Curvature and Convergence

G.K. Smyth, Statistics and Applied Probability Program, University of California, Santa Barbara

Abstract

Fisher's method of scoring is probably the most important general algorithm in statistics. This paper picks out those aspects of curvature, normal and statistical, which are relevant to its convergence properties. For any particular data set the convergence of the algorithm near the maximum likelihood estimate depends on the eigenvalues of the convergence matrix, the derivative of the iteration function. In the least squares case, these eigenvalues can be interpretted as normal curvatures of one-dimensional curves on the response surface. Statistical curvature is shown to provide a before-the-data estimate of the squared sizes of the components of the convergence matrix. Bates and Watts (1980)'s intrinsic curvature is shown to correspond to the expected size of the convergence matrix in particular directions. A theme of the paper is that there is a close relationship between the convergence properties of the method of scoring, and the statistical properties of the model being fitted. This relationship is in large part due to mutual dependence on curvature.

Citation: Smyth, G.K. (1987). Curvature and convergence. Proceedings of the Statistical Computing Section. American Statistical Association, Virginia, 278–283.

1 Introduction

Statisticians in general seem to have given surprisingly little attention to the convergence properties of the algorithms they use to estimate nonlinear models. Jennrich (1969), Ruhe and Wedin (1980), Osborne (1987) and Osborne and Smyth (1987a,b) are amongst the small number of papers which examine such questions. This paper concentrates on Fisher's method of scoring, perhaps the most important general algorithm in statistics, and argues that there is an intimate connection between the the algorithm's convergence properties and the statistical properties of the models being fitted. The connection is that both depend on statistical curvature. Efron (1975), Hamilton, Watts and Bates (1982) and Amari (1985) have demonstrated the relevance of statistical curvature to nonlinear statistical inference. The aim of this paper is to show its relevance to the convergence of the scoring algorithm. Such a relationship is not surprising, since both the method of scoring and statistical curvature are concerned with quadratic approximations to the log-likelihood function, and with approximating the log-likelihood hessian with the Fisher information matrix.

The following two examples will make clear the sort of phenomenon we are interested in, and illustrate behaviour which is familiar to many applied statisticians. Table 1 gives the average number of iterations required for convergence by the Gauss-Newton algorithm on a simulated problem for various sample sizes and standard deviations. Observations were simulated to have means

$$\mathbb{E}(y_i) = \frac{\alpha_1 + \alpha_2 t_i}{1 + \beta_1 t_i + \beta_2 t_i^2}$$

with the t_i equally spaced on the unit interval, and constant variance σ^2. The actual parameter values were $\alpha_1 = \alpha_2 = .5$, $\beta_1 = -.5$ and $\beta_2 = .1$. It can be seen that the number of iterations required decreases as σ decreases and as n increases. In other words, fewer iterations are required if there is more information in the data set. The fact no more than 4 iterations were required even for large σ and small n, reflects the fact that rational fitting is quite close to being a linear problem.

The second example is more dramatic. Data were simulated as before, except that the means were given by

$$\mathbb{E}(y_i) = \alpha_1 + \alpha_2 e^{-\beta_1 t_i} + \alpha_3 e^{\beta_2 t_i}.$$

Table 1: The number of iterations to convergence for the Gauss-Newton algorithm applied to a rational fitting problem. Median and maximum over 10 simulated data sets.

$n\backslash\sigma$	.030		.010		.003		.001	
32	4	5	3	3	3	3	2	3
64	3	5	3	4	3	3	2.5	3
128	3	5	3	3	2.5	3	2	3
256	3	3	2	3	2	3	2	2
512	3	3	2	3	2	3	2	2

Table 2: Median and maximum iteration counts for the Levenberg and Prony algorithms applied to exponential fitting. Results for the Levenberg algorithm are below those for Prony. The maximum number of iterations allowed for the Levenberg algorithm was 40.

