Electronic Transactions on Numerical Analysis.
Volume 28, pp. 1-15, 2007.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

SEPARABLE LEAST SQUARES, VARIABLE PROJECTION, AND THE GAUSS-NEWTON ALGORITHM*

M. R. OSBORNE†

Dedicated to Gene Golub on the occasion of his 75th birthday

Abstract. A regression problem is separable if the model can be represented as a linear combination of functions which have a nonlinear parametric dependence. The Gauss-Newton algorithm is a method for minimizing the residual sum of squares in such problems. It is known to be effective both when residuals are small, and when measurement errors are additive and the data set is large. The large data set result that the iteration asymptotes to a second order rate as the data set size becomes unbounded is sketched here. Variable projection is a technique introduced by Golub and Pereyra for reducing the separable estimation problem to one of minimizing a sum of squares in the nonlinear parameters only. The application of Gauss-Newton to minimize this sum of squares (the RGN algorithm) is known to be effective in small residual problems. The main result presented is that the RGN algorithm shares the good convergence rate behaviour of the Gauss-Newton algorithm on large data sets even though the errors are no longer additive. A modification of the RGN algorithm due to Kaufman, which aims to reduce its computational cost, is shown to produce iterates which are almost identical to those of the Gauss-Newton algorithm on the original problem. Aspects of the question of which algorithm is preferable are discussed briefly, and an example is used to illustrate the importance of the large data set behaviour.

Key words. nonlinear least squares, scoring, Newton's method, expected Hessian, Kaufman's modification, rate of convergence, random errors, law of large numbers, consistency, large data sets, maximum likelihood

AMS subject classifications. 62-07, 65K99, 90-08

1. Introduction.

The Gauss-Newton algorithm is a modification of Newton's method for minimization developed for the particular case when the objective function can be written as a sum of squares. It has a cost advantage in that it avoids the calculation of second derivative terms in estimating the Hessian. Other advantages possessed by the modified algorithm are that its Hessian estimate is generically positive definite, and that it actually has better transformation invariance properties than those possessed by the original algorithm. It has the disadvantage that it has a generic first order rate of convergence. This can make the method unsuitable except in two important cases:

1. *The case of small residuals*. This occurs when the individual terms in the sum of squares can be made small simultaneously so that the associated nonlinear system is consistent or nearly so.
2. *The case of large data sets*. An important application of the Gauss-Newton algorithm is to parameter estimation problems in data analysis. Nonlinear least squares problems occur in maximizing likelihoods based on the normal distribution. Here Gauss-Newton is a special case of the Fisher scoring algorithm [6]. In appropriate circumstances this asymptotes to a second order convergence rate as the number of independent observations in the data set becomes unbounded.

The large data set problem is emphasised here. This seeks to estimate the true parameter vector $\overline{\boldsymbol{\beta}} \in \mathbb{R}^p$ by solving the optimization problem

$$\min_{\boldsymbol{\beta}} F_n\left(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n\right), \tag{1.1}$$

where

$$F_n\left(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n\right) = \frac{1}{2n}\left\|\mathbf{f}^n\left(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n\right)\right\|^2,$$

*Received November 17, 2006. Accepted for publication February 21, 2007. Recommended by M. Gutknecht.
†Mathematical Sciences Institute, Australian National University, ACT 0200, Australia (mike.osborne@maths.anu.edu.au).

$\mathbf{f}^n \in \mathbb{R}^n$ is a vector of smooth enough functions $f_i^n(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n)$, $i = 1, 2, \ldots, n$, $\nabla_\beta \mathbf{f}^n$ has full column rank p in the region of parameter space of interest, and $\boldsymbol{\varepsilon}^n \in \mathbb{R}^n \sim N(0, \sigma^2 I_n)$ plays the role of observational error. The norm is assumed to be the Euclidean vector norm unless otherwise specified. It is assumed that the measurement process that generated the data set can be conceptualised for arbitrarily large n, and that the estimation problem is consistent in the sense that there exists a sequence $\left\{\widehat{\boldsymbol{\beta}}_n\right\}$ of local minimisers of (1.1) such that $\widehat{\boldsymbol{\beta}}_n \overset{a.s.}{\to} \overline{\boldsymbol{\beta}}$, $n \to \infty$. Here the mode of convergence is almost sure convergence. A good reference on asymptotic methods in statistics is [12].

REMARK 1.1. A key point is that the errors are assumed to enter the model additively. That is, the f_i^n, $i = 1, 2, \ldots, n$, have the functional form

$$f_i^n(\boldsymbol{\beta}, \boldsymbol{\varepsilon}) = y_i^n - \mu_i^n(\boldsymbol{\beta}),$$

where, corresponding to the case of observations made on a signal in the presence of noise,

$$y_i^n = \mu_i^n\left(\overline{\boldsymbol{\beta}}\right) + \varepsilon_i^n. \tag{1.2}$$

Thus differentiation of f_i^n removes the random component. Also F_n is directly proportional to the problem log likelihood and the property of consistency becomes a consequence of the other assumptions.

In a number of special cases there is additional structure in $\mathbf{f}^n$ so it becomes a legitimate question to ask if this can be used to advantage. A nonlinear regression model is called *separable* if the problem residuals $\mathbf{b}^n$ can be represented in the form

$$b_i^n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\varepsilon}^n) = y_i^n - \sum_{j=1}^m \phi_{ij}(\boldsymbol{\beta})\,\alpha_j, \quad i = 1, 2, \ldots, n.$$

Here the model has the form of a linear combination expressed by $\boldsymbol{\alpha} \in \mathbb{R}^m$ of nonlinear functions $\phi_{ij}(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathbb{R}^p$. The modified notation

$$f_i^n(\boldsymbol{\beta}, \boldsymbol{\varepsilon}) \to b_i^n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\varepsilon}^n),$$
$$\mu_i^n(\boldsymbol{\beta}) \to \sum_{j=1}^m \phi_{ij}(\boldsymbol{\beta})\,\alpha_j,$$

is used here to make this structure explicit. It is assumed that the problem functions are $\phi_{ij}(\boldsymbol{\beta}) = \phi_j(t_i^n, \boldsymbol{\beta})$, $j = 1, 2, \ldots, m$, where the t_i^n, $i = 1, 2, \ldots, n$, are sample points where observations on the underlying signal are made. There is no restriction in assuming $t_i^n \in [0, 1]$. One source of examples is provided by general solutions of the m'th order linear ordinary differential equation with fundamental solutions given by the $\phi_i(t, \boldsymbol{\beta})$. In [1] a systematic procedure (variable projection) is introduced for reducing the estimation problem to a nonlinear least squares problem in the nonlinear parameters $\boldsymbol{\beta}$ only. A recent survey of developments and applications of variable projection is [2]. To introduce the technique let $\Phi_n : \mathbb{R}^m \to \mathbb{R}^n$, $n > m$, be the matrix with components ϕ_{ij}. The rank assumption in the problem formulation now requires $\begin{bmatrix} \Phi_n & \nabla_\beta \Phi_n \boldsymbol{\alpha} \end{bmatrix}$ to have full column rank $m + p$. Also let $P_n(\boldsymbol{\beta}) : \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection matrix defined by

$$P_n(\boldsymbol{\beta})\,\Phi_n(\boldsymbol{\beta}) = 0. \tag{1.3}$$

Here $P_n(\boldsymbol{\beta})$ has the explicit representation

$$P_n(\boldsymbol{\beta}) = I_n - \Phi_n\left(\Phi_n^T \Phi_n\right)^{-1} \Phi_n^T.$$

