the derivatives of $r(\cdot)$ are given by $\dot{r}(\theta_2) = \dot{\ell}_2(\hat{\theta}_1(\theta_2), \theta_2)$ and $\ddot{r}(\theta_2) = \ddot{\ell}_{2.1}(\hat{\theta}_1(\theta_2), \theta_2)$ where $\ddot{\ell}_{2.1} = \ddot{\ell}_{22} - \ddot{\ell}_{21}\ddot{\ell}_{11}^{-1}\ddot{\ell}_{12}$. The profile Newton-Raphson iteration therefore is

$$\theta_2^{k+1} = \theta_2^k - \ddot{\ell}_{2.1}^{-1}\dot{\ell}_2$$

where all quantities on the right hand side are evaluated at θ_2^k and $\hat{\theta}_1(\theta_2^k)$.

If we wish to instead apply scoring to the profile likelihood we meet the difficulty that the expected value of $-\ddot{\ell}_{2.1}$ will seldom be simple or useful to evaluate. A more amenable alternative is to use the nested scoring algorithm, which is developed as follows. We first calculate the partial iteration functions F_1 and F_2 for Newton-type algorithms. It is useful to have the block Cholesky decomposition of A,

$$A = \begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & 0 \\ 0 & A_{2.1} \end{pmatrix} \begin{pmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix}$$

with $A_{2.1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$. This allows us to write

$$A^{-1} = \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{2.1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix}$$

so that the Newton-type iteration can be partitioned as

$$\begin{aligned} \theta_1^{k+1} &= F_1(\theta_1^k, \theta_2^k) &&= \theta_1^k + A_1^{-1}\dot{\ell}_1 - A_1^{-1}A_{12}A_{2.1}^{-1}\dot{\ell}_{2.1} \\ \theta_2^{k+1} &= F_2(\theta_1^k, \theta_2^k) &&= \theta_2^k + A_{2.1}^{-1}\dot{\ell}_{2.1} \end{aligned}$$

where $\dot{\ell}_{2.1} = \dot{\ell}_2 - A_{21}A_1^{-1}\dot{\ell}_1$. This shows that the nested Newton-type iteration is

$$\theta_2^{k+1} = F_2(\hat{\theta}_1(\theta_2^k), \theta_2^k) = \theta_2^k + A_{2.1}^{-1}\dot{\ell}_2$$

noting that $\dot{\ell}_{2.1} = \dot{\ell}_2$ at $\theta_1 = \hat{\theta}_1(\theta_2)$. Hence the nested Newton-Raphson iteration is equivalent to Newton-Raphson applied directly to the profile likelihood, while other Newton-type algorithms approximate $-\ddot{\ell}_{2.1}$ with the corresponding partitioned matrix $A_{2.1}$.

A nested iteration can be thought of as equivalent to a full Newton-type iteration with θ_1^{k+1} discarded and replaced by $\hat{\theta}_1(\theta_2)$ (see also Bates and Lindstrom, 1986). For some problems, it may be usefully programmed this way. In this sense it is a natural method of restricting the iteration to the locus $\{(\theta_1, \theta_2) : \theta_1 = \hat{\theta}_1(\theta_2)\}$. A possible complication is that $\hat{\theta}_1(\theta_2)$ can fail to be unique for some θ_2 even though the likelihood has a well-defined global maximum. In such cases the above locus is disconnected. While this does not prevent the nested iteration from being defined, nesting is most likely to be useful when $\hat{\theta}_1(\theta_2)$ is available in closed form or can otherwise be guaranteed to be unique.

Example 2.1. Suppose that $y_1, \ldots, y_n$ is a sample from a population such that $Z = (Y^\lambda - 1)/\lambda$ is $N(\mu, \sigma^2)$ for some $\lambda > 0$ (Box and Cox, 1964), and consider the problem of estimating λ. The conditional maximum likelihood estimators for μ and σ^2 are $\hat{\mu}(\lambda) = \bar{z}$ and $\hat{\sigma}^2(\lambda) = s_z^2 = \frac{1}{n}\sum(z_i^2 - \bar{z})^2$ respectively, and at these values $\ddot{\ell}_{\mu,\sigma^2} = 0$. The nested Newton-Raphson iteration for λ therefore is

$$\lambda^{k+1} = \lambda^k - \ddot{\ell}_{\lambda|\mu,\sigma^2}^{-1}\dot{\ell}_\lambda$$

with

$$\ddot{\ell}_{\lambda|\mu,\sigma^2} = \ddot{\ell}_\lambda - \ddot{\ell}_\mu^{-1}\ddot{\ell}_{\lambda\mu}^2 - \dot{\ell}_{\sigma^2}^{-1}\ddot{\ell}_{\lambda\sigma^2}^2$$

Explicit expressions for these derivatives are

$$\dot{\ell}_\lambda = \sum \log y_i - \sigma^{-2}\sum(z_i - \mu)\frac{dz_i}{d\lambda}$$

and

$$\ddot{\lambda}_{\lambda|\mu,\sigma^2} = -\frac{1}{\sigma^2}\sum\left[\left(\frac{dz_i}{d\lambda}\right)^2 + (z_i - \mu)\frac{d^2z_i}{d\lambda^2}\right] - \frac{1}{n\sigma^2}\left(\sum\frac{dz_i}{d\lambda}\right)^2 - \frac{2}{n\sigma^4}\left[\sum(z_i - \mu)\frac{dz_i}{d\lambda}\right]^2$$

The above iteration is equivalent to applying Newton-Raphson to the profile likelihood

$$r(\lambda) = \ell(\hat{\mu}(\lambda), \hat{\sigma}^2(\lambda), \lambda) = -\frac{n}{2}\log 2\pi s_z^2 - \frac{n}{2} + (\lambda - 1)\sum \log y_i$$

Example 2.2. Jørgensen (1987) has shown that there exists a generalized linear model distribution with power variance function $\mathrm{var}(Y) = \sigma^2\mu^\gamma$, $\mu = E(Y)$, for each $\gamma \geq 1$ or ≤ 0. If $1 < \gamma < 2$ then the distribution is a Poisson mixture of gamma distributions, $Y = \sum_{i=1}^N X_i$ where N is Poisson(λ) and the X_i are gamma(α,δ) with $\lambda = \mu^\gamma/\sigma^2(2-\gamma)$, $\alpha = (2-\gamma)/(\gamma - 1)$ and $\delta = \sigma^2(\gamma - 1)\mu^{\gamma-1}$

(Jørgensen, 1987, p. 140; Smyth, 1992). This is a mixed distribution, continuous on the positive half-line and with probability mass at zero. It is of considerable interest for modelling non-negative data with exact zeros, and in general it will be necessary to estimate γ. The log-density function can be written

$$\log f(y;\mu,\sigma^2,\gamma) = \begin{cases} -\frac{1}{\sigma^2} + \log \delta_0 & y = 0 \\ \frac{1}{\sigma^2}\left(y\frac{\mu^{1-\gamma}}{1-\gamma} - \frac{\mu^{2-\gamma}}{2-\gamma}\right) - \log y + \log W(y,\sigma^2,\gamma) & y > 0 \end{cases}$$

where δ_0 is the Dirac delta function at zero and W is Wright's Bessel function,

$$W(y,\sigma^2,\gamma) = \sum_{j=1}^{\infty} \frac{\lambda^j (y/\delta)^{j\alpha}}{j!\Gamma(j\alpha)}$$

(Tweedie, 1984, p. 592). In generalized linear model applications the mean of each observation will be assumed to be a function of a vector β of regression coefficients. For any given value of γ, maximum likelihood estimates of β can be calculated as for a generalized linear model (McCullagh and Nelder, 1989), although the estimation of σ^2 does require maximizing an infinite series. In calculating the complete information matrix for all parameters, β, σ^2 and γ, it is natural to put those derivatives involving β to their expectations, in particular $E(\ddot{\ell}_{\beta,\sigma^2}) = E(\ddot{\ell}_{\beta,\gamma}) = 0$. On the other hand it is more natural to use observed information for σ^2 and γ since $\ddot{\ell}_{\sigma^2}$, $\ddot{\ell}_{\gamma}$ and $\ddot{\ell}_{\sigma^2,\gamma}$ involve infinite series and have expectations which are not easily evaluated. The resulting information matrix is a mixture of observed and expected information. The corresponding nested Newton-type algorithm for γ is

$$\gamma^{k+1} = \gamma^k + A^{-1}_{\gamma|\beta,\sigma^2}\dot{\ell}_{\gamma}$$

where

$$-A_{\gamma|\beta,\sigma^2} = \ddot{\ell}_{\gamma} - \ddot{\ell}^{-1}_{\sigma^2}\ddot{\ell}^2_{\sigma^2,\gamma}$$

and all quantities are evaluated at $\hat{\beta}(\gamma)$ and $\hat{\sigma}^2(\gamma)$.

Example 2.3. The hyperbolic distribution has density

$$f(y;\mu,\delta,\alpha,\beta) = a(\delta,\alpha,\beta)\exp[\alpha\sqrt{\delta^2+(y-\mu)^2} + \beta(y-\mu)]$$

where $a(\delta,\alpha,\beta)$ is a normalizing constant, and is characterized by the fact that $\log f$ is a hyperbolic function of y (Barndorff-Nielsen, 1977). By contrast, the normal distribution gives a parabola. The hyperbolic distribution is often used to model particle size distributions. Let r_i be the proportion of a sample which is between sizes s_i and s_{i+1}, $i = 1,\ldots,k$, with $s_1 = 0$ and $s_{k+1} = \infty$. Then the parameters may be estimated by minimizing the Kullback-Lieber distance between the empirical and theoretical probabilities

$$\ell = \sum_{i=1}^{k} r_i \log p_i$$

where p_i is the probability mass between s_i and s_{i+1}. First and second derivatives of ℓ are given by Jensen (1988). Experience has shown that maximizing ℓ

can be a difficult numerical problem (Fieller, Flenley and Olbricht, 1992). The greatest difficulty is associated with estimation of the scale parameter δ, and the nested Newton-Raphson iteration for δ with the other three parameters nested has proved to be more stable that the full Newton-Raphson for all four parameters simultaneously (Jensen, 1988).

