Algorithms for Nonlinear Generalized Cross–Validation

Brian S. Yandell, University of Wisconsin–Madison

A variety of penalized nonlinear problems can be expressed as the iterated solution to a nonlinear minimization, in which the inner step involves minimizing a penalized weighted least squares expression. We propose algorithms when matrices in the least squares problem may depend on the unknown parameters. The problems in increasing complexity are (a) generalized linear models, (b) iterated reweighted least squares, and (c) general nonlinear problems. The algorithms are built around GCVPACK (Bates, Lindstrom, Wahba and Yandell, 1985), a package for generalized cross-validation, using a balance of Cholesky and singular value decompositions which is adjusted depending on the type of problem.

1. Introduction

A variety of penalized nonlinear problems can be expressed as the iteration to a solution of a nonlinear minimization, in which the inner step involves minimizing a quadratic form such as

$$\frac{1}{n} \| \mathbf{W}^{\mathrm{T}}(\mathbf{y}-\mathbf{S}\alpha-\mathbf{T}\beta-\mathbf{K}\delta) \|^2 + \lambda\delta^{\mathrm{T}}\mathbf{K}_U\delta \qquad (1.1)$$

in which **S**, **T** and **K** are the design matrices for the covariates, polynomial and "smooth" part of the model, and **y** and **W** are the responses and the weights. The simplest form is the partial spline model, or semi-parametric linear model,

$$y_i = S_i^{\mathrm{T}}\alpha + f(\mathbf{x}_i) + \varepsilon_i \ , i = 1, \cdots, n \ , \qquad (1.2)$$

in which $f(\cdot)$ is some "smooth" function and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^{\mathrm{T}}$ has covariance matrix $(\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}$ which is usually diagonal. We present three situations and proposed computational solutions when matrices in the above linearized problem may depend on the unknown parameters. The problems in increasing degree of complexity are:

(1) Semi-parametric generalized linear models, in which **S**, **T**, **K** and $\mathbf{K}_U$ are constant, while **W** and **y** may change with each iteration.

(2) Iteratively reweighted least squares, in which only $\mathbf{K}_U$ remains constant.

(3) General nonlinear problems (remote sensing, for example), in which all matrices may change with each iteration.

Different compromises are suggested by each problem. Clearly, one would like to decompose the constant matrices exactly once and would like to keep decompositions of the changing matrices as cheap as possible. The method proposed here combines the advantages of SVD in locating the generalized cross validation choice of λ with Cholesky decompositions which are relatively cheap once λ is fixed. While the decompositions suggested are not new, the combination of approaches appears to be an unexplored area. The basic strategy is as follows:

(1) guess at initial λ ($=\infty$) and $(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \delta^{\mathrm{T}})$
(2) CD: iterate (part-way) to solution for fixed λ
(3) linearize the problem as in (1.1)
(4) SVD: pick optimal λ via GCV
(5) iterate (2)-(4) to convergence

Convergence criteria can include absolute or relative convergence of the regularization functional and/or the parameter estimates, and absolute convergence of $\log(n\lambda)$. The number of iterations in (2) may be restricted, leading to rough estimates which are fed into (3).

We do not assume any special structure to the design or the matrices, except that we suppose that **W** is of full rank, and computationally invertible. In many cases, **W** is actually diagonal, but this will not be explicitly used in the linear algebra.

Algorithms for the linear model (1.2) have been given by many authors, most recently in the multivariate form by Bates et al. (1985). The algorithms below are extensions of Bates et al. (1985), building on their Fortran77 package, GCVPACK.

2. Semi-Parametric Generalized Linear Models

For semi-parametric generalized linear models (SGLM), one has a parameter vector θ which consists of a parametric piece and a "smooth" nonparametric piece,

$$\theta_i = S_i^{\mathrm{T}}\alpha + f(\mathbf{x}_i) \ , \ i = 1, \cdots, n \ .$$

One can formulate the problem as minimizing, for fixed λ,

$$S_\lambda(\theta) = L(\theta) + \lambda J(\theta) \ .$$

in which L is the log likelihood and J is the smoothing penalty (see Good and Gaskins (1971); Leonard (1982); Green, Jennison and Seheult (1983); O'Sullivan, Yandell and Raynor (1986); Green and Yandell (1985)). We know from O'Sullivan (1983) that if $L(\theta)$ is suitably convex and $J(\theta)$ is a quadratic form (*e.g.*, the squared norm of a projection), then $S_\lambda(\theta)$ has a unique minimum for each λ. These conditions appear to hold for many generalized linear models.