$n\backslash\sigma$	0.030		0.010		0.003		0.001	
32	6	11	4	6	3	4	3	3
	40	40	33	40	26	40	16	40
64	4	8	3	4	2	3	2	2
	32.5	40	31.5	40	20	40	13	22
128	3	3	2	3	2	2	1.5	2
	16.5	40	10	40	8	34	6	18
256	2	3	2	2	1	1	1	1
	30	30	20	40	14	32	10	12
512	1	1	1	1	1	1	1	1
	36.5	40	19.5	40	13	22	7.5	12

The actual parameter values were $\alpha_1 = .5$, $\alpha_2 = 2$, $\alpha_3 = -1.5$, $\beta_1 = 4$ and $\beta_2 = 7$. Exponential function fitting is a highly nonlinear problem, with which the straightfoward application of the Gauss-Newton algorithm would have had little success. The algorithms used instead were the Levenberg modification of Gauss-Newton (Osborne, 1976), and a more special purpose algorithm based on Prony's parametrization described by Osborne and Smyth (1987b). Table 2 shows the same pattern in the number of iterations required as was observed for the rational example. The decrease in the number of iterations for large n is especially dramatic for the modified Prony algorithm. For the Levenberg algorithm this effect is somewhat obscured by its very careful convergence criterion, which caused the algorithm to return increasingly precise estimates for sample sizes over 128.

2 The Method of Scoring

Consider the iterative process

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

in which the iteration function F updates the current estimate $\boldsymbol{\theta}^k \in \mathbb{R}^p$ to the new value $\boldsymbol{\theta}^{k+1}$. The convergence of this process depends on the derivative matrix G with elements

$$G_{ij} = \frac{\partial F_i}{\partial \theta_j}.$$

Ostrowski's Theorem (Ortega and Rheinboldt, 1970) asserts that a stationary point $\hat{\boldsymbol{\theta}}$ is a *point of attraction* of the iterative process if the spectral radius $\rho(G(\hat{\boldsymbol{\theta}}))$ is less than one. In that case, $\rho(G(\hat{\boldsymbol{\theta}}))$ is the ultimate rate of convergence $\lim_{k\to\infty} \|\boldsymbol{\theta}^{k+1} - \hat{\boldsymbol{\theta}}\|/\|\boldsymbol{\theta}^k - \hat{\boldsymbol{\theta}}\|$.

The Newton-Raphson iteration to maximize a log-likelihood function $\ell(\boldsymbol{\theta})$ can be defined by

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

where $\dot{\ell}$ is the gradient or score vector and $\ddot{\ell}$ is the hessian. Then $\boldsymbol{\theta}^{k+1}$ maximizes the quadratic expansion of ℓ at $\boldsymbol{\theta}^k$. The method of scoring iteration

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \mathcal{I}^{-1}\dot{\ell} \tag{1}$$

replaces $-\ddot{\ell}$ by its expectation, the Fisher information matrix $\mathcal{I}$. Differentiating (1) at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ gives the convergence matrix of the scoring iteration as

$$\begin{aligned} G &= I - \mathcal{I}^{-1}\dot{\mathcal{I}}\mathcal{I}^{-1}\dot{\ell} + \mathcal{I}^{-1}\ddot{\ell} \\ &= \mathcal{I}^{-1}(\ddot{\ell} + \mathcal{I}) \end{aligned}$$

since $\dot{\ell}(\hat{\boldsymbol{\theta}}) = 0$. The eigenvalues of G are invariant under reparametrization, which means that reparametrization cannot change the ultimate rate of convergence of the scoring iteration.

The log-likelihood kernel for a normal sample (ignoring the variance) is

$$\ell = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{y} - \boldsymbol{\mu})$$

so that

$$\begin{aligned} \dot{\ell} &= \dot{\boldsymbol{\mu}}^T(\mathbf{y} - \boldsymbol{\mu}), \\ \ddot{\ell} &= -\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}} + \ddot{\boldsymbol{\mu}}^T(\mathbf{y} - \boldsymbol{\mu}) \end{aligned}$$

and

$$\mathcal{I} = \dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}.$$

Here $\mathbf{y}$ is the vector of observations, $\boldsymbol{\mu}$ is the vector of fitted values, $\dot{\boldsymbol{\mu}}$ is the $n \times p$ gradient matrix with elements

$$\dot{\boldsymbol{\mu}}_{ij} = \frac{\partial \mu_i}{\partial \theta_j}$$

and $\ddot{\boldsymbol{\mu}}$ is a 3-dimensional array such that

$$\left(\ddot{\boldsymbol{\mu}}^T(\mathbf{y} - \boldsymbol{\mu})\right)_{jk} = \sum_{i=1}^{n} \frac{\partial^2 \mu_i}{\partial \theta_j \partial \theta_k}(y_i - \mu_i).$$