3 Zigzag Algorithms

The zigzag iteration (also known as the Gauss-Seidel method) given in the introduction extends to any number of parameter subsets. Let $\theta_1, \ldots, \theta_m$ be, possibly vector, parameters. The zigzag method for maximizing the log-likelihood ℓ maximizes ℓ with respect to each θ_j in turn. Ortega and Rheinboldt (1970) and Thisted (1987) consider the possibility of cycling through the θ_j in different orders, for example choosing at each stage the θ_j for which is $\partial\ell/\partial\theta_j$ is farthest from zero, or cycling first $\theta_1, \ldots, \theta_m$ then $\theta_m, \ldots, \theta_1$, but in this paper it will be assumed that the θ_j are cycled in index order. Maximizing ℓ with respect to θ_j will itself in general require iteration. This will be called inner iteration, while one cycle through the $\theta_1, \ldots, \theta_m$ will be called an outer iteration. If, for example, scoring is used for the inner iterations, the algorithm will be called zigzag scoring.

In each inner iteration the likelihood is considered to be a function of θ_j only, the other parameters being fixed. This may be called the submodel corresponding to θ_j (Smyth, 1989). If $\theta_1, \ldots, \theta_m$ are orthogonal, then standard errors and score tests calculated in any submodel are correct for the complete likelihood.

If the parameters are orthogonal, the leapfrog scoring iteration for $\theta_1, \ldots, \theta_m$ is

$$\begin{aligned}
\theta_1^{k+1} &= \theta_1^k + \mathcal{I}_1^{-1}\dot{\ell}_1(\theta_1^k, \ldots, \theta_m^k) \\
\theta_2^{k+1} &= \theta_2^k + \mathcal{I}_2^{-1}\dot{\ell}_2(\theta_1^{k+1}, \theta_2^k, \ldots, \theta_m^k) \\
&\vdots \\
\theta_m^{k+1} &= \theta_m^k + \mathcal{I}_m^{-1}\dot{\ell}_m(\theta_1^{k+1}, \ldots, \theta_{m-1}^{k+1}, \theta_m^k).
\end{aligned}$$

Zigzag scoring consists of iterating each of the m equations to convergence before going on to the next.

Example 3.1. Consider the multiple regression model, $E(Y_i) = \beta_1 x_{1i} + \ldots + \beta_m x_{mi}$ with $\text{var}(Y_i) = \sigma^2$. The zigzag iteration for the β_j estimates β_1 by regressing $y - \beta_2^k x_2 - \ldots - \beta_m x_m$ on x_1, and so on through the other β_j. In Section 5 it will be shown that this iteration is very slow unless the covariates are orthogonal.

Example 3.2. Consider the rational model $E(Y_i) = (\alpha_1 x_{1i} + \ldots + \alpha_p x_{pi})/(\beta_1 z_{1i} + \ldots + \beta_p x_{pi})$. If the β_j are considered fixed, the α_j may be estimated by linear regression. If the α_j are fixed, the β_j may be estimated using a generalized linear model with reciprocal link function.

Example 3.3. Suppose that the Y_i are normal, gamma or inverse-Gaussian, with a link-linear regression model

$$g(\mu_i) = \beta^T x_i$$

for the means. Suppose that each Y_i has it's own dispersion parameter ϕ and suppose a link linear model

$$h(\phi_i) = \gamma^T z_i$$

for the dispersions. Then the submodels for the mean and dispersion parameters respectively are themselves generalized linear models (Smyth, 1989). This includes the often used heteroscedastic regression model with log-linear variances (Aitkin, 1987; Verbyla, 1993). When the mean model is just a linear regression, there is no inner iteration in the mean submodel. In that case the leapfrog iteration for β and γ is equivalent to a nested iteration with the iteration for β nested within that for γ, while the zigzag iteration for β and γ is equivalent to a nested iteration with the iteration for γ nested within that for γ.

Example 3.4. Let Y_i be normal with means μ_i and autoregressive errors. Let α be the autoregressive parameters and let σ^2 be the variances of the innovations. In general, the means, variances and autoregressive parameters will themselves be functions of a smaller number unknown parameters. The mean parameters, the variance parameters, and the autoregressive parameters are mutually orthogonal, suggesting that a zigzag iteration may be useful for maximum likelihood estimation.

4 Applications to Regression

4.1 Separable Least Squares

The oldest and most common application of partitioned algorithm ideas is to regression. If the data vector y is from a normal distribution with mean $\mu(\theta)$ and covariance matrix $I\sigma^2$, then the scoring iteration for θ specializes to the well known Gauss-Newton iteration for solving nonlinear least squares problems

$$\theta^{k+1} = \theta^k + (\dot{\mu}^T\dot{\mu})^{-1}\dot{\mu}^T(y - \mu)$$

where $\dot{\mu}$ is the gradient matrix of partial derivatives $\partial\mu_i/\partial\theta_j$. Letting $(\dot{\mu}_1\,\dot{\mu}_2)$ be the partition of $\dot{\mu}$ into derivatives with respect to θ_1 and θ_2 respectively, the nested Gauss-Newton iteration is

$$\theta_2^{k+1} = \theta_2^k + \left[\dot{\mu}_2^T(I - P_1)\dot{\mu}_2\right]^{-1}\dot{\mu}_2^T(y - \mu)$$

where $P_1 = \dot{\mu}_1(\dot{\mu}_1^T\dot{\mu}_1)^{-1}\dot{\mu}_1^T$ is the orthogonal projection onto the column space of $\dot{\mu}_1$ and all terms are evaluated at $\theta_1 = \hat{\theta}_1(\theta_2)$. For simplicity, $\dot{\mu}$ is assumed to be of full rank in the parameter region of interest; otherwise generalized inverses replace inverses. The nested iteration is most useful when θ_1 can be chosen to be conditionally linear, i.e., $\mu = X(\theta_2)\theta_1$ where X is a known matrix function of θ_2, so that $\dot{\mu}_1 = X$ and $\hat{\theta}_1(\theta_2) = (X^TX)^{-1}X^Ty$ is available in closed form.

The use of the closed form expression for $\hat{\theta}_1(\theta_2)$ to concentrate the least squares estimation problem on θ_2 is known as separable regression in the numerical literature. Although the idea of eliminating the linear parameters seems to go back to

the earliest treatments on nonlinear models in regression (Hartley, 1948; Stevens, 1951; Pimentel-Gomes, 1953), the earliest explicit substitution of $\hat{\theta}_1(\theta_2)$ into the sum of squares seems to be Richards (1961). The first example of the above nested Gauss-Newton iteration is Walling (1968). Barham and Drane (1972) applied it to a variety of examples and Kaufman (1975) gave detailed calculations. A closely related algorithm is given by Golub and Pereyra (1973, 1976), who also give a detailed treatment of the case when $X(\theta_2)$ is not of full rank. Ruhe and Wedin (1980) were the first to consider nested Gauss-Newton for μ not necessarily linear in θ_1.

Although an old idea, separable regression seems to be less used in practice than it might be. Its use seems to have been delayed by the computation burden in differentiating $S(\hat{\theta}_1(\theta_2), \theta_2)$ with respect to θ_2. See Golub and Pereyra (1973, 1976) and Harville (1973). The nested Gauss-Newton iteration has the attraction that it avoids the need for this derivative. Ruhe and Wedin (1980) show that nested Gauss-Newton has the same computation count per iteration as does the unseparated algorithm, and that it is the simplest scheme for which convergence is almost quadratic when σ^2 is small.

Example 4.1. Many nonlinear functions that are fitted to data by least squares arise as solutions to homogeneous differential equations. Typically the systematic component of the process, $\mu(t)$ say, satisfies

$$\sum_{k=0}^{q} c_k(t;\beta)\frac{\partial^k \mu(t)}{\partial t^k} = 0$$

in which the c_k are coefficients and β is a vector of unknown parameters. If $g_j(t;\beta)$, $j = 1, \ldots, q$ are distinct special solutions for μ, then the general solution is of the form

$$\mu(t;\alpha,\beta) = \sum_{j=1}^{q} \alpha_j g_j(t;\beta)$$

The empirical values y_i observed for the process at times t_i are assumed to be independent with means $\mu_i = \mu(t_i;\alpha,\beta)$, producing a separable regression problem in which the α_j are the linear parameters.

Example 4.2. In the previous example, if the c_k are constant, i.e., do not depend on t, and the polynomial with coefficients c_k has distinct real roots, then the differential equation has general solution

$$\mu(t;\alpha,\beta) = \sum_{j=1}^{q} \alpha_j \exp(\beta_j t)$$

where the α_j are arbitrary and the β_j are the roots of the polynomial. Many of the early nonlinear regression problems in the literature were of this form. Osborne and Smyth (1991) show how to express this model as a separable regression with the c_k themselves as the nonlinear parameters, and this is important when, as is usually the case, complex or repeated roots for the polynomial cannot be ruled out.