ETNA
Kent State University
etna@mcs.kent.edu

Then

$$F_n = \frac{1}{2n}\left\{\|P_n\mathbf{y}^n\|^2 + \|(I_n - P_n)\,\mathbf{b}^n\|^2\right\}.$$

The first term on the right of this equation is independent of $\boldsymbol{\alpha}$ and the second can be reduced to zero by setting

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\beta}) = \left(\Phi_n^T\Phi_n\right)^{-1}\Phi_n^T\mathbf{y}^n. \tag{1.4}$$

Thus an equivalent formulation of (1.1) in the separable problem is

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|P_n(\boldsymbol{\beta})\,\mathbf{y}^n\|^2, \tag{1.5}$$

which is a sum of squares in the nonlinear parameters $\boldsymbol{\beta}$ only so that, at least formally, the Gauss-Newton algorithm can be applied. However, now the random errors do not enter additively but are coupled with the nonlinear parameters in setting up the objective function.

The plan of the paper is as follows. The large data set rate of convergence analysis appropriate to the Gauss-Newton method in the case of additive errors is summarized in the next section. The third section shows why this analysis cannot immediately be extended to the RGN algorithm. Here the rather harder work needed to arrive at similar conclusions is summarised. Most implementations of the variable projection method use a modification due to Kaufman [4] which serves to reduce the amount of computation needed in the RGN algorithm. This modified algorithm also shows the favourable large data set rates despite being developed using an explicit small residual argument. However, it is actually closer to the additive Gauss-Newton method than is the full RGN algorithm. A brief discussion of which form of algorithm is appropriate in particular circumstances is given in the final section. This is complemented by an example of a classic data fitting problem which is used to illustrate the importance of the large sample convergence rate.

2. Large data set convergence rate analysis. The basic iterative step in Newton's method for minimizing F_n defined in (1.1) is

$$\begin{aligned}\boldsymbol{\beta}_{i+1} &= \boldsymbol{\beta}_i - \mathcal{J}_n(\boldsymbol{\beta}_i)^{-1}\nabla_\beta F_n(\boldsymbol{\beta}_i)^T,\\ \mathcal{J}_n(\boldsymbol{\beta}_i) &= \nabla_\beta^2 F_n(\boldsymbol{\beta}_i).\end{aligned} \tag{2.1}$$

In the case of additive errors the scoring/Gauss-Newton method replaces the Hessian with an approximation which is constructed as follows. The true Hessian is

$$\mathcal{J}_n(\boldsymbol{\beta}) = \frac{1}{n}\left\{\{\nabla_\beta\mathbf{f}^n\}^T\{\nabla_\beta\mathbf{f}^n\} + \sum_{i=1}^n f_i^n\nabla_\beta^2 f_i^n\right\}. \tag{2.2}$$

The stochastic component enters only through (1.2) so taking expectations gives

$$\mathcal{E}\{\mathcal{J}_n\}(\boldsymbol{\beta}) = \mathcal{I}_n(\boldsymbol{\beta}) - \frac{1}{n}\sum_{i=1}^n\left(\mu_i(\boldsymbol{\beta}) - \mu_i(\overline{\boldsymbol{\beta}})\right)\nabla_\beta^2 f_i^n(\boldsymbol{\beta}),$$

where

$$\mathcal{I}_n(\boldsymbol{\beta}) = \frac{1}{n}\left\{\{\nabla_\beta\mathbf{f}^n\}^T\{\nabla_\beta\mathbf{f}^n\}\right\}. \tag{2.3}$$

The Gauss-Newton method replaces $\mathcal{J}_n(\boldsymbol{\beta})$ with $\mathcal{I}_n(\boldsymbol{\beta})$ in (2.1). The key point to notice is

$$\mathcal{I}_n(\overline{\boldsymbol{\beta}}) = \mathcal{E}\left\{\mathcal{J}_n(\overline{\boldsymbol{\beta}})\right\}. \tag{2.4}$$

Several points can be made here:

1. It follows from the special form of (2.3) that the Gauss-Newton correction $\boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i$ solves the linear least squares problem

$$\min_{\mathbf{t}} \left\| \mathbf{y}^n - \boldsymbol{\mu}^n(\boldsymbol{\beta}) - \nabla_\beta \boldsymbol{\mu}^n(\boldsymbol{\beta})\,\mathbf{t} \right\|^2 . \tag{2.5}$$

2. It is an important result, conditional on an appropriate experimental setup, that $\mathcal{I}_n(\boldsymbol{\beta})$ is generically a bounded, positive definite matrix for all n large enough [6]. A similar result is sketched in Lemma 3.2.
3. The use of the form of the expectation which holds at the true parameter values is a characteristic simplification of the scoring algorithm and is available for more general likelihoods [7]. Here it leads to the same result as ignoring small residual terms in (2.2).

The full-step Gauss-Newton method has the form of a fixed point iteration:

$$\begin{aligned} \boldsymbol{\beta}_{i+1} &= Q_n(\boldsymbol{\beta}_i), \\ Q_n(\boldsymbol{\beta}) &= \boldsymbol{\beta} - \mathcal{I}_n(\boldsymbol{\beta})^{-1} \nabla_\beta F_n(\boldsymbol{\beta})^T . \end{aligned}$$

The condition for $\widehat{\boldsymbol{\beta}}_n$ to be an attractive fixed point is

$$\varpi\left(Q_n'\left(\widehat{\boldsymbol{\beta}}_n\right)\right) < 1,$$

where ϖ denotes the spectral radius of the variational matrix Q_n'. This quantity determines the first order convergence multiplier of the Gauss-Newton algorithm. The key to the good large sample behaviour is the result

$$\varpi\left(Q_n'\left(\widehat{\boldsymbol{\beta}}_n\right)\right) \xrightarrow{a.s.} 0, \quad n \to \infty. \tag{2.6}$$

which shows that the algorithm tends to a second order convergent process as $n \to \infty$. The derivation of this result will now be outlined. As $\nabla_\beta F_n\left(\widehat{\boldsymbol{\beta}}_n\right) = 0$, it follows that

$$Q_n'\left(\widehat{\boldsymbol{\beta}}_n\right) = I_p - \mathcal{I}_n\left(\widehat{\boldsymbol{\beta}}_n\right)^{-1} \nabla_\beta^2 F_n\left(\widehat{\boldsymbol{\beta}}_n\right).$$