In this case maximum likelihood estimation is equivalent to least squares, and the method of scoring is called the Gauss-Newton algorithm. It has convergence matrix

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

3 Geometry of Least Squares

The least squares problem consists of finding that point $\hat{\boldsymbol{\mu}}$ on the response surface $\{\boldsymbol{\mu}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ closest to $\mathbf{y}$ in n-dimensional Euclidean space. An important role is played by the tangent plane $T = \mathcal{R}(\dot{\boldsymbol{\mu}})$ to the response surface at $\hat{\boldsymbol{\mu}}$. It is clear that curvature in this situation means curvature of the response surface. To quantify this we define our notion of curvature, and consider one-dimensional curves on the response surface.

Let f be a function mapping $\mathbb{R}$ into $\mathbb{R}^n$. The range of f then defines a one-dimensional curve in n-space. Consider the limiting circle through the points $f(\alpha - \epsilon)$, $f(\alpha)$ and $f(\alpha + \epsilon)$ as $\epsilon \to 0$. The *normal curvature* at α may be defined to be the inverse radius of this limiting circle, and can be calculated as

$$\frac{\|P_N\ddot{f}\|}{\dot{f}^T\dot{f}}$$

where P_N is the projection onto $N = \mathcal{R}(\dot{f}(\alpha))^{\perp}$ (Johansen, 1984).

Now consider one-dimensional curves on the response surface. Corresponding to any direction $\mathbf{v} \in \mathbb{R}^p$ there is a line $\boldsymbol{\theta}(\alpha) = \hat{\boldsymbol{\theta}} + \alpha\mathbf{v}$ from $\hat{\boldsymbol{\theta}}$ in the parameter space, and a lifted one-dimensional curve on the response surface defined by $\boldsymbol{\mu}(\boldsymbol{\theta}(\alpha))$. We have that

$$\frac{d\boldsymbol{\mu}}{d\alpha} = \dot{\boldsymbol{\mu}}\mathbf{v}$$

and

$$\frac{d^2\boldsymbol{\mu}}{d\alpha^2} = \mathbf{v}^T\ddot{\boldsymbol{\mu}}\mathbf{v}.$$

Furthermore the projection of $\mathbf{v}^T\ddot{\boldsymbol{\mu}}\mathbf{v}$ onto $\mathcal{R}(\dot{\boldsymbol{\mu}}\mathbf{v})$ is the same as its projection onto $\mathcal{R}(\dot{\boldsymbol{\mu}})$. So the normal curvature of the lifted curve is

$$\kappa(\mathbf{v}) = \frac{\mathbf{v}^T P_N \ddot{\boldsymbol{\mu}}\mathbf{v}}{\mathbf{v}^T\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}\mathbf{v}} \tag{2}$$

where P_N is now the projection onto $\mathcal{R}(\dot{\boldsymbol{\mu}})^\perp$. (In the terminology of differential geometry, $P_N\ddot{\boldsymbol{\mu}}$ is the *second fundamental form* of the response surface, the information matrix $\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}$ being the first.) An essentially equivalent derivation of (2) would have arisen had we defined the one-dimensional curve as the intersection of $\mathcal{R}(\dot{\boldsymbol{\mu}}\mathbf{v})$ with the solution locus, that is as a *normal cut*. That approach would exhibit κ as a function of the tangent direction $\dot{\boldsymbol{\mu}}\mathbf{v}$. As a function of $\dot{\boldsymbol{\mu}}\mathbf{v}$, κ is a geometric invariant.