Example 4.3. Stevens (1951) considered nonlinear regression with $\mu_i = \alpha + \beta\rho^{x_i}$, and proposed a separable regression iteration which is an leapfrog rather than a nested iteration. He showed that the Gauss-Newton iteration produces updated estimates α^{k+1} and β^{k+1} which depend on ρ^k but not on α^k or β^k, and hence recommended a modified iteration which does not require starting values for α or β. In our terminology, Steven's iteration is the leapfrog Gauss-Newton iteration with $\theta_1 = (\alpha, \beta)^T$ and $\theta_2 = \rho$. The update function F_1 for θ_1 depends only on θ_2, so the leapfrog iteration $\theta_2^{k+1} = F_2(F_1(\theta_2^k), \theta_2^k)$ effectively eliminates the linear parameters. This result generalizes to any separable regression in which each component of θ_2 appears in at most one column of $X(\theta_2)$, for example

$$\mu = \alpha_1 g_1(x; \beta_1) + \ldots + \alpha_p g_p(x; \beta_p)$$

where the α_j and β_j are scalar parameters, and to some other cases. See also Ross (1990, Section 5.4). It can be shown that the complete characterization of regression models for which $F_1(\theta_1, \theta_2)$ does not depend on θ_1 is very similar to Khuri's (1984) condition that the optimal subset design for θ_1 not depend on θ_1.

4.2 Generalized linear models

The ideas of separable regression extend to generalized linear models. Suppose that each y_i is an observation from some generalized linear model distribution with mean μ_i, $i = 1, \ldots, n$ (McCullagh and Nelder, 1989; Jorgensen, 1987). Then $\text{var}(y_i) = \phi v(\mu_i)$, where $v(\cdot)$ is a known function which characterizes the distribution and ϕ is a proportionality constant. The μ_i are assumed to depend on known covariates x_{ij} and unknown regression parameters β_j. The score vector is

$$\dot{\ell} = \dot{\mu}^T V^{-1}(y - \mu)$$

where as usual $\dot{\mu}$ is the gradient matrix of derivatives $\partial\mu_i/\partial\beta_j$ and $V = \text{diag}[v(\mu_i)]$. The expected information is

$$\mathcal{I} = \dot{\mu}^T V^{-1} \dot{\mu}$$

Now let θ_1, θ_2 be a partition of β, and let $\dot{\mu}_1$ and $\dot{\mu}_2$ be the corresponding partition of $\dot{\mu}$. We have

$$\dot{\ell}_2 = \dot{\mu}_2^T V^{-1}(y - \mu)$$

and

$$\mathcal{I}_{2.1} = \dot{\mu}_2^T V^{-1} \dot{\mu}_2 - \dot{\mu}_2^T V^{-1} \dot{\mu}_1 \left(\dot{\mu}_1^T V^{-1} \dot{\mu}_1\right)^{-1} \dot{\mu}_1^T V^{-1} \dot{\mu}_2 = \dot{\mu}_{2.1}^T V^{-1} \dot{\mu}_{2.1}$$

where

$$\dot{\mu}_{2.1} = \dot{\mu}_2 - \dot{\mu}_1 \left(\dot{\mu}_1^T V^{-1} \dot{\mu}_1\right)^{-1} \dot{\mu}_1^T V^{-1} \dot{\mu}_2$$

The nested scoring iteration for θ_2 may be written

$$\theta_2^{k+1} = \left(\dot{\mu}_{2.1}^T V^{-1} \dot{\mu}_{2.1}\right)^{-1} \dot{\mu}_{2.1}^T V^{-1}(y - \mu + \dot{\mu}_{2.1}\theta_2^k)$$

where the right hand side is evaluated at $\theta_1 = \hat{\theta}_1(\theta_2^k)$. Here $\dot{\mu}_{2.1}$ is the residual vector from regression of $\dot{\mu}_2$ on $\dot{\mu}_1$ with weight matrix V^{-1}. The nested iteration may be performed by two weighted linear regressions, one of $\dot{\mu}_2$ on $\dot{\mu}_1$ and the other of $z = y - \mu + \dot{\mu}_{2.1}\theta_2$ on $\dot{\mu}_{2.1}$.

Example 4.4. Now suppose that $\mu_i = h(\eta_i;\gamma)$, where $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, or equivalently $g(\mu_i;\gamma) = \eta_i$ where $g(\cdot;\gamma)$ is the inverse of $h(\cdot;\gamma)$. Here g is a parametric link function as studied by Pregibon (1980), Scallan (1982) and many others. In this case

$$\dot{\mu}_1 = \mathrm{diag}(\frac{\partial h}{\partial \eta})X, \quad \dot{\mu}_2 = \frac{\partial h}{\partial \gamma}$$

The nested iteration may be programmed in for example GLIM, and is closely related to but simpler than the algorithm given by Scallan, Gilchrist and Green (1984).

5 Local Convergence

5.1 Contraction factors

In this section we look at the limiting rate of convergence as an iteration approaches a stationary value. For Newton-type algorithms this turns out to depend on the accuracy with which the observed information is approximated. For the EM algorithm it is determined by the proportion of incomplete to complete data.

Consider a general iterative process defined by

$$\theta^{k+1} = F(\theta^k)$$

with a stationary point at the desired estimate $\hat{\theta}$. A first order Taylor series gives

$$\begin{aligned} \theta^{k+1} - \hat{\theta} &= F(\theta^k) - \hat{\theta} \\ &= F(\hat{\theta}) + \dot{F}(\tilde{\theta})(\theta^k - \hat{\theta}) - \hat{\theta} \\ &= \dot{F}(\tilde{\theta})(\theta^k - \hat{\theta}) \\ &\approx \dot{F}(\hat{\theta})(\theta^k - \hat{\theta}) \end{aligned}$$

where $\tilde{\theta}$ is a point on the line between θ^k and $\hat{\theta}$ and $\dot{F}$ is the $p \times p$ iteration derivative with components $\partial F_i/\partial\theta_j$. So the iteration behaves locally as a multivariate geometric series, and will converge to $\hat{\theta}$ from some neighbourhood if the matrix $G = \dot{F}(\hat{\theta})$ defines a contraction mapping. In fact, sufficiently close to the stationary value, the largest absolute eigenvalue R of G will dominate, and it is a sufficient condition for the iteration to have a point of attraction at $\hat{\theta}$ that R be less than one. Except for indeterminacy at $R = 1$, this condition is necessary as well as sufficient (Ostrowski, 1960, Section 4.2; Ortega and Rheinboldt, 1970, Section 10.1). If the algorithm converges, R is the limiting contraction factor $\lim_{k\to\infty} \|\theta^{k+1} - \hat{\theta}\|/\|\theta^k - \hat{\theta}\|$.

5.2 A Nested EM Algorithm

Many statistical estimation problems can be viewed as incomplete data problems so that application can be made of the EM algorithm of Dempster *et al* (1977). The EM algorithm is not a partitioned algorithm in that it does not reduce the dimension of the parameter space. Instead it imbeds the original problem in a

complete data problem which, although still of the same dimension, is simpler in form. Dempster *et al* (1977) showed that iteration derivative of the EM algorithm is determined by the ratio of the incomplete to complete information matrices, i.e., the iteration derivative is $G = \mathcal{I}_{\rm u}(\hat{\theta})\mathcal{I}_c^{-1}(\hat{\theta})$ where $\mathcal{I}_{\rm c}$ is the Fisher information for the complete data and $\mathcal{I}_{\rm u}$ is the Fisher information for the unobserved component. Convergence of the EM algorithm can therefore be accelerated by separating out from the iteration those parameters about which there is relatively least information in the incomplete data.

Let the EM iteration for a parameter vector $\theta = (\theta_1, \theta_2)$ be defined by

$$\begin{aligned} \theta_1^{k+1} &= F_1(\theta_1^k, \theta_2^k) \\ \theta_2^{k+1} &= F_2(\theta_1^{k+1}, \theta_2^k) \end{aligned}$$

We define a nested EM algorithm, with θ_1 nested within θ_2, to be the iteration

$$\theta_2^{k+1} = F_2(\tilde{\theta}_1(\theta_2^k), \theta_2^k)$$

where $\tilde{\theta}_1(\theta_2)$ is the solution with respect to θ_1 of $\theta_1 = F_1(\theta_1, \theta_2)$, assumed to be unique. The contract factor of the nested algorithm is $\mathcal{I}_{\rm u2}(\hat{\theta})\mathcal{I}_{c2}^{-1}(\hat{\theta})$ where $\mathcal{I}_{\rm u2}$ and $\mathcal{I}_{c2}$ are the information matrices for the non-nested parameters. Further improvement can often be obtained by applying Aitken acceleration, which converts linear into quadratic convergence, to the nested algorithm.

Example 5.1. Let $y = (y_1, y_2, y_3)^T$ be trivariate normal with mean μ and covariance Σ. Suppose that the experimental design is such that (y_1, y_2) is observed n_1 times, (y_1, y_3) is observed n_2 times, but y_2 and y_3 are never observed together. Then the data contains zero information about σ_{23}, the covariance of y_2 and y_3. All values of σ_{23} for which Σ is positive definite yield the same value of the likelihood. If the complete data are taken to consist of $n_1 + n_2$ trivariate observations, then $\mathcal{I}_{\rm u}(\hat{\theta})\mathcal{I}_c^{-1}(\hat{\theta})$ has an eigenvalue equal to one corresponding to σ_{23}. The convergence of the EM algorithm is, ultimately, infinitely slow.

For this problem the stationary value $\tilde{\sigma}_{23}$ is that value of σ_{23} which makes the corresponding component of Σ^{-1} zero. Equivalently, $\tilde{\sigma}_{23}$ is the midpoint of the interval of values for σ_{23} which make Σ positive definite. An algorithm for determining such values for covariance matrices of arbitrary dimension is given by Wermuth and Scheidt (1977). The nested EM algorithm with σ_{23} set to $\tilde{\sigma}_{23}$ at each iteration converges linearly with contraction factor $R = \max(n_1, n_2)/(n_1 + n_2)$.

5.3 Scoring

The convergence of scoring and other Newton-type algorithms depends on two approximations: the quadratic approximation to the log-likelihood function, and the approximation of observed with expected information. In this section we look at the information approximation and find that it affects the convergence rate near the solution. In Section 6 we look at the quadratic approximation and find that it affects convergence from any starting value.