Now define $W_n(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}^p$ by

$$W_n(\boldsymbol{\beta}) = \mathcal{I}_n(\boldsymbol{\beta})^{-1}\left\{\mathcal{I}_n(\boldsymbol{\beta}) - \nabla_\beta^2 F_n(\boldsymbol{\beta})\right\}. \tag{2.7}$$

Then

$$W_n\left(\widehat{\boldsymbol{\beta}}_n\right) = Q_n'\left(\widehat{\boldsymbol{\beta}}_n\right) = W_n\left(\overline{\boldsymbol{\beta}}\right) + O\left(\left\|\widehat{\boldsymbol{\beta}}_n - \overline{\boldsymbol{\beta}}\right\|\right), \tag{2.8}$$

by consistency. By (2.4),

$$W_n\left(\overline{\boldsymbol{\beta}}\right) = -\mathcal{I}_n\left(\overline{\boldsymbol{\beta}}\right)^{-1}\left\{\nabla_\beta^2 F_n\left(\overline{\boldsymbol{\beta}}\right) - \mathcal{E}\left\{\nabla_\beta^2 F_n\left(\overline{\boldsymbol{\beta}}\right)\right\}\right\}. \tag{2.9}$$

It has been noted that $\mathcal{I}_n\left(\overline{\boldsymbol{\beta}}\right)$ is bounded, positive definite. Also, a factor $\frac{1}{n}$ is implicit in the second term of the right hand side of (2.9), and the components of $\nabla_\beta^2 F_n\left(\overline{\boldsymbol{\beta}}\right)$ are sums of independent random variables. Thus it follows by an application of the law of large numbers [12], that $W_n\left(\overline{\boldsymbol{\beta}}\right) \xrightarrow{a.s.} 0$ component-wise as $n \to \infty$. An immediate consequence is that

$$\varpi\left(W_n\left(\overline{\boldsymbol{\beta}}\right)\right) \xrightarrow{a.s.} 0, \quad n \to \infty.$$

The desired convergence rate result (2.6) now follows from (2.8). Note that the property of consistency that derives from the maximum likelihood connection is an essential component of the argument. Also, that this is not a completely straightforward application of the law of large numbers because a sequence of sets of observation points $\{t_i^n,\ i = 1, 2, \ldots, n\}$ is involved. For this case see [13].

3. Rate estimation for separable problems. Variable projection leads to the nonlinear least squares problem (1.5) where

$$\mathbf{f}^n(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n) = P_n(\boldsymbol{\beta})\,\mathbf{y}^n,$$
$$F_n(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n) = \frac{1}{2n}(\mathbf{y}^n)^T P_n(\boldsymbol{\beta})\,\mathbf{y}^n.$$

Implementation of the Gauss-Newton algorithm (RGN algorithm) has been discussed in detail in [11]. It uses an approximate Hessian computed from (2.3) and requires derivatives of $P_n(\boldsymbol{\beta})$. The derivative of P in the direction defined by $\mathbf{t} \in \mathbb{R}^p$ is

$$\nabla_\beta P[\mathbf{t}] = -P\nabla_\beta \Phi[\mathbf{t}]\,\Phi^+ - \left(\Phi^+\right)^T \nabla_\beta \Phi^T[\mathbf{t}]\,P \tag{3.1}$$
$$= A(\boldsymbol{\beta}, \mathbf{t}) + A^T(\boldsymbol{\beta}, \mathbf{t}), \tag{3.2}$$

where $A \in \mathbb{R}^n \to \mathbb{R}^n$, the matrix directional derivative $\frac{d\Phi}{dt}$ is written $\nabla_\beta \Phi[\mathbf{t}]$ to emphasise both the linear dependence on $\mathbf{t}$ and that $\mathbf{t}$ is held fixed in this operation, explicit dependence on both n and $\boldsymbol{\beta}$ is understood, and Φ^+ denotes the generalised inverse of Φ. Note that $\Phi^+ P = \Phi^+ - \Phi^+\Phi\Phi^+ = 0$ so the two components of $\nabla_\beta P[\mathbf{t}]$ in (3.2) are orthogonal. Define matrices $K, L : \mathbb{R}^p \to \mathbb{R}^n$ by

$$A(\boldsymbol{\beta}, \mathbf{t})\,\mathbf{y} = K(\boldsymbol{\beta}, \mathbf{y})\,\mathbf{t},$$
$$A^T(\boldsymbol{\beta}, \mathbf{t})\,\mathbf{y} = L(\boldsymbol{\beta}, \mathbf{y})\,\mathbf{t}.$$

Then the RGN correction solves

$$\min_{\mathbf{t}} \|P\mathbf{y} + (K + L)\,\mathbf{t}\|^2, \tag{3.3}$$

where

$$L^T K = 0 \tag{3.4}$$

as a consequence of the orthogonality noted above.

REMARK 3.1. Kaufman [4] has examined these terms in more detail. We have

$$\mathbf{t}^T K^T K \mathbf{t} = \mathbf{y}^T A^T A \mathbf{y} = O\left(\|\boldsymbol{\alpha}\|^2\right),$$
$$\mathbf{t}^T L^T L \mathbf{t} = \mathbf{y}^T A A^T \mathbf{y} = O\left(\|P\mathbf{y}\|^2\right).$$

If the orthogonality noted above is used then the second term in the design matrix in (3.3) corresponds to a small residual term when $\|P\mathbf{y}\|^2$ is relatively small and can be ignored . The resulting correction solves

$$\min_{\mathbf{t}} \|P\mathbf{y} + K\mathbf{t}\|^2. \tag{3.5}$$

This modification was suggested by Kaufman. It can be implemented with less computational cost, and it is favoured for this reason. Numerical experience is reported to be very satisfactory [2].