Now we return to the convergence matrix. The eigenvalues of G are the stationary values of the Rayleigh quotient

$$q(\mathbf{v}) = \frac{\mathbf{v}^T\ddot{\boldsymbol{\mu}}^T(\mathbf{y}-\boldsymbol{\mu})\mathbf{v}}{\mathbf{v}^T\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}\mathbf{v}}. \tag{3}$$

Let $\mathbf{e} = (\mathbf{y}-\boldsymbol{\mu})/\|\mathbf{y}-\boldsymbol{\mu}\|$, and let P_e be the projection onto $\mathcal{R}(\mathbf{e})$. Then at the least squares estimate

$$q(\mathbf{v}) = \frac{\|\mathbf{v}^T P_e\ddot{\boldsymbol{\mu}}\mathbf{v}\|}{\mathbf{v}^T\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}\mathbf{v}}\hat{\sigma}$$

Let $\lambda_1 < \ldots < \lambda_p$ be the eigenvalues of G with eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$. We see that $\lambda_i/\hat{\sigma}$ is the normal curvature at $\alpha = 0$ of the curve $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}} + \alpha\mathbf{x}_i)$ imbedded in the space spanned by $\dot{\boldsymbol{\mu}}$ and $\mathbf{e}$. We call the $\lambda_i/\hat{\sigma}$ the *normal curvatures* of the solution locus at $\hat{\boldsymbol{\theta}}$ relative to $\mathbf{e}$.

4 The Convergence Matrix

From the form of G in the least squares case, we can immediately explain the pattern of convergence behaviour observed in the examples in the introduction:

(i) G is directly proportional to σ.

(ii) $G \to 0$ as $n \to \infty$. This can be seen from the expansion

$$G = \left(\frac{1}{n}\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}\right)^{-1}\frac{1}{n}\ddot{\boldsymbol{\mu}}^T(\mathbf{y}-\boldsymbol{\mu})$$

in which $\frac{1}{n}\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}$ converges to a positive definite limit, while $\frac{1}{n}\ddot{\boldsymbol{\mu}}^T(\mathbf{y}-\boldsymbol{\mu})$ converges to zero by the law of large numbers.

In general, $G \to 0$ under the same sort of conditions that imply that $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal (see Jennrich, 1969). Consider the approximation, resulting from a quadratic expansion of the log-likelihood, which is usually used to establish the asymptotic distribution of the maximum likelihood estimators

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx \ddot{\ell}(\boldsymbol{\theta}_0)^{-1}\dot{\ell}(\boldsymbol{\theta}_0).$$

The standard result is to apply some form of the central limit theorem to $n^{-1/2}\dot{\ell}$, and some form of the law of large numbers to $\frac{1}{n}\ddot{\ell}$ to show that

$$\frac{1}{n}(\ddot{\ell}(\boldsymbol{\theta}_0) + \mathcal{I}(\boldsymbol{\theta}_0)) \to 0$$

and $\frac{1}{n}\mathcal{I}$ has a nonsingular limit. Under these conditions,

$$G = \mathcal{I}^{-1}(\ddot{\ell} + \mathcal{I}) \to 0.$$

5 Generalized Linear Models

Generalized linear models assume that the mean and variance structure of the observations are given by

$$g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$$

and

$$\mathrm{var}(y_i) = \phi v(\mu_i)$$

where X is some regression matrix, g is the *link function* and v is some *variance function* determined by the distribution. The derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ are

$$\dot{\ell}_\beta = X^T\mathrm{diag}\left(\frac{1}{\phi\dot{g}v}\right)(\mathbf{y}-\boldsymbol{\mu})$$

and

$$\ddot{\ell}_\beta = -X^T\mathrm{diag}\left(\tfrac{1}{\phi\dot{g}^2v}\right)X - X^T\mathrm{diag}\left(\left(\tfrac{\ddot{g}}{\dot{g}^2v\phi} + \tfrac{\dot{v}}{\dot{g}v^2\phi}\right)(y_i-\mu_i)\right)X$$

where v, g, $\dot{g}$ and $\ddot{g}$ are understood to be functions of μ_i. The curvature matrix represents the size of the second term of $\ddot{\ell}_\beta$ relative to the first. If g is the canonical link then the second term is exactly zero, which can be most easily seen by substituting $\dot{g} = 1/v$ into $\dot{\ell}_\beta$. This reflects the fact that the scoring iteration for generalized linear models with canonical link is quadratically convergent.