The iteration derivative for a Newton-type iteration can be found to be

$$G = A^{-1}(\ddot{\ell} + A) \tag{3}$$

This is the relative difference between $-\ddot{\ell}$ and A at $\hat{\theta}$, and is a measure of the accuracy with which the observed information is approximated. In the case of least squares this specializes to

$$G = (\dot{\mu}^T \dot{\mu})^{-1} \ddot{\mu}^T (y - \mu)$$

for the Gauss-Newton iteration with identity weight matrix. Here $\ddot{\mu}^T(y - \mu)$ is the $p \times p$ matrix with j, kth element $\sum_{i=1}^{n} (\partial \ddot{\mu}_i / \partial \theta_j \partial \theta_k)(y_i - \mu_i)$. Convergence of the Gauss-Newton algorithm depends therefore on the size of the residual vector in the direction of $\ddot{\mu}$. See also Varah (1991) on this point.

Note that, although G is not in general symmetric, it will have all real eigenvalues if A is symmetric. In that case convergence will be ultimately monotonic if R is achieved by a positive eigenvalue of G and will be ultimately oscillatory is R is achieved by a negative eigenvalue.

Several properties of the scoring iteration derivative are encapsulated in three small theorems below. The first is an expression of the well known property that the scoring iteration tends to converge rapidly when the number of observations is large relative to the number of parameters. This theorem is given without proof, but is an application of the law of large numbers and assumes standard regularity conditions. A rigorous proof of strong convergence to zero in the case of least squares is given by Jennrich (1969). The second is that the spectrum of the scoring iteration derivative depends on the shape of the statistical model, not on the particular parametrization. This is a surprising result given empirical evidence (Ratkowsky, 1983; Ross, 1990) that reparametrization can make scoring less sensitive to starting values. The theorem implies that such benefits are a result of improving the quadratic approximation to the log-likelihood function as discussed in Section 6 rather than of improving the approximation of observed by expected information. The third theorem shows that if the scoring iteration diverges from close to a maximum, it is likely to do so in an oscillatory manner. This and subsequent proofs make use of Rayleigh quotients which are described by Golub and van Loan (1983, page 308).

Theorem 1 *If Fisher information is proportional to the sample size n so that $\lim_{n\to\infty} n^{-1}\mathcal{I}$ is positive definite at the true parameter value, then the components of the iteration derivative for scoring are $O(n^{-1/2})$.*

Theorem 2 *The ultimate rate of convergence of the scoring algorithm cannot be improved by reparametrization.*

Proof. Let J be the Jacobian of the reparametrization. The new Fisher information and observed information matrices can be expressed in terms of the old as $J^T \mathcal{I} J$ and $J(\hat{\theta})^T \ddot{\ell}(\hat{\theta}) J(\hat{\theta})$ respectively. Hence the new iteration derivative can be related to the old through the similarity transformation

$$J^{-1}\mathcal{I}^{-1}(\ddot{\ell} + \mathcal{I})J$$

which leaves its spectrum unchanged. □

Theorem 3 *The spectrum of the iteration derivative for scoring is bounded above by one, strictly so if $\ddot{\ell}(\hat{\theta})$ is negative definite.*

Proof. The iteration derivative for scoring has Rayleigh quotient

$$1 + \frac{z^T \ddot{\ell} z}{z^T \mathcal{I} z}$$

which is less than one for all z if $\ddot{\ell}$ is negative definite. □

5.4 Partitioned Scoring

The iteration derivative for nested scoring is obtained by differentiating $F(\theta_2) = \theta_2 + \mathcal{I}^{-1}\dot{\ell}_2(\hat{\theta}_1(\theta_2), \theta_2)$ with respect to θ_2, and observing that $\partial\hat{\theta}_1(\theta_2)/\partial\theta_2 = -\ddot{\ell}_1^{-1}\ddot{\ell}_{12}$. The iteration derivative is

$$G_n = \mathcal{I}_{2.1}^{-1}(\ddot{\ell}_{2.1} + \mathcal{I}_{2.1}) \tag{4}$$

which is the relative difference between observed and expected information for θ_2 adjusted for θ_1.

Similarly the iteration derivative for zigzag maximum likelihood emerges from differentiating $\theta_2 = \hat{\theta}_2(\hat{\theta}_1(\theta_2))$ as

$$G_z = \ddot{\ell}_2^{-1}\ddot{\ell}_{21}\ddot{\ell}_1^{-1}\ddot{\ell}_{12}. \tag{5}$$

This type of expression is familiar from canonical correlation analysis. If $-\ddot{\ell}$ is viewed as a partitioned covariance matrix, G_z is a measure of the strength of relationship between the two groups of variables.

The iteration derivative for leapfrog scoring is slightly less straightforward. If the convergence matrix for the full scoring iteration is written as

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial \theta_1} & \frac{\partial F_1}{\partial \theta_2} \\ \frac{\partial F_2}{\partial \theta_1} & \frac{\partial F_2}{\partial \theta_2} \end{pmatrix}$$

then the leapfrog iteration derivative matrix is

$$G_l = \begin{pmatrix} G_1 & G_{12} \\ G_{21}G_1 & G_{21}G_{12} + G_2 \end{pmatrix}.$$

Some conclusions for large n can immediately be drawn. Nested and leapfrog iteration derivatives simply inherit the convergence rates given in Theorem 1, while for zigzag iterations orthogonality is everything.

Corollary 1 *The elements of the iteration derivatives for nested and leapfrog scoring are also $O(n^{-1/2})$. For zigzag iteration derivative is $O(1)$ unless θ_1 and θ_2 are orthogonal, in which case it is $O(n^{-1})$.*

Note that if the elements of G are $O(n^{-1/2})$ then $G_{21}G_1$ and $G_{21}G_{12}$ are $O(n^{-1})$. So for large n, G_l is approximately equal to

$$\begin{pmatrix} G_1 & G_{12} \\ 0 & G_2 \end{pmatrix}$$

the eigenvalues of which are those of G_1 and G_2. For large n therefore, the contraction factor for leapfrog iterations is simply the maximum of the contraction factors for scoring for θ_1 and θ_2 alone, and is therefore less than or equal to that of the full scoring iteration.

If θ_1 and θ_2 are orthogonal, and the scoring iterations for θ_1 and θ_2 separately are convergent, then G_n, G_z and G_l are special cases of a more general expression. For each k, let the sequence $\{\theta_1^{k,j} : j \geq 0\}$ be defined by $\theta_1^{k,0} = \theta_1^k$ and $\theta_1^{k,j+1} = F_1(\theta_1^{k,j}, \theta_2^k)$, and consider the process

$$\begin{aligned} \theta_1^{k+1} &= \theta_1^{k,m} \\ \theta_2^{k+1} &= F_2(\theta_1^{k+1}, \theta_2^k) \end{aligned}$$

for different values of m. This corresponds to updating θ_1 m times before turning to θ_2. The iteration derivative of this process is

$$\begin{pmatrix} G_1^m & (I - G_1)^{-1}(I - G_1^m)G_{12} \\ G_{21}G_1^m & G_{21}(I - G_1)^{-1}(I - G_1^m)G_{12} + G_2 \end{pmatrix}$$

which reduces to G_l for $m = 1$ and to G_n for $m = \infty$. Zigzag iterations could be incorporated also by considering repeated updating of θ_2. Choosing $m > 1$ in the above iteration is likely to be worthwhile if θ_1^k is converging more slowly than θ_2^k.

We now consider how partitioned algorithms may benefit the limiting convergence rate for a given sample. Nesting is found to generally improve the accuracy with which observed information is approximated. The zigzag iteration is found to admit a global convergence result. An ideal situation for partitioned algorithms turns out to be that in which θ_1 and θ_2 are orthogonal with exponential family submodels.

Theorem 4 *The contraction factor for nested scoring is bounded above by that for the full scoring iteration.*

Proof. We show that the extreme eigenvalues of $\mathcal{I}_{2.1}^{-1}\ddot{\ell}_{2.1}$ are bounded by those of $\mathcal{I}^{-1}\ddot{\ell}$. This is sufficient for the result, since adding identity matrices to both expressions simply increments all the eigenvalues without altering the relative result.

Let P be the $p \times p_2$ matrix $(0\ I)^T$. Then $\mathcal{I}_{2.1}^{-1} = P^T\mathcal{I}^{-1}P$ and $\ddot{\ell}_{2.1}^{-1} = P^T\ddot{\ell}^{-1}P$, so $\mathcal{I}_{2.1}^{-1}\ddot{\ell}_{2.1}$ has Rayleigh quotient

$$\frac{z^T P^T \mathcal{I}^{-1} P z}{z^T P^T \ddot{\ell}^{-1} P z} \tag{6}$$

for $z \in \mathbb{R}^{p_2}$. The extrema of (6) are constrained extrema of

$$\frac{v^T \mathcal{I}^{-1} v}{v^T \ddot{\ell}^{-1} v}$$

or equivalently of

$$\frac{v^T \ddot{\ell} v}{v^T \mathcal{I} v} \tag{7}$$

over $v \in \mathbb{R}^p$, and hence are bounded by the unconstrained extrema. Observing that (7) is the Rayleigh quotient of $\mathcal{I}^{-1}\ddot{\ell}$ completes the proof. □

This generalizes some of the results of Ruhe and Wedin (1980) who examined contraction factors for several separable least squares algorithms including nested Gauss-Newton and the variable projection algorithm of Golub and Pereyra (1973). Both algorithms were found to have contraction factors bounded by that of the unseparated Gauss-Newton iteration. In practice the contraction factors of nested and full scoring iterations are often very similar, because the nesting is applied to parameters which are conditionally easy to estimate, as in separable regression. A greater reduction in the contraction factor could be achieved by applying nesting to the most nonlinear parameters. This reduction would, however, have to be balanced against the increase in computation per iteration.