One can choose λ to minimize the GCV criterion (Craven and Wahba, 1979). which is "close" to minimizing the predictive mean square error (see Craven and Wahba (1979); Speckman (1985); Cox (1983)). What we propose to do here is to iterate on θ *and* λ, to find the $\hat{\lambda}$ which is the GCV minimizer and the $\hat{\theta}$ which minimizes $S_{\hat{\lambda}}(\hat{\theta})$. It is not known whether such a procedure will converge, but we conjecture that, if the GCV minimizer is bounded away from 0 and ∞ and L is suitable convex, then it does converge.

The log likelihood can be written in an iterative form using pseudo-values **y** and pseudo-weights **W**,

$$\begin{aligned} \mathbf{W}\mathbf{W}^{\mathrm{T}} &= E\left[-\frac{\partial^2 L}{\partial\theta\partial\theta^{\mathrm{T}}}\right]_{\theta^o} \\ \mathbf{y} &= \theta^o + (\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\left[\frac{\partial L}{\partial\theta}\right]_{\theta^o} , \end{aligned} \qquad (2.1)$$

based on θ^o from the previous iteration. Note that for the independent normal model, $\mathbf{W}^{-1}$ is a diagonal matrix of the standard deviations and **y** is the vector of observed responses. The linearized log likelihood is

$$L(\theta) = \frac{1}{n} \| \mathbf{W}^{\mathrm{T}}(\mathbf{y}-\theta) \|^2 \ .$$

The penalty J can often be written in a nonnegative definite quadratic form in δ (see Green and Yandell (1985)). We follow the spline literature and formulate it as

Reprinted from Computer Science and Statistics: 1986
Proceedings of the 18th Symposium on the Interface

$$J(\theta) = J(\delta) = \delta^T \mathbf{K}_U \delta \text{ subject to } \mathbf{T}_U^T \delta = \mathbf{0} .$$

Typically the $k \times k$ matrix $\mathbf{K}_U$ and $k \times t$ matrix $\mathbf{T}_U$ are either derived from the unique design points or from a set of user-supplied basis nodes (see Appendix 2 of Bates et al. (1985)). If we write the parameter vector as

$$\theta = \mathbf{S}\alpha + \mathbf{T}\beta + \mathbf{K}\delta$$

in which $\mathbf{S}$ is the $n \times c$ covariate matrix, $\mathbf{T}$ is the $n \times t$ polynomial matrix, and $\mathbf{K}$ is the $n \times k$ smooth matrix, the linearized problem becomes (1.1).

We can locate the unique design points $\mathbf{T}_U$, and the corresponding unique covariates $\mathbf{S}_{1U}$, and form a QR decomposition

$$[\mathbf{T}_U : \mathbf{S}_{1U}] = \tilde{\mathbf{F}}\tilde{\mathbf{G}} = \tilde{\mathbf{F}}_1\tilde{\mathbf{G}}_1 .$$

From this we construct the (unweighted) design

$$\mathbf{X} = [\mathbf{T} : \mathbf{S} : \mathbf{K}\tilde{\mathbf{F}}_2] \tag{2.2}$$

and penalty

$$\Sigma = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{F}}_2^T \mathbf{K}_U \tilde{\mathbf{F}}_2 \end{pmatrix} . \tag{2.3}$$

We decompose Σ using a pivoted Cholesky followed by a Householder,

$$\mathbf{E}^T \Sigma \mathbf{E} = \mathbf{L}^T \mathbf{L} \quad \text{and} \quad \mathbf{L}^T = \mathbf{Q}\mathbf{R} = \mathbf{Q}_1 \mathbf{R}_1 , \tag{2.4}$$

and construct

$$\mathbf{Z} = [\mathbf{Z}_1 : \mathbf{Z}_2] = \mathbf{X}\mathbf{E}\mathbf{Q} \begin{pmatrix} \mathbf{R}_1^{-T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} . \tag{2.5}$$