Some cancellation occurs in the curvature matrix for the variance stabilizing link also (for example for the log-link for the gamma distribution) for which $\dot{g} = 1/v^{1/2}$. For such a link we have

$$\ddot{\ell}_\beta = -\frac{1}{\phi}X^TX - \frac{1}{\phi}X^T\mathrm{diag}\left(\frac{\dot{v}(y_i-\mu_i)}{2v^{3/2}}\right)X.$$

6 Nested Iterations

Consider the problem of fitting the function

$$\mu(t) = \alpha_1 e^{-\beta_1 t} + \alpha_2 e^{-\beta_2 t}$$

to data by least squares. The mean vector can be written

$$\boldsymbol{\mu} = A(\boldsymbol{\beta})\boldsymbol{\alpha}$$

where A is the $n \times p$ matrix function with elements $A_{ij} = e^{-\beta_j t_i}$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. The sum of squares

$$\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{y}-\boldsymbol{\mu})$$

is maximized for fixed $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}) = (A^TA)^{-1}A^T\mathbf{y}.$$

Substituting this into the sum of squares, gives the reduced object function

$$\psi(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}) = \mathbf{y}^T(I - P_A)\mathbf{y}$$

where P_A is the orthogonal projection onto $\mathcal{R}(A)$, which can be maximized with respect to $\boldsymbol{\beta}$. This is an example of *separable regression* in which the α_i are *linear parameters* and the β_j are *nonlinear*.

In general, consider maximum likelihood estimation of the partitioned parameter vector

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$$

and let $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ maximize the likelihood for fixed $\boldsymbol{\theta}_2$. Then the Newton-Raphson iteration to maximize the reduced likelihood $\ell(\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$ is

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

with $\ddot{\ell}_{2.1} = \ddot{\ell}_2 - \ddot{\ell}_{21}\ddot{\ell}_{11}^{-1}\ddot{\ell}_1$ (Richards, 1961). Replacing $-\ddot{\ell}$ with the Fisher information matrix $\mathcal{I}$ gives *nested iterations*

$$\boldsymbol{\theta}_2^{k+1} = \boldsymbol{\theta}_2^k + \mathcal{I}_{2.1}^{-1}\dot{\ell}_2$$

with $\mathcal{I}_{2.1} = \mathcal{I}_2 - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}$ (Smyth, 1985). The nested iteration process has the convergence matrix

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

which has spectral radius not greater than that of the full method of scoring applied to the full parameter vector (Smyth, 1985). In effect, in separating out $\boldsymbol{\theta}_1$ we have restricted the relevant curvatures to the curve $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ rather than having the consider to full response surface.

7 Expected Curvature

The concept of *statistical curvature*, as defined by Efron (1975), Bates and Watts (1980), Amari (1985) and others, is a property of a family of probability distributions rather than a property of a particular observed data set. It corresponds to the question: before we observe the data, how large should we expect $\rho(G)$ to be, given the probability model? We work with the symmetric curvature

$$B = \mathcal{I}^{-\frac{1}{2}}(\ddot{\ell} + \mathcal{I})\mathcal{I}^{-\frac{T}{2}},$$

to which G is similar, and investigate the expected sizes of its eigenvalues. There are at least two possibilities:

(i) The components B_{ij} of B have expected value zero. If we can calculate their expected squared sizes, that is their variances, then we can calculate the expected Frobenius norm

$$S_2 = \mathbb{E}(\|B\|_F) = \sum_{i,j=1}^{p} \mathbb{E}(B_{ij}^2)$$

which can in turn be used to bound the maximum eigenvalue by

$$\frac{1}{p}S_2 \le \mathbb{E}(\rho^2(B)) \le S_2.$$

In many cases, S_2 is a tight bound for $\rho(G)^2$.