Example 5.2. As a numerical example of nesting, consider the following simulated data set to which the rational function $y = (1+\theta_1 x)/(1+\theta_2 x^2)$ was fitted by least squares. The problem has been chosen small to make plotting of the iteration function possible.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y	0.8280	0.5232	0.5510	0.6087	0.3365	0.3150	0.1629	0.2490	−0.0330	0.1965

The least squares estimates are $\hat{\theta}_1 = -0.6900$ and $\hat{\theta}_2 = 3.4055$. The contraction factor for Gauss-Newton is -0.49 while that for nested Gauss-Newton with θ_1 nested within θ_2 is -0.20. Figure 1 plots $F(\theta_2)$ against θ_2 for nested Gauss-Newton, so that the contraction factor is represented by the slope of the curve at $\theta_2 = \hat{\theta}_2$. Also given is a plot of $F(\theta_1, \theta_2)$ versus θ_2 for the Gauss-Newton algorithm, where θ_1 is chosen so that $(\theta_1, \theta_2)^T$ lies in the dominant eigenspace of the iteration derivative, as it would in practice after repeated iteration. The plot shows that the nested iteration converges from starting values between about -1 and 8, while the full iteration converges from values between about -1 and 6.

Example 5.3. It can be shown that for regression models of the Stevens type, discussed in Example 4.3, the contraction factors for nested and complete Gauss-Newton are identical. This includes for example exponential function fitting as in Example 4.2. Practical benefits do often come from the nested Gauss-Newton algorithm for such models, but this is related to the quadratic approximation to the log-likelihood rather than to the difference between observed and expected information; see Section 6.

Theorem 5 *Zigzag iterations have a point of attraction at any maxima of the likelihood for which $-\ddot{\ell}$ is positive definite. Furthermore, the convergence is monotonic.*

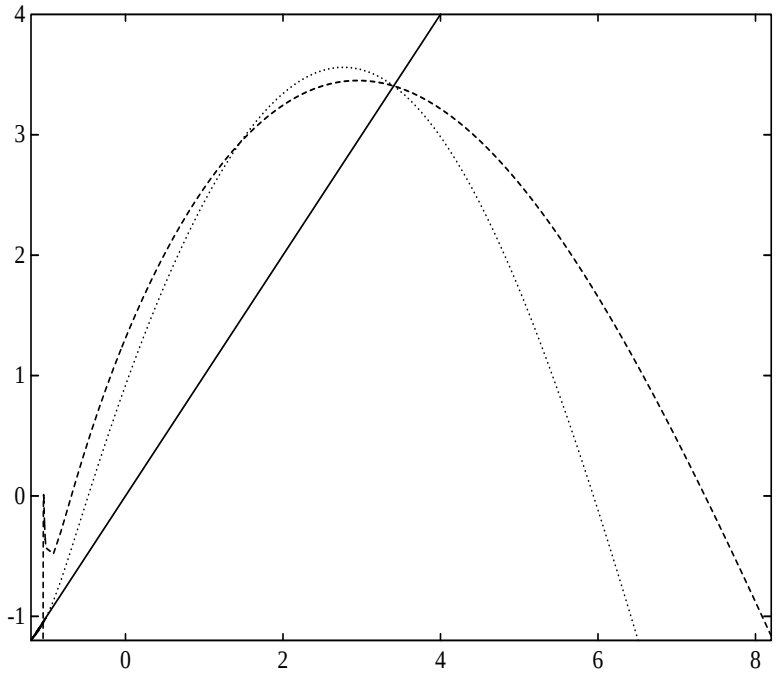

Figure 1: Plot of θ_2 versus $F(\theta_2)$, where F is an iteration function, for the data of Example 5.2. Dashed line is for nested Gauss-Newton, dotted line is for full Gauss-Newton with θ_1 chosen so that $(\theta_1, \theta_2)^T$ lies in the dominant eigenspace of $\dot{F}$.

Proof. If $-\ddot{\ell}$ is positive definite, then the eigenvalues of (5) may be recognized as the squared canonical correlations calculated from the partitioned covariance matrix

$$\begin{pmatrix} -\ddot{\ell}_{11} & -\ddot{\ell}_{12} \\ -\ddot{\ell}_{21} & -\ddot{\ell}_{22} \end{pmatrix}.$$

All the eigenvalues, therefore, lie between 0 and 1 (Rao, 1973, Section 8f.1). □

Zigzag iterations have the advantage that the likelihood is non-decreasing with each iteration. This allows a global convergence result, given by Ortega and Rheinboldt (1970, page 516) as the Global SOR Theorem. This has been exploited by Jensen, Johansen and Lauritzen (1991) to give a globally convergent algorithm for maximizing exponential family likelihoods.

Example 5.4. Consider the zigzag iteration for β_1 and β_2 in the multiple linear regression model $E(Y_i) = \beta_1 x_{1i} + \beta_2 x_{2i}$. The contraction factor for this iteration is the squared correlation between x_1 and x_2.

Example 5.5. As an example of both the reliability of the zigzag iteration and its slowness for nonorthogonal parameters, consider the application of the method described in Example 3.2 to data from Kowalik and Osborne (1968, page 104). The model is $E(Y_i) = (\alpha_1 x_i^2 + \alpha_2 x_i)/(x_i^2 + \beta_1 x_i + \beta_2)$, and there are 11 observations. This problem is difficult for Newton methods in that it requires superb starting values

for β_1 and β_2. Instead, cycling between the numerator and denominator models as described in Example 3.2 produces an algorithm which converges from the usual generalized linear model starting values, although very slowly. Convergence in the sum of squares to 5 significant figures required about 300 iterations for the author's implementation in GLIM. Parameter values suitable as starting values for Gauss-Newton were obtained far more quickly.

The final two theorems give circumstances in which leapfrog scoring, which is the least expensive iteration, is a suitable substitute for nested scoring.

Theorem 6 *If θ_1 and θ_2 are orthogonal, and the scoring iteration for θ_1 with θ_2 fixed is quadratically convergent, then leapfrog and nested scoring have the same contraction factor.*

Proof. If $\mathcal{I}_{12} = 0$ and $\ddot{\ell}_1 + \mathcal{I}_1 = 0$, then the iteration derivative for leapfrog scoring is

$$\begin{pmatrix} 0 & G_{12} \\ 0 & G_{21}G_{12} + G_2 \end{pmatrix} = \begin{pmatrix} 0 & \mathcal{I}_1^{-1}\ddot{\ell}_{12} \\ 0 & \mathcal{I}_2^{-1}\ddot{\ell}_{21}\mathcal{I}_1^{-1}\ddot{\ell}_{12} + \mathcal{I}_2^{-1}(\ddot{\ell}_2 + \mathcal{I}_2) \end{pmatrix}$$

the eigenvalues of which are those of

$$\mathcal{I}_2^{-1}\ddot{\ell}_{21}\mathcal{I}_1^{-1}\ddot{\ell}_{12} + \mathcal{I}_2^{-1}(\ddot{\ell}_2 + \mathcal{I}_2) = \mathcal{I}_2^{-1}(\ddot{\ell}_{2.1} + \mathcal{I}_2).$$

□

Example 5.6. Let $y_1, \ldots, y_n$ be independent, and let y_i have a gamma distribution with mean μ_i and variance $\phi_i\mu_i^2$, where

$$\frac{1}{\mu_i} = x_i^T\beta$$

and

$$\log \phi_i = z_i^T\gamma$$

where the x_i and z_i are known vectors of covariates. Then β and γ are orthogonal, and the scoring iteration for β alone is quadratically convergent since the reciprocal link function is canonical for the gamma family. So leapfrog scoring for β and γ has the same contraction factor as nested scoring for γ.

The final theorem shows that, under ideal circumstances, one iteration of any of the partitioned algorithms is equivalent to two iterations of the full scoring algorithm.

Theorem 7 *If θ_1 and θ_2 are orthogonal, and the scoring iterations for θ_1 and θ_2 separately are quadratically convergent, then leapfrog, nested and zigzag scoring all have the same contraction factor, which is the square of that of the full scoring iteration.*

Proof. The iteration derivative of the full scoring iteration is

$$G = \begin{pmatrix} 0 & \mathcal{I}_1^{-1}\dot{\ell}_{12} \\ \mathcal{I}_2^{-1}\ddot{\ell}_{21} & 0 \end{pmatrix},$$

the eigenvalues of which are the same as of those of

$$G = \begin{pmatrix} 0 & \mathcal{I}_1^{-1/2}\ddot{\ell}_{12}\mathcal{I}_2^{-T/2} \\ \mathcal{I}_2^{-1/2}\ddot{\ell}_{21}\mathcal{I}_1^{-T/2} & 0 \end{pmatrix}$$

which are the singular values of $\mathcal{I}_1^{-1/2}\ddot{\ell}_{12}\mathcal{I}_2^{-T/2}$. The iteration derivative for leapfrog scoring is

$$\begin{pmatrix} 0 & \mathcal{I}_1^{-1}\ddot{\ell}_{12} \\ 0 & \mathcal{I}_2^{-1}\ddot{\ell}_{21}\mathcal{I}_1^{-1}\ddot{\ell}_{12} \end{pmatrix}$$

the eigenvalues of which are those of $\mathcal{I}_2^{-1}\ddot{\ell}_{21}\mathcal{I}_1^{-1}\ddot{\ell}_{12}$ or of $\mathcal{I}_2^{-1/2}\ddot{\ell}_{21}\mathcal{I}_1^{-1}\ddot{\ell}_{12}\mathcal{I}_2^{-T/2}$, that is the squares of the singular values of $\mathcal{I}_1^{-1/2}\ddot{\ell}_{12}\mathcal{I}_2^{-T/2}$. The corresponding results for leapfrog and nested scoring follow directly by substituting $\mathcal{I}_{12} = 0$, $\mathcal{I}_1 = -\ddot{\ell}_1$ and $\mathcal{I}_2 = -\ddot{\ell}_2$ into (5) and (4). □

Example 5.7. Let y_i be inverse-Gaussian, with mean μ_i and variance $\phi_i\mu_i^3$, where

$$\frac{1}{\mu_i^2} = x_i^T\beta$$

and

$$\frac{1}{\phi_i} = z^T\gamma.$$

In this case, β and γ are orthogonal, and the scoring iterations for θ_1 or θ_2 separately are quadratically convergent. One iteration of nested, zigzag or leapfrog scoring for β and γ is equivalent in this case to two iterations of the full scoring algorithm.