ETNA
Kent State University
etna@mcs.kent.edu

The terms in the sum of squares in the reduced problem (1.5) are

$$f_i = \sum_{i=1}^{n} P_{ij} y_j, \quad i = 1, 2, \ldots, n.$$

Now, because the noise ε is coupled with the nonlinear parameters and so does not disappear under differentiation, $\mathcal{I}_n$ is quadratic in the noise contributions. An immediate consequence is that

$$\mathcal{I}_n \neq \frac{1}{n} \mathcal{E} \left\{ \nabla_\beta \mathbf{f}^T \nabla_\beta \mathbf{f} \right\}.$$

Thus it is not possible to repeat exactly the rate of convergence calculation of the previous section. Instead it is convenient to rewrite equation (2.7):

$$W_n (\boldsymbol{\beta}) = - \left(\frac{1}{n} \nabla_\beta \mathbf{f}^T \nabla_\beta \mathbf{f} \right)^{-1} \frac{1}{n} \left\{ \sum_{i=1}^{n} f_i \nabla_\beta^2 f_i \right\}, \tag{3.6}$$

where the right hand side is evaluated at $\boldsymbol{\beta}$. The property of consistency is unchanged so the asymptotic convergence rate is again determined by $\varpi \left(W_n \left(\overline{\boldsymbol{\beta}} \right) \right)$. We now examine this expression in more detail.

LEMMA 3.2.

$$\frac{1}{n} \Phi_n^T \Phi_n \to G, \quad n \to \infty, \tag{3.7}$$

where

$$G_{ij} = \int_0^1 \phi_i (t) \phi_j (t) \varrho (t) \, dt, \quad 1 \leq i, j \leq m,$$

and the density ρ is determined by the asymptotic properties of the method for generating the sample points t_i^n, $i = 1, 2, \ldots, n$, for large n. The Gram matrix G is bounded and generically positive definite. Let $T_n = I - P_n$. Then

$$(T_n)_{ij} = \frac{1}{n} \boldsymbol{\phi}_i^T G^{-1} \boldsymbol{\phi}_j + o \left(\frac{1}{n} \right), \tag{3.8}$$

where

$$\boldsymbol{\phi}_i = \left[\begin{array}{cccc} \phi_1 (t_i) & \phi_2 (t_i) & \cdots & \phi_m (t_i) \end{array} \right]^T.$$

This gives an $O\left(\frac{1}{n}\right)$ component-wise estimate which applies also to derivatives of both P_n and T_n with respect to $\boldsymbol{\beta}$.

Proof. The result (3.7) is discussed in detail in [6]. It follows from

$$\left(\frac{1}{n} \Phi_n^T \Phi_n \right)_{ij} = \frac{1}{n} \sum_{k=1}^{n} \phi_i (t_k) \phi_j (t_k) = G_{ij} + O \left(\frac{1}{n} \right)$$

by interpreting the sum as a quadrature formula. Positive definiteness is a consequence of the problem rank assumption. To derive (3.8) note that

$$\begin{aligned} T_n &= \Phi_n \left(\Phi_n^T \Phi_n \right)^{-1} \Phi_n^T \\ &= \frac{1}{n} \Phi_n G^{-1} \Phi_n^T + o \left(\frac{1}{n} \right). \qquad \square \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu

The starting point for determining the asymptotics of the convergence rate of the RGN algorithm as $n \to \infty$ is the computation of the expectations of the numerator and denominator matrices in (3.6). The expectation of the denominator is bounded and generically positive definite. The expectation of the numerator is $O\left(\frac{1}{n}\right)$ as $n \to \infty$. This suggests strongly that the spectral radius of $Q'\left(\overline{\boldsymbol{\beta}}\right) \to 0,\ n \to \infty$, a result of essentially similar strength to that obtained for the additive error case. To complete the proof requires showing that both numerator and denominator terms converge to their expectations with probability 1.

Consider first the denominator term.

LEMMA 3.3. *Fix $\boldsymbol{\beta} = \overline{\boldsymbol{\beta}}$.*

$$\frac{1}{n}\mathcal{E}\left\{\nabla_\beta \mathbf{f}^T \nabla_\beta \mathbf{f}\right\} = \sigma^2 M_1 + M_2,$$

where $M_1 = O\left(\frac{1}{n}\right),\ n \to \infty$, and M_2 tends to a limit which is a bounded, positive definite matrix when the problem rank assumption is satisfied. In detail, these matrices are

$$M_1 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ij},$$

$$M_2 = \frac{1}{n}\left\{\sum_{j=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_j - \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_k T_{jk}\right\}.$$

Proof. Set

$$\begin{aligned}\nabla \mathbf{f}^T \nabla \mathbf{f} &= \sum_{i=1}^{n}\nabla f_i^T \nabla f_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla P_{ij}\right)^T y_j \sum_{k=1}^{n}\nabla P_{ik} y_k.\end{aligned}$$

To calculate the expectation note that it follows from equation (1.2) that

$$\mathcal{E}\left\{y_j y_k\right\} = \sigma^2 \delta_{jk} + \mu_j\left(\overline{\boldsymbol{\beta}}\right)\mu_k\left(\overline{\boldsymbol{\beta}}\right), \tag{3.9}$$

where

$$\mu_j\left(\boldsymbol{\beta}\right) = \mathbf{e}_j^T \Phi \boldsymbol{\alpha}\left(\boldsymbol{\beta}\right).$$

It follows that

$$\begin{aligned}\frac{1}{n}\mathcal{E}\left\{\nabla_\beta \mathbf{f}^T \nabla_\beta \mathbf{f}\right\} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2\sum_{j=1}^{n}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ij} + \sum_{j=1}^{n}\sum_{k=1}^{n}\mu_j\mu_k\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}\right\} \\ &= \sigma^2 M_1 + M_2\end{aligned}$$

To show $M_1 \to 0$ is a counting exercise. M_1 consists of the sum of n^2 terms each of which is an $p \times p$ matrix of $O(1)$ gradient terms divided by n^3 as a consequence of Lemma 3.2. M_2 can be simplified somewhat by noting that $\sum_{j=1}^{n} P_{ij}\mu_j = 0$ identically in β by (1.3) so that

$$\sum_{j=1}^{n}\mu_j \nabla_\beta P_{ij} = -\sum_{j=1}^{n}\nabla_\beta \mu_j P_{ij}.$$

This gives, using the symmetry of $P = I - T$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mu_j\mu_k\left(\nabla_\beta P_{ij}\right)^T\nabla_\beta P_{ik} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta\mu_j^T\nabla_\beta\mu_k P_{ij}P_{ik}$$
$$(3.10) \qquad = \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta\mu_j^T\nabla_\beta\mu_k P_{jk}$$
$$= \sum_{j=1}^{n}\nabla_\beta\mu_j^T\nabla_\beta\mu_j - \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta\mu_j^T\nabla_\beta\mu_k T_{jk}.$$