(ii) Under regularity conditions, the B_{ij} are asymptotically normal with mean zero. Given the dispersion matrix of $\mathrm{vec}(B)$ and assuming normality, one can in principle write down the distribution of the eigenvalues of B as described by Muirhead (1982). Obtaining a practically useful distribution for the maximum eigenvalue is difficult however, so the approach is not taken further in this paper.

Consider first the case of least squares. Hamilton, Watts and Bates (1982) call

$$B = (\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}})^{-\frac{1}{2}}\ddot{\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu})(\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}})^{-\frac{T}{2}}$$

the *effective residual curvature matrix*, and show that its eigenvalues can be used to obtain improved approximations to inference regions in nonlinear least squares. We seek the expected squared sizes of its components B_{ij} given the fact that a particular value of $\boldsymbol{\theta}$ is the least squares estimate. Now $\boldsymbol{\theta}$ will be the least squares estimate if $\mathbf{y}$ solves $\dot{\boldsymbol{\mu}}^T(\mathbf{y} - \boldsymbol{\mu}) = 0$, and the conditional distribution of $\mathbf{y}$ given this is $N(\boldsymbol{\mu}, P_N\sigma^2)$ where

$$P_N = I - \dot{\boldsymbol{\mu}}(\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}})^{-1}\dot{\boldsymbol{\mu}}^T$$

is the projection matrix onto the linear space orthogonal to $\dot{\boldsymbol{\mu}}$. Under this distribution, the elements of B are normal, and the variances and covariances of its elements are given by

$$\mathbb{E}(B\otimes B) = (L^{-1}\otimes I)(I\otimes L^{-1})\ddot{\boldsymbol{\mu}}^T P_N \ddot{\boldsymbol{\mu}}(I\otimes L^{-T})(L^{-T}\otimes I)$$

where L is the Choleski factor satifying

$$\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}} = LL^T.$$

The expected squared size of B in a particular direction $\mathbf{v}$ is given by

$$\mathbb{E}(q(\mathbf{v})^2) = \sigma^2\frac{\|\mathbf{v}^T P_N\ddot{\boldsymbol{\mu}}\mathbf{v}\|^2}{(\mathbf{v}^T\dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}\mathbf{v})^2} = \sigma^2\kappa^2(\mathbf{v}),$$

that is by the normal curvature in that direction. Except for scaling, the $\kappa(\mathbf{v})$ is the *intrinsic curvature* in the direction $\mathbf{v}$ of Bates and Watts (1980). Their *relative* intrinsic curvature is

$$\gamma(\mathbf{v}) = \sigma p^{1/2}\kappa(\mathbf{v}).$$

They construct summary measures of (relative) instrinsic curvature by maximizing or averaging over the possible directions to obtain

$$\Gamma = \max_v \gamma(\mathbf{v})$$

and

$$\gamma_{RMS}^2 = \mathbb{E}_v\gamma^2(\mathbf{v}),$$

the expectation being taken over $\mathbf{v} \sim N(0, \dot{\boldsymbol{\mu}}^T\dot{\boldsymbol{\mu}}^{-1})$. Since

$$\max_v \mathbb{E}(q(\mathbf{v})^2) \le \mathbb{E}(\max_v q(\mathbf{v})^2)$$

we have that

$$\frac{1}{p}\gamma_{RMS}^2 \le \frac{1}{p}\Gamma^2 \le \mathbb{E}(\rho(B)^2).$$

Table 3: Maximum intrinsic curvature, and the extreme eigenvalues and spectral radius of the convergence matrix, for 18 data sets.

Data set	Γ	$\lambda_1(B)$	$\lambda_p(B)$	$\rho(B)$
1	.03	−.00	.00	.00
2	.06	−.00	.00	.00
3	.08	−.17	.00	.17
4	.07	−.10	.00	.10
5	.21	−.00	.10	.10
9	.18	−.06	.02	.06
13	.01	−.00	.00	.00
14	.15	−.00	.08	.08
15	.04	−.04	.00	.04
16	.04	−.00	.02	.02
17	.25	−.00	.15	.15
18	.00	−.00	.00	.00
19	.02	−.02	.00	.02
20	.02	−.06	.00	.06
21	.90	−.06	.26	.26
22	.08	−.19	.00	.19
23	.09	−.02	.10	.10
24	.37	−.04	.10	.10