6 The Quadratic Approximation to the Log-likelihood

Unlike the EM algorithm, Newton-type algorithms often take very many iterations to asymptote to their limiting contraction factors, and to understand their behaviour on practical problems it is necessary to study convergence more globally. An ideal iteration function F, which converges to $\hat{\theta}$ in one step from any starting value, has a derivative $\dot{F}$ which is identically zero. The Newton-Raphson iteration, being quadratically convergent, has $\dot{F} = 0$ at $\theta = \hat{\theta}$, but for the iteration to converge reliably in practice it is necessary that $\dot{F}$ remain small as far from $\hat{\theta}$ as possible. More precisely, $F(\theta)$ will be closer to $\hat{\theta}$ than θ if the average derivative $\dot{F}$ between θ and $\hat{\theta}$ has spectral radius less than one.

The iteration derivative for Newton-Raphson at an arbitrary point involves the third derivative of the log-likelihood, which will be expressed as a trilinear form,

$\dddot{\ell}\,[\cdot,\cdot,\cdot]$, in notation similar to Dieudonné (1960). If u, v and w are vectors then $\dddot{\ell}\,[u,v,w]$ is the scalar $\sum_{ijk}\ \dddot{\ell}_{ijk}\ u_i v_j w_k$, where the $\dddot{\ell}_{ijk}$ are the individual partial derivatives, and, for example, $\ddot{\ell}\,[u,\cdot,\cdot]$ is a bilinear form which can be identified with a matrix. The iteration derivative for Newton-Raphson is

$$\dot{F} = \ddot{\ell}^{-1}\frac{\partial \ddot{\ell}}{\partial \theta}\ddot{\ell}^{-1}\dot{\ell} = \ddot{\ell}\,[\ddot{\ell}^{-1}, \ddot{\ell}^{-1}\dot{\ell}, \cdot]$$

i.e., the matrix whose (a,b) element is $\sum_{ijk}\ddot{\ell}^{ai}\ \dddot{\ell}_{ijb}\ \ddot{\ell}^{jk}\dot{\ell}_k$, where $\ddot{\ell}^{ai}$, $\dddot{\ell}_{ijb}$ and $\dot{\ell}_k$ are components of $\ddot{\ell}^{-1}$, $\dddot{\ell}$ and $\dot{\ell}$ respectively. The largest eigenvalue of this matrix is a measure of how effective the quadratic approximation to the log-likelihood is at this point. The matrix is always zero at $\theta = \hat{\theta}$, reflecting the fact that the quadratic approximation is always satisfactory sufficiently close to $\hat{\theta}$.

Now consider a reparametrization to ϕ defined by $\theta = H\phi$, where H is a non-singular constant matrix. We show that the eigenvalues of the iteration derivative are invariant with respect to such a linear reparametrization. The Newton-Raphson iteration in terms of ϕ, $F_\phi(\phi) = \phi - \ddot{\ell}_\phi^{-1}\dot{\ell}_\phi$ has derivative

$$\dot{F}_\phi = \ddot{\ell}_\phi\,[\ddot{\ell}_\phi^{-1}, \ddot{\ell}_\phi^{-1}\dot{\ell}_\phi, \cdot]$$

where $\dot{\ell}_\phi$, $\ddot{\ell}_\phi$ and $\dddot{\ell}_\phi$ are the derivatives of the log-likelihood with respect to ϕ. Now $\dot{\ell}_\phi = H^T\dot{\ell}$, $\ddot{\ell}_\phi = H^T\ddot{\ell}H$ and $\dddot{\ell}_\phi\,[\cdot,\cdot,\cdot] = \dddot{\ell}\,[H,H,H]$, so

$$\dot{F}_\phi = \ddot{\ell}_\phi\,[H^{-1}\ddot{\ell}^{-1}H^{-T}, H^{-1}\ddot{\ell}^{-1}H^{-T}H^T\dot{\ell}, \cdot] = \ddot{\ell}_\phi\,[H^{-1}\ddot{\ell}^{-1}H^{-T}, H^{-1}\ddot{\ell}^{-1}\dot{\ell}, \cdot]$$

$$= \dddot{\ell}\,[\ddot{\ell}^{-1}H^{-T}, \ddot{\ell}^{-1}\dot{\ell}, H]$$

which is a bilinear form corresponding to the matrix $H^{-1}\dot{F}H$. The iteration derivatives for the original and transformed parameters are related through a similarity relation.

We are now in a position to prove Theorem 8, which shows that nesting improves globally the rate of convergence of the Newton-Raphson algorithm, at least close to the locus $\theta_1 = \hat{\theta}_1(\theta_2)$. Nesting improves the quadratic approximation to the log-likelihood by reducing the number of dimensions over which inaccuracies may manifest themselves.

Theorem 8 *The eigenvalues of the iteration derivative for nested Newton-Raphson are bounded by those of the iteration derivative for the full Newton-Raphson iteration at $\theta_1 = \hat{\theta}_1(\theta_2)$.*

Proof. The nested Newton-Raphson iteration is

$$F_n(\theta_2) = \theta_2 - \ddot{\ell}_{2.1}^{-1}(\hat{\theta}_1(\theta_2), \theta_2)\dot{\ell}_2(\hat{\theta}_1(\theta_2), \theta_2)$$

which has the derivative with respect to θ_2

$$\dot{F}_n = \ddot{\ell}_{2.1}^{-1}\frac{\partial \ddot{\ell}_{2.1}}{\partial \theta_2}\ddot{\ell}_{2.1}^{-1}\dot{\ell}_2$$

Now

$$\frac{\partial \ddot{\ell}_{2.1}}{\partial \theta_2} = \frac{\partial \ddot{\ell}_{22}}{\partial \theta_2} - \frac{\partial \ddot{\ell}_{21}}{\partial \theta_2}\ddot{\ell}^{-1}\ddot{\ell}_{12} + \ddot{\ell}_{21}\ddot{\ell}_{11}^{-1}\frac{\partial \ddot{\ell}_{11}}{\partial \theta_2}\ddot{\ell}_{11}^{-1}\ddot{\ell}_{12} - \ddot{\ell}_{21}\ddot{\ell}_{11}^{-1}\frac{\partial \ddot{\ell}_{12}}{\partial \theta_2}$$

Writing $K = \partial\hat{\theta}_1(\theta_2)/\partial\theta_2 = -\ddot{\ell}_{21}\ddot{\ell}_{11}^{-1}$, and remembering that the matrices are evaluated at $(\hat{\theta}_1(\theta_2), \theta_2)$, this is

$$\frac{\partial \ddot{\ell}_{2.1}}{\partial \theta_2} = \dddot{\ell}_{222}[\cdot,\cdot,\cdot] + \dddot{\ell}_{221}[\cdot,\cdot,K] + \dddot{\ell}_{212}[\cdot,K,\cdot] + \dddot{\ell}_{211}[\cdot,K,K]$$

$$+ \dddot{\ell}_{112}[K,K,\cdot] + \dddot{\ell}_{111}[K,K,K] + \dddot{\ell}_{122}[K,\cdot,\cdot] + \dddot{\ell}_{121}[K,\cdot,K]$$

which is $\dddot{\ell}[J,J,J]$ with $J = (-\ddot{\ell}_{21}\ddot{\ell}_{11}^{-1} \;\; I)^T$. Therefore

$$\dot{F}_n = \dddot{\ell}[J\ddot{\ell}_{2.1}^{-1}, J\ddot{\ell}_{2.1}^{-1}\dot{\ell}_2, J]$$

Recall that $\dot{\ell}_1(\hat{\theta}_1(\theta_2), \theta_2) = 0$, so, using the factorization for $\ddot{\ell}$ given in Section 2,

$$\ddot{\ell}^{-1}\dot{\ell}(\hat{\theta}_1(\theta_2), \theta_2) = \begin{pmatrix} I & -\ddot{\ell}_{11}^{-1}\ddot{\ell}_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} \ddot{\ell}_{11}^{-1} & 0 \\ 0 & \ddot{\ell}_{2.1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\ddot{\ell}_{21}\ddot{\ell}_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} 0 \\ \dot{\ell}_2 \end{pmatrix}$$

$$= \begin{pmatrix} I & -\ddot{\ell}_{11}^{-1}\ddot{\ell}_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 \\ \ddot{\ell}_{2.1}^{-1}\dot{\ell}_2 \end{pmatrix} = J\ddot{\ell}_{2.1}^{-1}\dot{\ell}_2$$

Also write $H_2 = J\ddot{\ell}_{2.1}^{-1/2}$ and

$$H = \begin{pmatrix} H_1 & H_2 \end{pmatrix} = \begin{pmatrix} I & -\ddot{\ell}_{11}^{-1}\ddot{\ell}_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} \ddot{\ell}_{11}^{-1/2} & 0 \\ 0 & \ddot{\ell}_{2.1}^{-1/2} \end{pmatrix}$$

so that $H = \ddot{\ell}^{-T/2}$ and H_2 holds the last $\dim(\theta_2)$ columns. Now we can write

$$\dot{F}_n = \ddot{\ell}_{2.1}^{-1/2}\,\dddot{\ell}[H_2, \ddot{\ell}^{-1}\dot{\ell}, H_2]\ddot{\ell}_{2.1}^{1/2}$$

which is similar to the symmetric matrix

$$\dddot{\ell}[H_2, \ddot{\ell}^{-1}\dot{\ell}, H_2]$$

which is the trailing diagonal submatrix of

$$\dddot{\ell}[H, \ddot{\ell}^{-1}\dot{\ell}, H]$$

which is itself symmetric. Finally, noting that $H = \ddot{\ell}^{-1}H^{-T}$, this last matrix is

$$\dddot{\ell}[\ddot{\ell}^{-1}H^{-T}, \ddot{\ell}^{-1}\dot{\ell}, H]$$

which is similar to $\dot{F}$. □

A similar result shows that the Newton-Raphson iteration for any submodel (i.e., with a subset of the parameters held fixed) has spectral radius less than or equal to that of the full iteration. This result is relevant for the inner iterations for zigzag iterations.