Boundedness of M_2 as $n \to \infty$ now follows using the estimates for the size of the T_{ij} computed in Lemma 3.2. To show that M_2 is positive definite note that it follows from (3.10) that

$$\mathbf{t}^T M_2 \mathbf{t} = \frac{d\boldsymbol{\mu}}{dt}^T \{I - T\}\frac{d\boldsymbol{\mu}}{dt} \geq 0.$$

As $\left\|T\frac{d\boldsymbol{\mu}}{dt}\right\| \leq \left\|\frac{d\boldsymbol{\mu}}{dt}\right\|$, this expression can vanish only if there is a direction $\mathbf{t} \in \mathbb{R}^p$ such that $\frac{d\boldsymbol{\mu}}{dt} = \gamma\boldsymbol{\mu}$ for some $\gamma \neq 0$. This requirement is contrary to the Gauss-Newton rank assumption that $\left[\begin{array}{cc}\Phi & \nabla_\beta\Phi\boldsymbol{\alpha}\end{array}\right]$ has full rank $m + p$. □

LEMMA 3.4. *The numerator in the expression (3.6) defining $W_n(\overline{\boldsymbol{\beta}})$ is*

$$\sum_{i=1}^{n} f_i\nabla_\beta^2 f_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} y_j y_k P_{ij}\nabla_\beta^2 P_{ik}.$$

Let $M_3 = \frac{1}{n}\mathcal{E}\left\{\sum_{i=1}^{n} f_i\nabla_\beta^2 f_i\right\}$ then

$$M_3 = \frac{1}{n}\sum_{i=1}^{n}\sigma^2\left\{\nabla_\beta^2 P_{ii} - \sum_{j=1}^{n} T_{ij}\nabla_\beta^2 P_{ij}\right\},$$

and $M_3 \to 0,\ n \to \infty$.

Proof. This is similar to that of Lemma 3.3. The new point is that the contribution to M_3 from the signal terms $\mu_j(\overline{\boldsymbol{\beta}})$ in the expectation (3.9) is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mu_j\mu_k P_{ij}\nabla_\beta^2 P_{ik} = 0,$$

by summing over j keeping i and k fixed. The previous counting argument can be used again to give the estimate $M_3 = O\left(\frac{1}{n}\right),\ n \to \infty$. □

The final step required is to show that the numerator and denominator terms in (3.6) approach their expectations as $n \to \infty$. Only the case of the denominator is considered here.

LEMMA 3.5.

$$\left(\frac{1}{n}\nabla_\beta\mathbf{f}^T\nabla_\beta\mathbf{f}\right) \xrightarrow{a.s.} M_2, \quad n \to \infty.$$

ETNA
Kent State University
etna@mcs.kent.edu

Proof. The basic quantities are:

$$\begin{aligned}\left(\frac{1}{n}\nabla_\beta \mathbf{f}^T\nabla_\beta \mathbf{f}\right) &= \frac{1}{n}\sum_{i=1}^n \nabla_\beta f_i^T \nabla_\beta f_i\\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \left(\nabla_\beta P_{ij}\right)^T y_j \sum_{k=1}^n \nabla_\beta P_{ik} y_k\\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \left\{\mu_j\mu_k + (\mu_j\varepsilon_k + \mu_k\varepsilon_j) + \varepsilon_j\varepsilon_k\right\}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}.\end{aligned}$$

The first of the three terms in this last expansion is M_2. Thus the result requires showing that the remaining terms tend to 0. Let

$$\boldsymbol{\pi}_i^n = \sum_{j=1}^n \varepsilon_j \left(\nabla_\beta P_{ij}\right)^T, \quad \boldsymbol{\pi}_i^n \in \mathbb{R}^p.$$

As, by Lemma 3.2, the components of $\nabla_\beta P_{ij} = O\left(\frac{1}{n}\right)$, it follows by applications of the law of large numbers that

$$\boldsymbol{\pi}_i^n \overset{a.s.}{\to} 0, \quad n \to \infty,$$

componentwise. Specifically, given $\delta > 0$, there is an n_0 such that

$$\forall i, \quad \left\|\boldsymbol{\pi}_i^n\right\|_\infty < \delta \quad \forall n > n_0 \text{ with probability } 1.$$

Consider the third term. Let

$$\begin{aligned}S_n &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \varepsilon_j\varepsilon_k \left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}\\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\pi}_i^n \left(\boldsymbol{\pi}_i^n\right)^T.\end{aligned}$$

Then, in the maximum norm, with probability 1 for $n > n_0$,

$$\|S_n\|_\infty \le p\delta^2,$$

showing that the third sum tends to 0, $n \to \infty$ almost surely. A similar argument applies to the second term which proves to be $O(\delta)$. □

These results can now be put together to give the desired convergence result.

THEOREM 3.6.

$$W_n\left(\overline{\boldsymbol{\beta}}\right) \overset{a.s.}{\to} 0, \quad n \to \infty.$$

Proof. The idea is to write each component term Ω in (3.6) in the form

$$\Omega = \mathcal{E}\left\{\Omega\right\} + \left(\Omega - \mathcal{E}\left\{\Omega\right\}\right),$$

and then to appeal to the asymptotic convergence results established in the preceding lemmas. □

REMARK 3.7. This result when combined with consistency suffices to establish the analogue of (2.6) in this case. The asymptotic convergence rate of the RGN algorithm can be expected to be similar to that of the full Gauss-Newton method. While the numerator expectation in the Gauss-Newton method is 0, and that in the RGN algorithm is $O\left(\frac{1}{n}\right)$ by Lemma 3.4, these are both smaller than the discrepancies $\left(\Omega - \mathcal{E}\left\{\Omega\right\}\right)$ between their full expressions and their expectations. Thus it is these discrepancy terms that are critical in determining the convergence rates. Here these correspond to law of large numbers rates for which a scale of $O\left(n^{-1/2}\right)$ is appropriate.