One might use Γ^2 itself as a conservative estimate of $\mathbb{E}(\rho(B)^2)$, a relationship which seems credible from Table 3. (Table 3 was constructed from tables in Bates and Watts (1980) and Hamilton, Watts and Bates (1982).) It does not seem possible to relate the Bates and Watts summary measures of curvature to the size of the convergence matrix more closely than this. Bates and Watts (1980) themselves compare Γ^2 with $1/F(p, n-p; .95)$, which can be seen to be quite conservative from a convergence point of view.

8 A Numerical Example

Suppose that observations are independent and normally distributed with means

$$\mu(t) = \alpha e^{-\beta t}$$

for times $t_1, \ldots, t_n$ equally spaced on the unit interval, and constant variance σ^2. The function $\mu(t)$ has partial derivatives

$$\begin{aligned} \dot{\mu}_\alpha &= e^{-\beta t} \\ \dot{\mu}_\beta &= -t\alpha e^{-\beta t} \\ \ddot{\mu}_\alpha &= 0 \\ \ddot{\mu}_{\alpha\beta} &= -te^{-\beta t} \\ \ddot{\mu}_\beta &= t^2\alpha e^{-\beta t}, \end{aligned}$$

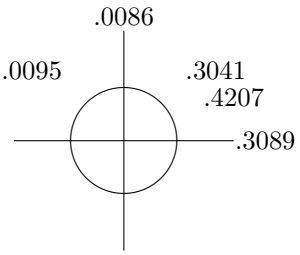
which shows that $\ddot{\mu}_\beta$ is the only second derivative which in not linearly dependent on the first derivatives. Because of this, the matrix $\ddot{\boldsymbol{\mu}} P_N \ddot{\boldsymbol{\mu}}$ has just one non-zero element. For $n = 8$ and $\alpha = \beta = \sigma = 1$ we have

$$\ddot{\boldsymbol{\mu}} P_N \ddot{\boldsymbol{\mu}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .01312 \end{pmatrix}$$

and $\mathbb{E}(\|B\|_F^2) = .4207$. For these parameter values, convergence will be a problem even very close to the least squares estimate a significant proportion of the time. The expected Frobenius norm depends on α and β in the following way:

$\beta\backslash\alpha$	.50	1.00	2.00
.50	.76	.26	.16
1.00	1.34	.42	.21
2.00	4.16	1.18	.46

We see that the curvature decreases with α but increases with β — that is, it decreases with the signal to noise ratio. Returning to $\alpha = \beta = 1$, the following picture displays (squared) curvatures in particular directions,



the directions being in terms of $\mathbf{d} = (\dot{\boldsymbol{\mu}}^T \dot{\boldsymbol{\mu}})^{1/2}\mathbf{v}$. In this case the maximum curvature Γ^2 is equal to the expected Frobenius norm, because B has only one nonzero eigenvalue.

9 Statistical Curvature

In a seminal paper, Efron (1975) defined *statistical curvature* γ for one parameter probability families, to be

$$\gamma^2 = \frac{\mathrm{var}(\ddot{\ell}|\dot{\ell})}{\mathcal{I}^2}$$

where

$$\ddot{\ell}|\dot{\ell} = \ddot{\ell} - \frac{\mathrm{cov}(\dot{\ell}, \ddot{\ell})}{\mathrm{var}(\dot{\ell})}\dot{\ell}$$

is the log-likelihood hessian $\ddot{\ell}$ corrected for linear regression on the first derivative $\dot{\ell}$. One motivation for the definition is that the statistical curvature of one parameter *curved exponential families* with log-densities

$$\boldsymbol{\theta}(\alpha)^T \mathbf{y} - \psi(\boldsymbol{\theta}) + f(\mathbf{y})$$

where $\boldsymbol{\theta}$ is an n-dimensional function of the scalar parameter α, is the ordinary normal curvature of the function $\boldsymbol{\theta}(\cdot)$ relative the the inner product defined by the covariance matrix of $\mathbf{y}$.