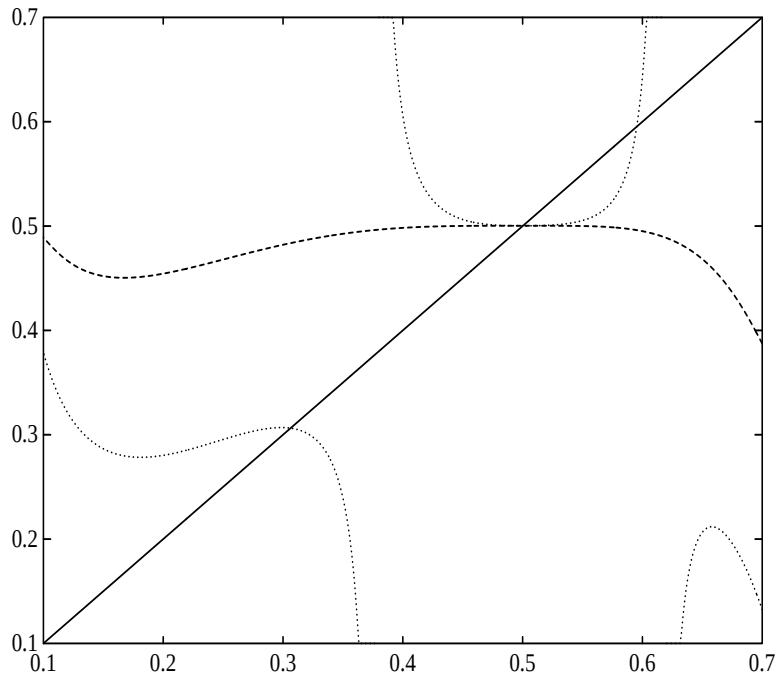

Figure 2: Plot of β versus $F(F(\beta))$, where F is an iteration function, for the data of Example 6.1. Dashed line is for nested Newton-Raphson, dotted line is for full Newton-Raphson with α initially at $\hat{\alpha}(\beta)$.

Example 6.1. Lawton and Sylvestre (1971) give a small data set to which they fit the growth model $E(Y) = \alpha \exp(\beta x)$ by least squares. Figure 2 plots β versus the two-step update $F_n(F_n(\beta))$, where F_n is the nested Newton-Raphson iteration function. Nested Newton-Raphson converges rapidly to the least squares estimate of about 0.5 for all values in the plot, from 0.1 to 0.7. Also plotted in the two-step update for β using the full Newton-Raphson iteration, where α is set equal to $\hat{\alpha}(\beta)$ initially. Note that, because of this choice for α, both nested and full iterations would give the same update for β after one step. The figure shows that the Newton-Raphson iteration oscillates wildly for β near 0.4 and 0.6, and converges to an incorrect stationary value for β near 0.3.

Theorem 8 can be extended to show that the iteration derivative for nested scoring is bounded in spectrum by that of full scoring. Intuitively, the two necessary pieces of the result have already been established in that Theorem 8 shows that nesting improves the quadratic approximation to the log-likelihood while Theorem 4 shows that nesting also improves the accuracy with which expected information approximates observed information. The iteration derivative for Fisher scoring may be written

$$\dot{F} = \mathcal{I}^{-1}(\ddot{\ell} + \mathcal{I}) - \dot{\mathcal{I}}[\mathcal{I}^{-1}, \mathcal{I}^{-1}\dot{\ell}, \cdot]$$

where $\dot{\mathcal{I}}$ is the trilinear form obtained by differentiating the Fisher information matrix. In this expression the first term is a measure of the difference between

Full Gauss-Newton				
Iteration	β_1	β_2	$\rho(\dot{F})$	$\frac{\|F(\theta)-\hat{\theta}\|}{\|\theta-\hat{\theta}\|}$
0	2.600000	5.200000		
1	1.831228	13.043768	1.624803	3.001813
2	0.572247	14.354976	2.134172	1.072768
3	4.112035	12.767053	1.820660	0.975636
4	4.593813	6.841691	2.066168	0.163049
5	3.219695	8.166753	0.341903	1.194553
6	3.828292	7.506807	0.154307	0.466922
7	4.092792	7.100365	0.014724	0.175763
8	4.118335	7.078389	0.027882	0.010473
9	4.117614	7.079257	0.028973	0.016849
10	4.117611	7.079264	0.028973	0.575726

Nested Gauss-Newton				
Iteration	β_1	β_2	$\rho(\dot{F})$	$\frac{\|F(\theta)-\hat{\theta}\|}{\|\theta-\hat{\theta}\|}$
0	2.600000	5.200000		
1	1.831228	13.043768		2.644429
2	3.198500	6.633438		0.159922
3	3.769804	7.963231		0.929900
4	4.037648	7.167256		0.125157
5	4.115179	7.082125		0.031516
6	4.117600	7.079288		0.004778

Table 1: Successive iterates obtained using the full and nested Gauss-Newton algorithms to solve the least squares problem of Example 6.2. The table also gives the spectral radius of the iteration derivative and the observed contraction factor at each iteration. The conditional least squares estimates $\hat{\alpha}(\beta)$ were used as starting values for α_0, α_1 and α_2. The iterates for the α_i are omitted from the table.

expected and observed information and the second is analogous to the iteration derivative for Newton-Raphson.

Although nesting improves the quadratic approximation, the reduced log-likelihood can still be far from quadratic for certain parameter values, as Ross (1990; page 124) points out in the case of the separable regression. In separable regression the reduced sum of squares is bounded above by $\sum y_i^2$ whereas a quadratic function would be unbounded. Ross argues that the quadratic approximation can be further improved by replacing the reduced sum of squares $r(\theta_2)$ with $r(\theta_2)/\{\sum y_i^2 - r(\theta_2)\}$.

Example 6.2. The model $E(Y_i) = \alpha_0 + \alpha_1 \exp(-\beta_1 x_i) + \alpha_2 \exp(-\beta_2 x_i)$ is fitted by least squares to a data set of 33 observations from Osborne (1972). Least squares estimates are $(\hat{\alpha}^T, \hat{\beta}^T) = (0.37541, 1.9358, -1.4647, 4.1176, 7.0793)$ with the x_i normalized so that $x_{33} = 1$. The fit is excellent, the residual root mean square

being 0.0013970, which is only slightly more than the accuracy with which the responses were recorded. Table 1 gives successive iterates for β_1 and β_2 for both full and nested Gauss-Newton algorithms starting from the values $(\beta_1, \beta_2) = (2.6, 5.2)$. For the full Gauss-Newton algorithm, the conditional least squares estimates $\hat{\alpha}(\beta)$ were used for the α_i. On this problem, both full and nested Gauss-Newton have the same limiting contraction factor, namely $R = 0.0290$, since the model is of the Stevens type discussed in Examples 4.3 and 5.2. Differences between the algorithms can therefore be interpreted as mainly due to differences in the quality of the quadratic approximation to the log-likelihood.

The nested algorithm has a radius of convergence 40 or 50% greater than that of the full algorithm. For example, the full algorithm converges if started from $\alpha = \hat{\alpha}(\beta)$, $\beta_2 = \hat{\beta}_2$ and β_1 in the interval $(1.4, 4.8)$, and for other values of β_1 it diverges. For the nested algorithm, the interval in which β_1 must lie is $(1.0, 5.9)$. Similarly, the full algorithm converges if started from $\alpha = \hat{\alpha}(\beta)$, $\beta_1 = \hat{\beta}_1$ and β_2 in the interval $(6.1, 20.5)$. For the nested Gauss-Newton algorithm, the interval in which β_2 must lie is $(4.9, 25.5)$.

Counting the number of iterations to convergence in cases where convergence in achieved gives an inflated impression of the performance of the Gauss-Newton algorithm, since, like Newton-Raphson, it tends to converge rapidly or not at all. By including a line search or trust region modification to ensure convergence, the algorithm may be started from more difficult initial values and the number of iterations to convergence can become far greater. On this problem for example, the full Gauss-Newton algorithm with a Levenberg-Marquardt modification requires 526 iterations to convergence from the starting value $\beta = (5.1, 5.2)^T$, $\alpha = \hat{\alpha}(\beta)$, while nested Gauss-Newton with the same modification requires only 11 iterations. Both unmodified algorithms diverge from this starting value.