4. The Kaufman modification. As the RGN algorithm possesses similar convergence rate properties to Gauss-Newton in large sample problems, and, as the Kaufman modification is favoured in implementation, it is of interest to ask if it too shares the same good large sample convergence rate properties. Fortunately the answer is in the affirmative. This result can be proved in the same way as the main lemmas in the previous section. This calculation is similar to the preceding and is relegated to the Appendix. In this section the close connection between the modified algorithm and the full Gauss-Newton method is explored. That both can be implemented with the same amount of work is shown in [11]. First note that equation (2.5) for the Gauss-Newton correction here becomes

$$\min_{\delta\boldsymbol{\alpha},\delta\boldsymbol{\beta}} \left\| \mathbf{y} - \Phi\boldsymbol{\alpha} - \left[\begin{array}{cc} \Phi & \nabla_\beta\left(\Phi\boldsymbol{\alpha}\right) \end{array}\right] \left[\begin{array}{c} \delta\boldsymbol{\alpha} \\ \delta\boldsymbol{\beta} \end{array}\right] \right\|^2 .$$

Introducing the variable projection matrix P permits this to be written:

$$\min_{\delta\boldsymbol{\beta}} \left\| P\mathbf{y} - P\nabla_\beta\left(\Phi\boldsymbol{\alpha}\right)\delta\boldsymbol{\beta} \right\|^2 + \min_{\delta\boldsymbol{\alpha}} \left\| (I-P)\left(\mathbf{y} - \nabla_\beta\left(\Phi\boldsymbol{\alpha}\right)\delta\boldsymbol{\beta}\right) - \Phi\left(\boldsymbol{\alpha}+\delta\boldsymbol{\alpha}\right) \right\|^2 .$$

Comparison with (3.1) shows that the first minimization is just

$$\min_{\delta\boldsymbol{\beta}} \left\| P\mathbf{y} - K\delta\boldsymbol{\beta} \right\| .$$

Thus, given $\boldsymbol{\alpha}$, the Kaufman search direction computed using (3.5) is exactly the Gauss-Newton correction for the nonlinear parameters. If $\boldsymbol{\alpha}$ is set using (1.4) then the second minimization gives

$$\begin{aligned} \delta\boldsymbol{\alpha} &= -\Phi^+\nabla_\beta\left(\Phi\boldsymbol{\alpha}\right)\delta\boldsymbol{\beta} \\ &= -\Phi^+\nabla_\beta\Phi\left[\delta\boldsymbol{\beta}\right]\Phi^+\mathbf{y}, \end{aligned} \tag{4.1}$$

while the increment in $\boldsymbol{\alpha}$ arising from the Kaufman correction is

$$\boldsymbol{\alpha}\left(\boldsymbol{\beta}+\delta\boldsymbol{\beta}\right) - \boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) = \left(\nabla_\beta\Phi^+\mathbf{y}\right)\delta\boldsymbol{\beta} + O\left(\left\|\delta\boldsymbol{\beta}\right\|^2\right).$$

Note this increment is not computed as part of the algorithm. To examine (4.1) in more detail we have

$$\begin{aligned} \frac{d\Phi^+}{dt} &= -\left(\Phi^T\Phi\right)^{-1}\left(\frac{d\Phi^T}{dt}\Phi + \Phi^T\frac{d\Phi}{dt}\right)\left(\Phi^T\Phi\right)^{-1}\Phi^T + \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt} \\ &= -\left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}T - \Phi^+\frac{d\Phi}{dt}\Phi^+ + \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt} \\ &= \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}P - \Phi^+\frac{d\Phi}{dt}\Phi^+. \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu

The second term in this last equation occurs in (4.1). Thus, setting $\delta\boldsymbol{\beta} = \|\delta\boldsymbol{\beta}\|\,\mathbf{t}$,

$$\begin{aligned}\delta\boldsymbol{\alpha} - \left(\nabla_\beta \Phi^+ \mathbf{y}\right)\delta\boldsymbol{\beta} &= -\|\delta\boldsymbol{\beta}\|\left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}P\mathbf{y} + O\left(\|\delta\boldsymbol{\beta}\|^2\right),\\ &= \frac{\|\delta\boldsymbol{\beta}\|}{n}G^{-1}\left(\frac{d\Phi^T}{dt}P\left(\Phi(\overline{\boldsymbol{\beta}}) - \Phi(\boldsymbol{\beta})\right)\overline{\boldsymbol{\alpha}} - \Phi^T\frac{dP}{dt}\boldsymbol{\varepsilon}\right) + O\left(\|\delta\boldsymbol{\beta}\|^2\right).\end{aligned}$$

The magnitude of this resulting expression can be shown to be small almost surely compared with $\|\delta\boldsymbol{\beta}\|$ when n is large enough using the law of large numbers and consistency as before. The proximity of the increments in the linear parameters plus the identity of the calculation of the nonlinear parameter increments demonstrates the close alignment between the Kaufman and Gauss-Newton algorithms. The small residual result is discussed in [11].

5. Discussion. It has been shown that both of the variants of the Gauss-Newton algorithm considered possess similar convergence properties in large data set problems. However, that does not help resolve the question of the method of choice in any particular application. There is agreement that the Kaufman modification of the RGN algorithm has an advantage in being cheaper to compute, but it is not less expensive than the full Gauss-Newton algorithm [11]. Thus a choice between variable projection and Gauss-Newton must depend on other factors. These include flexibility, ease of use, and global behaviour. Flexibility tends to favour the full Gauss-Newton method because it can be applied directly to solve a range of maximum likelihood problems [7] so it has strong claims to be provided as a general purpose procedure. Ease of use is just about a draw. While Gauss-Newton requires starting values for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, given $\boldsymbol{\beta}$ the obvious approach is to compute $\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right)$ by solving the linear least squares problem. Selecting between the methods on some a priori prediction of effectiveness appears much harder. It is argued in [2] that variable projection can take fewer iterations in important cases. There are two significant points to be made here.

1. Nonlinear approximation families need not be closed. Especially if the data is inadequate then the iterates generated by the full Gauss-Newton may tend to a function in the closure of the family. In this case some parameter values will tend to ∞ and divergence is the correct answer. The nonlinear parameters can be bounded so it is possible for variable projection to yield a well determined answer. However, it still needs to be interpreted correctly. An example involving the Gauss-Newton method is discussed in [7].
2. There is some evidence that strategies which eliminate the linear parameters in separable models can be spectacularly effective in exponential fitting problems with small numbers of variables [5], [9]. Similar behaviour has not been observed for rational fitting [8] which is also a separable regression problem. It seems there is something else going on in the exponential fitting case as ill-conditioning of the computation of the linear parameters affects directly both the conditioning of the linear parameter correction in Gauss-Newton and the accuracy of the calculation of P_n in variable projection in both these classes of problems. It should be noted that maximum likelihood is not the way to estimate frequencies which are just the nonlinear parameters in a closely related problem [10]. Some possible directions for developing modified algorithms are considered in [3].