Efron (1975) demonstrated the relevance of γ^2 to second order efficiency and other aspects of statistical inference. In particular, the variance of a bias-corrected maximum likelihood estimate can be written as the Cramèr-Rao lower bound plus a term that depends in γ plus the Bhattacharyya correction or *naming curvature* plus terms of $O(1/n^3)$. The Cramér-Rao term is $O(1/n)$ while the second two are $O(1/n^2)$. Intuitively, $\gamma^2/\mathcal{I}$ is the amount of information lost when summarizing the data with the maximum likelihood estimate $\hat{\alpha}$. Reeds (1975) and Dawid (1975) suggested generalizations to multi-parameter models, which have been taken further by Amari (1982, 1985).

From the point of view of convergence, we need to know the expected (squared) size of the symmetric convergence matrix

$$B = \mathcal{I}^{-1/2}(\ddot{\ell} + \mathcal{I})\mathcal{I}^{-T/2}.$$

So let $\gamma_{ij,kl}$ be the covariance between the components B_{ij} and B_{kl} of B, taken over the conditional distribution of $\ddot{\ell}$ corrected for linear regression on $\dot{\ell}$. Then the $O(1/n^2)$ correction due to statistical curvature to the variance of the maximum likelihood estimator is given by the positive definite matrix A, with components

$$A_{mn} = \sum_{j,k=1}^{p} \mathcal{I}^{jk}\gamma_{ij,kl}$$

where the $\mathcal{I}^{jk}$ are the components of $\mathcal{I}^{-1}$ (Reeds, 1975). Asymptotically, $\ddot{\ell}$ and $\dot{\ell}$ are normally distributed, so that the distribution of $\ddot{\ell}$ corrected for linear regression on $\dot{\ell}$ is asymptotically the conditional distribution of $\ddot{\ell}$ given $\dot{\ell}$. Since the maximum likelihood estimate satisfies $\dot{\ell} = 0$, we are effectively taking expections conditional on the maximum likelihood estimate being a specified value.

This γ is a natural generalization of Efron (1975)'s one-dimensional measure of statistical curvature, and is essentially equivalent to the *exponential curvature* of Amari (1985). It holds all the second order information about the expected size of the convergence matrix for the method of scoring. In particular, the expected Frobenius norm is

$$\mathbb{E}_{\ddot{\ell}|\dot{\ell}}(\|B\|_F) = \sum_{ij}^{p} \gamma_{ii,jj}$$

which can be used to bound the expected squared spectral radius.

References

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.

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

Dawid (1975). Discussion on Professor Efron's paper. *Ann. Statist.* 3:1231–1234.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3:1189–1242.

Hamilton, D. C., Watts, D. G. and Bates, D. M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. *Ann. Statist.* 10:386–393.

Johansen, S, (1984). *Functional relations, random coefficients and nonlinear regression with application to kinetic data.* Springer-Verlag, New York.

Jennrich, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* 40:633–643.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.* Wiley, New York.

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

Osborne, M. R. (1976). Nonlinear least squares — the Levenberg algorithm revisited. *J. Austral. Math. Soc. B* 19:343–357.

Osborne, M. R. (1987). Estimating nonlinear models by maximum likelihood for the exponential family. *SIAM J. Sci. Statist. Comp.* 8:446–456.

Osborne, M. R. and Smyth, G. K. (1987). A modified Prony algorithm I: rational fitting. Technical Report No. 22, Statistics & Applied Probability, University of California, Santa Barbara.

Osborne, M. R. and Smyth, G. K. (1987). A modified Prony algorithm II: exponential function fitting. Technical Report No. 30, Statistics & Applied Probability, University of California, Santa Barbara.

Reeds, J. (1975). Discussion on Professor Efron's paper. *Ann. Statist.* 3:1234–1238.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *J. R. Statist. Soc. B* 23:469-475.

Ruhe, A. and Wedin, P. A. (1980). Algorithms for separable nonlinear least squares problems. *SIAM Rev.* 22:318–337.

Smyth, G. K. (1985). Coupled and separable iterations in nonlinear estimation. PhD thesis, Australian National University, Canberra.