7 Discussion

Nested scoring is in several ways the natural counterpart to full scoring in the presence of nuisance parameters. For example, it is well known that the score test statistic of a simple hypothesis, $H_0 : \theta = \theta_0$, is equal to the Wald test statistic using the estimator obtained after one scoring iteration starting from θ_0. Nested iterations play the same role for composite hypotheses. The score test statistic of the composite hypothesis $H_0 : \theta_2 = \theta_{20}$ is $\mathcal{I}_{2.1}^{-1/2}\dot{\ell}_2(\hat{\theta}_1(\theta_{20}), \theta_{20})$, which is equal to the Wald test statistic $\mathcal{I}_{2.1}^{1/2}(\tilde{\theta}_2 - \theta_{20})$ using the one-step estimator $\tilde{\theta}_2 = \theta_{20} + \mathcal{I}_{2.1}^{-1}\dot{\ell}_2(\hat{\theta}_1(\theta_{20}), \theta_{20})$. As another related example, the result that one-step scoring estimators starting from consistent estimates are fully efficient can be shown to carry over to show that one-step nested scoring estimators starting from consistent estimators are consistent in the presence of nuisance parameters.

The relative difference between observed and expected information, used in Section 5 to construct contraction factors, has eigenvalues which are geometric invariants. In the least squares context they relate to Bates and Watts' (1988) intrinsic curvature. For more general models they relate to statistical curvature as defined by Efron (1975), Amari (1982, 1985) and others. This relationship

discussed briefly in Smyth (1987).

The symmetrized version of the iteration derivative $\dot{F}$ used in Section 6 to evaluate the quadratic approximation to the log-likelihood can be represented as the symmetric trilinear form $\dddot{\ell}\ [\mathcal{I}^{-1/2}, \mathcal{I}^{-1/2}, \mathcal{I}^{-1/2}]$ acting on the standardized score vector $\mathcal{I}^{-1/2}\dot{\ell}$. Evaluated at $\hat{\theta}$, the trilinear form is the multiparameter version of the statistic $\dddot{\ell}\ (-\ddot{\ell})^{-3/2}$ used by Sprott (1973) and by Kass and Slate (1992) as a diagnostic in large sample estimation theory.

The idea of cycling between fits for submodels as in the zigzag iteration is very general and can be applied outside the maximum likelihood context. For example Schall (1991) estimated generalized linear models with random effects by cycling between and a scoring iteration for the mean parameters and an EM iteration for the dispersion parameters. Weisberg and Welsh (1991) estimate generalized linear model with nonparametric link by cycling between scoring for the mean parameters and kernel estimation for the link. Results similar to those in this paper will apply when there is an objective function that applies to both submodels, as in the likelihood context.

Barham and Drane (1972) suggest that nesting the Gauss-Newton iteration is more important in conjunction with a line search method than with a Levenberg-Marquardt modification. The author's experience, as Example 6.2 shows, has been that nesting can be very important in the latter case also. The question of whether the benefits of partitioning are diminished or increased when algorithms are modified for practical use has been necessarily beyond the scope of this paper, but is an important one which would repay further investigation.

Acknowledgements

This paper was substantially completed at School of Statistics, University of Minnesota, while the author was on a Special Study Program from the University of Queensland. The author is indebted to an anonymous referee for comments which improved the paper.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.*, **36**, 332–9.

Amari, S. (1982). Differential geometry of curved exponential families — curvatures and information loss. *Ann. Statist.* 10:357–385.

Amari, S. (1985). *Differential geometrical methods in statistics. Lecture notes in statistics 28.* Springer-Verlag, Heidelberg.

Barham, R. H. and Drane, W. (1972). An algorithm for least estimation of nonlinear parameters when some of the parameters are linear. *Technometrics*, **14**, 757–66.

Barndorff-Nielsen, O. E. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proc. Roy. Statist. Soc. A*, **353**, 401–9.

Bates, D. M. and Lindstrom, M. J. (1986). Nonlinear least squares with conditionally linear parameters. In Proceedings Statistical Computing Section, New York: American Statistical Association.

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity. *J. R. Statist. Soc. B* 42:1–25.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications.* New York: Wiley.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation (with discussion). *J. R. Statistic. Soc. B*, **26**, 211–52.

Chambers, J. M. and Hastie, T. J. (ed.) (1992). Statistical models in S. Wadsworth & Brooks/Cole, Pacific Grove, California.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–37.

Dieudonné, J. A. E. (1960). *Foundations of modern analysis.* New York: Academic Press.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3:1189–242.

Fieller, N. R. J., Flenley, E. C. and Olbricht, W. (1992). Statistics of particle size data. *Applied Statist.*, **41**, 127–46.

Gallant, A. R. (1987). *Nonlinear statistical models.* New York: Wiley.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.* 10:413–32.

Golub, G. H. and Pereyra, V. (1976). 'The differentiation of pseudo-inverses, separable nonlinear least square problems and other tales' in *Generalized Inverses and Applications*, Academic Press, New York, pp. 303–24.

Golub, G. H. and van Loan, C. F. (1983). *Matrix computations.* Johns Hopkins, Baltimore, Maryland.

Hartley, H. O. (1961). The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, **3**, 269–80.

Harville, D. A. (1973). Fitting partially linear models by weighted least squares. *Technometrics*, **15**, 509–515.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* 40:633–43.

Jensen, J. (1988). Maximum likelihood estimation of hyperbolic parameters from grouped observations. *Comput. Geosci.*, **14**, 380–408.

Jensen, S. T., Johansen, S. and Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78**, 867–77.

Jørgensen, B. (1984). The delta algorithm and GLIM. *Int. Statist. Rev.*, **52**, 282–300.

Jørgensen, B. (1987). Exponential dispersion models. *J. R. Statist. Soc. B*, **49**, 127–162.

Kass, R. E. and Slate, E. H. (1992). Reparameterization and diagnostics of posterior non-normality. Department of Statistics, Carnegie-Mellon University, Pittsburgh.

Kaufmann, L. (1975). A variable projection method for solving separable nonlinear least squares problems. *BIT* 15:49–57.

Khuri, A. I. (1984). A note on D-optimal designs for partially nonlinear regression models. *Technometrics*, **26**, 59–61.

Kowalik, J. and Osborne, M. R. (1968). *Methods for unconstrained optimization problems.* American Elsevier, New York.

Lawton, W. H. and Sylvestre, E. A. (1971). Elimination of linear parameters in nonlinear regression. *Technometrics* 13:461–67.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, **2**, 164–8.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, **11**, 431–41.

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

Nelson, D. L. and Lewis, T. O. (1970). A method for the solution of non-linear least squares problems when some of the parameters are linear. *Texas J. Science*, **21**, 480.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables.* Academic Press, New York.

Osborne, M. R. (1972). Some aspects of nonlinear least squares calculations. In Lootsma, F. (ed.), *Numerical methods for nonlinear optimization,* Academic Press, New York.

Osborne, M. R. (1987). Estimating nonlinear models by maximum likelihood for the exponential family. *SIAM J. Sci. Statist. Comp.* 8:446–56.

Osborne, M. R. (1992). Fisher's method of scoring. *Int. Statist. Rev.,* **60**, 99–117.

Osborne, M.R. and Smyth, G.K. (1991). A modified Prony algorithm for fitting functions defined by difference equations. *SIAM J. Sci. Statist. Comp.* 12: 362–382.

Ostrowski, A.M. (1960). *Solutions of equations and systems of equations.* Academic Press: New York.

Pimentel-Gomes, F. (1953). The use of Mitscherlich's regression law in the analysis of experiments with fertilizers. *Biometrics*, **9**, 498–516.

Pregibon, D. (1980). Goodness of link tests for generalised linear models. *Appl. Statist.*, **29**, 15–24.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Ratkowsky, D. A. (1983). *Nonlinear regression modelling, a unified practical approach.* New York: Dekker.

Ratkowsky, D. A. (1989). *Handbook of nonlinear regression models.* New York: Dekker.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *J. R. Statist. Soc. B* 23:469-75.

Ross, G. J. S. (1970). The efficient use of function minimization in non-linear maximum-likelihood estimation. *Appl. Statist.* 19:205–21.

Ross, G. J. S. (1990). *Nonlinear estimation.* New York: Springer-Verlag.

Ruhe, A. and Wedin, P. A. (1980). Algorithms for separable nonlinear least squares problems. *SIAM Rev.* 22:318–37.

Scallan, A. (1982). Some aspects of parametric link functions. In: R. Gilchrist, Ed., *GLIM 82*, New York: Springer-Verlag.

Scallan, A., Gilchrist, R. and Green, M. (1984). Fitting parametric link functions in generalised linear models. *Comput. Statist. Data Anal.*, **2**, 37–49.

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

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear regression.* New York: Wiley.

Smyth, G. K. (1987). Curvature and convergence. *1987 Proceedings of the Statistical Computing Section.* American Statistical Association, Virginia, 278–83.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R. Statist. Soc. B*, **51**, 47–60

Smyth, G. K. (1992). Using Poisson-gamma generalized linear models to model data with exact zeros. Technical Report, Department of Mathematics, University of Queensland.

Sprott, D. A. (1973). Normal likelihoods and their relation to large sample theory estimation. *Biometrika*, **60**, 457–465.

Stevens, W. L. (1951). Asymptotic regression. *Biometrics*, **7**, 247–67.

Thisted, R. (1988). *Statistical Computing.* New York: Chapman and Hall.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* (Eds. J. K. Ghosh and J. Roy), pp. 579–604. Calcutta: Indian Statistical Institute.

Varah, J.M. (1990). Relative sizes of the Hessian terms in nonlinear parameter estimation. *SIAM J. Sci. Stat. Comput.*, **11**, 174–179.

Verbyla, A. P. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *J. Roy. Statist. Soc. B*, **55**, 493–508.

Walling, D. (1968). Non-linear least squares curve fitting when some parameters are linear. *Texas J. Science*, **20**, 119–24.

Weisberg, S. and Welsh, A. H. (1991). Adapting for the missing link. Department of Statistics, The Faculties, Australian National University, 39pp.

Wermuth, N. and Scheidt, E. (1977). Algorithm AS105: Fitting a Covariance Selection Model to a Matrix. *Appl. Statist.* **26**, 88–92.