The importance of large sample behaviour, and the need for appropriate instrumentation for data collection are consequences of the result that maximum likelihood parameter estimates have the property that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \overline{\boldsymbol{\beta}}\right)$ is asymptotically normally distributed [12]. The effect of sample size on the convergence rate of the Gauss-Newton method is illustrated in Table 5.1 for an estimation problem involving fitting three Gaussian peaks plus an exponential background term. Such problems are common in scientific data analysis and are well

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 5.1
Iteration counts for peak fitting with exponential background

n	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$
64	7	16	nc
256	11	21	50
1024	7	17	18
4096	6	6	7
16384	6	6	7

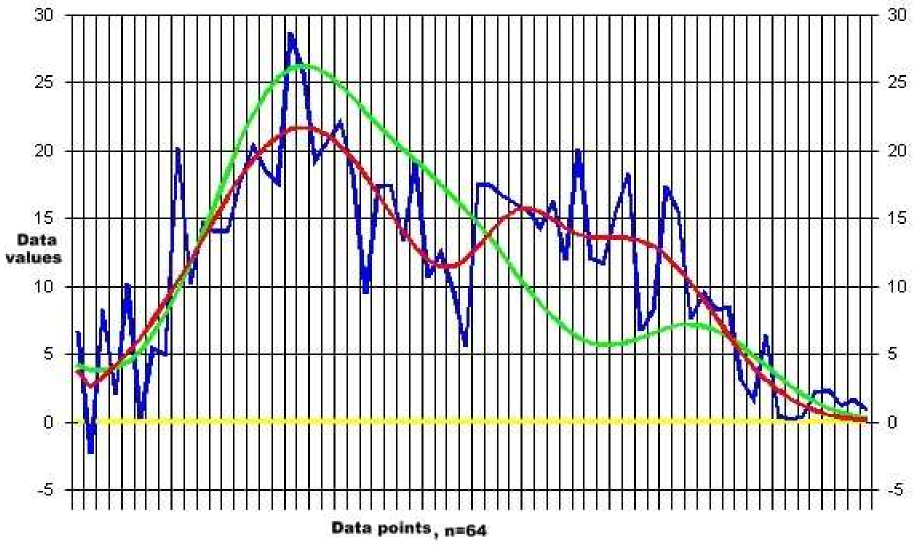

FIG. 5.1. *No convergence: fit after* 50 *iterations case* $\sigma = 4,\ n = 64$

enough conditioned if the peaks are reasonably distinct. In such cases it is relatively easy to set adequate initial parameter estimates. Here the chosen model is

$$\mu(\mathbf{x}, t) = 5e^{-10t} + 18e^{-\frac{(t-.25)^2}{.015}} + 15e^{-\frac{(t-.5)^2}{.03}} + 10e^{-\frac{(t-.75)^2}{.015}}.$$

Initial conditions are chosen such that there are random errors of up to 50% in the background parameters and peak heights, 12.5% in peak locations, and 25% in peak width parameters. Numbers of iterations are reported for an error process corresponding to a particular sequence of independent, normally distributed random numbers, standard deviations $\sigma = 1,\ 2,\ 4$, and equispaced sample points $n = 64,\ 256,\ 1024,\ 4096,\ 16384$. The most sensitive parameters prove to be those determining the exponential background, and they trigger the lack of convergence that occurred when $\sigma = 4,\ n = 64$. The apparent superior convergence behaviour in the $n = 64$ case over the $n = 256$ case for the smaller σ values can be explained by the sequence of random numbers generated producing more favourable residual values in the former case. The sequence used here corresponds to the first quarter of the sequence for $n = 256$.

Plots for the fits obtained for $\sigma = 4,\ n = 64$ and $\sigma = 4,\ n = 256$ are given in Figure 5.1 and Figure 5.2, respectively. The difficulty with the background estimation in the former shows up in the sharp kink in the fitted (red) curve near $t = 0$. This figure gives

ETNA
Kent State University
etna@mcs.kent.edu

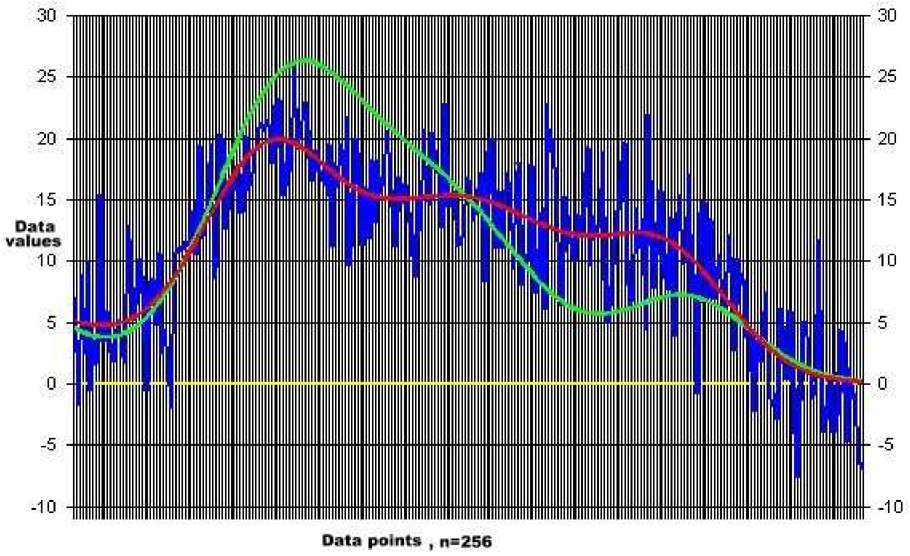

FIG. 5.2. *Fit obtained: case* $\sigma = 4,\ n = 256$

the result after 50 iterations when $x(1) = 269$ and $x(2) = 327$ so divergence of the background parameters is evident. However, the rest of the signal is being picked up pretty well. The quality of the signal representation suggests possible non-compactness, but the diverging parameters mix linear and nonlinear making interpretation of the cancelation occurring difficult. A similar phenomenon is discussed in [7]. This involves linear parameters only, and it is easier to see what is going on. The problem is attributed to lack of adequate parameter information in the given data. The green curves give the fit obtained using the initial parameter values and is the same in both cases. These curves manage to hide the middle peak fairly well, so the overall fits obtained are quite satisfactory. The problem would be harder if the number of peaks was not known a priori.

Appendix. The variational matrix whose spectral radius evaluated at $\widehat{\boldsymbol{\beta}}_n$ determines the convergence rate of the Kaufman iteration is

$$\begin{aligned} Q' &= I - \left(\frac{1}{n}K^T K\right)^{-1} \nabla_\beta^2 F \\ &= -\left(\frac{1}{n}K^T K\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} f_i \nabla_\beta^2 f_i + \frac{1}{n}L^T L\right). \end{aligned}$$

It is possible here to draw on work already done to establish the key convergence rate result (2.6). Lemmas 3.3 and 3.5 describe the convergence behaviour of $\mathcal{I}_n = \frac{1}{n}\left\{K^T K + L^T L\right\}$ as $n \to \infty$. Here it proves to be possible to separate out the properties of the individual terms by making use of the orthogonality of K and L, cf. (3.4), once it has been shown that $\frac{1}{n}\mathcal{E}\left\{L\left(\overline{\boldsymbol{\beta}}, \boldsymbol{\varepsilon}\right)^T L\left(\overline{\boldsymbol{\beta}}, \boldsymbol{\varepsilon}\right)\right\} \overset{a.s.}{\to} 0, n \to \infty$. This calculation can proceed as follows. Let

ETNA
Kent State University
etna@mcs.kent.edu

$\mathbf{t} \in \mathbb{R}^p$. Then

$$\begin{aligned}
\mathcal{E}\left\{\frac{1}{n}\mathbf{t}^T L^T L\mathbf{t}\right\} &= \frac{1}{n}\mathcal{E}\left\{\boldsymbol{\varepsilon}^T P\nabla_\beta \Phi\left[\mathbf{t}\right]\Phi^+\left(\Phi^+\right)^T \nabla_\beta \Phi\left[\mathbf{t}\right]^T P\boldsymbol{\varepsilon}\right\} \\
&= \frac{1}{n}\mathcal{E}\left\{\boldsymbol{\varepsilon}^T P\nabla_\beta \Phi\left[\mathbf{t}\right]\left(\Phi^T\Phi\right)^{-1} \nabla_\beta \Phi\left[\mathbf{t}\right]^T P\boldsymbol{\varepsilon}\right\} \\
&= \frac{1}{n^2}\,\mathrm{trace}\left\{\nabla_\beta \Phi\left[\mathbf{t}\right] G^{-1}\nabla_\beta \Phi\left[\mathbf{t}\right]^T P\mathcal{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right\}P\right\} + \text{ smaller terms} \\
&= \frac{\sigma^2}{n^2}\,\mathrm{trace}\left\{\nabla_\beta \Phi\left[\mathbf{t}\right] G^{-1}\nabla_\beta \Phi\left[\mathbf{t}\right]^T (I-T)\right\} + \text{ smaller terms.}
\end{aligned}$$

This last expression breaks into two terms, one involving the unit matrix and the other involving the projection T. Both lead to terms of the same order. The unit matrix term gives

$$\mathrm{trace}\left\{\nabla_\beta \Phi\left[\mathbf{t}\right] G^{-1}\nabla_\beta \Phi\left[\mathbf{t}\right]^T\right\} = \mathbf{t}^T\left\{\sum_{i=1}^n \Psi_i G^{-1}\Psi_i^T\right\}\mathbf{t},$$

where

$$\left(\Psi_i\right)_{jk} = \frac{\partial \phi_{ij}}{\partial \beta_k}, \quad \Psi_i : \mathbb{R}^m \to \mathbb{R}^p.$$

It follows that

$$\frac{\sigma^2}{n^2}\sum_{i=1}^n \Psi_i G^{-1}\Psi_i^T = O\left(\frac{1}{n}\right), \quad n \to \infty.$$

To complete the story note that the conclusion of Lemma 3.5 can be written

$$\frac{1}{n}\left(K^T K + L^T L\right) \overset{a.s.}{\to} \mathcal{E}\left\{\frac{1}{n}K^T K + \frac{1}{n}L^T L\right\}, \quad n \to \infty.$$

If $\frac{1}{n}K^T K$ is bounded, positive definite then, using the orthogonality (3.4),

$$\frac{1}{n}K^T K\left(\frac{1}{n}K^T K - \mathcal{E}\left\{\frac{1}{n}K^T K\right\}\right) \overset{a.s.}{\to} \frac{1}{n}K^T K\mathcal{E}\left\{\frac{1}{n}L^T L\right\}, \quad n \to \infty.$$

This shows that $\frac{1}{n}K^T K$ tends almost surely to its expectation provided it is bounded, positive definite for n large enough and so can be cancelled on both sides in the above expression. Note first that the linear parameters cannot upset boundedness.

$$\begin{aligned}
\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) &= \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{y} \\
&= \overline{\boldsymbol{\alpha}} + \frac{1}{n}\left(G^{-1} + O\left(\frac{1}{n}\right)\right)\Phi^T\boldsymbol{\varepsilon} \\
&= \overline{\boldsymbol{\alpha}} + \boldsymbol{\delta}, \quad \left\|\boldsymbol{\delta}\right\|_\infty = o\left(1\right),
\end{aligned}$$

where $\overline{\boldsymbol{\alpha}}$ is the true vector of linear parameters. Positive definiteness follows from

$$\begin{aligned}
\mathbf{t}K^T K\mathbf{t} &= \boldsymbol{\alpha}\left(\boldsymbol{\beta}\right)^T \frac{d\Phi}{dt}^T P\frac{d\Phi}{dt}\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) \\
&= \left\|\frac{d\Phi}{dt}\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right)\right\|^2 - \left\|T\frac{d\Phi}{dt}\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right)\right\|^2 \geq 0.
\end{aligned}$$

Equality can hold only if there is $\mathbf{t}$ such that $\frac{d\Phi}{dt}\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) = \gamma\Phi\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right)$. This condition was met also in Lemma 3.3.

REFERENCES

[1] G. GOLUB AND V. PEREYRA, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Numer. Anal., 10 (1973), pp. 413–432.

[2] ———, *Separable nonlinear least squares: the variable projection method and its applications*, Inverse Problems, 19 (2003), pp. R1–R26.

[3] M. KAHN, M. MACKISACK, M. OSBORNE, AND G. SMYTH, *On the consistency of Prony's method and related algorithms*, J. Comput. Graph. Statist., 1 (1992), pp. 329–350.

[4] L. KAUFMAN, *Variable projection method for solving separable nonlinear least squares problems*, BIT, 15 (1975), pp. 49–57.

[5] M. OSBORNE, *Some special nonlinear least squares problems*, SIAM J. Numer. Anal., 12 (1975), pp. 119–138.

[6] ———, *Fisher's method of scoring*, Internat. Statist. Rev., 86 (1992), pp. 271–286.

[7] ———, *Least squares methods in maximum likelihood problems*, Optim. Methods Softw., 21 (2006), pp. 943–959.

[8] M. OSBORNE AND G. SMYTH, *A modified Prony algorithm for fitting functions defined by difference equations*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 362–382.

[9] ———, *A modified Prony algorithm for exponential fitting*, SIAM J. Sci. Comput., 16 (1995), pp. 119–138.

[10] B. QUINN AND E. HANNAN, *The Estimation and Tracking of Frequency*, Cambridge University Press, Cambridge, United Kingdom, 2001.

[11] A. RUHE AND P. WEDIN, *Algorithms for separable nonlinear least squares problems*, SIAM Rev., 22 (1980), pp. 318–337.

[12] K. SEN AND J. SINGER, *Large Sample Methods in Statistics*, Chapman and Hall, New York, 1993.

[13] W. STOUT, *Almost Sure Convergence*, Academic Press, New York, 1974.