Finally, the original parameters are transformed to

$$\begin{pmatrix} \beta \\ \alpha \\ \tilde{\mathbf{F}}_2^T \delta \end{pmatrix} = \mathbf{E}\mathbf{Q} \begin{pmatrix} \mathbf{R}_1^{-T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \gamma \\ \omega \end{pmatrix} . \tag{2.6}$$

In the usual case that $\tilde{\mathbf{F}}_2^T \mathbf{K}_U \tilde{\mathbf{F}}_2$ is full rank, $\mathbf{E}\mathbf{Q}_2$ is an $n \times (c+t)$ matrix which permutes the coefficients α and β, i.e., $\omega^T = (\beta^T : \alpha^T : \mathbf{0})\mathbf{E}\mathbf{Q}_2$. The objective functional can now be reparameterized as

$$\frac{1}{n} \| \mathbf{W}^T(\mathbf{y} - \mathbf{Z}_2\omega - \mathbf{Z}_1\gamma) \|^2 + \lambda\gamma^T\gamma . \tag{2.7}$$

At this point, we have done all the "one-time" decompositions. The following steps must be redone each time $\mathbf{W}$ and $\mathbf{y}$ change, or simply once for the linear (normal) model. We form a QR decomposition of

$$\mathbf{W}^T\mathbf{Z}_2 = \mathbf{F}\mathbf{G} = \mathbf{F}_1\mathbf{G}_1 ,$$

and create

$$\mathbf{J} = [\mathbf{J}_1 : \mathbf{J}_2] = [\mathbf{F}_1^T : \mathbf{F}_2^T]\mathbf{W}^T\mathbf{Z}_1 ,$$

leading to the minimization of

$$\frac{1}{n} \| \mathbf{F}_1^T\mathbf{W}^T\mathbf{y} - \mathbf{G}_1\omega - \mathbf{J}_1\gamma \|^2 \tag{2.8}$$

$$+ \frac{1}{n} \| \mathbf{F}_2^T\mathbf{W}^T\mathbf{y} - \mathbf{J}_2\gamma \|^2 + \lambda\gamma^T\gamma .$$

The first term can be made zero by solving for ω, with any given γ,

$$\mathbf{G}_1\omega = \mathbf{F}_1^T\mathbf{W}^T\mathbf{y} - \mathbf{J}_1\gamma . \tag{2.9}$$

The estimate of γ is found by solving

$$\mathbf{M}\gamma = \mathbf{J}_2^T\mathbf{F}_2^T\mathbf{W}^T\mathbf{y} , \tag{2.10}$$

with

$$\mathbf{M} = \mathbf{J}_2^T\mathbf{J}_2 + n\lambda\mathbf{I} .$$

The "hat" matrix can be formally written as

$$\mathbf{A}(\lambda) = \mathbf{W}^{-T}\mathbf{F} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2\mathbf{M}^{-1}\mathbf{J}_2^T \end{pmatrix} \mathbf{F}^T\mathbf{W}^T \tag{2.11}$$

provided we can invert $\mathbf{M}$. Naturally, one would iterate to new pseudo-values and pseudo-weights using (2.1) and repeat the minimization of the objective function (2.7). At convergence, one can obtain the estimates of the original parameters via (2.6).

One may approach the above solution for γ and the "hat" matrix $\mathbf{A}(\lambda)$ in different ways, depending on whether one wishes to choose a new λ, say via generalized cross validation, or whether one wishes to leave λ fixed.

2.1. SVD approach

One way to choose a new λ is based on generalized cross validation for the linearized problem (2.7). This is basically the ridge regression problem of Golub, Heath and Wahba (1979). Form a singular value decomposition of

$$\mathbf{J}_2 = \mathbf{U}\mathbf{D}\mathbf{V}^T ,$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{D}$ is diagonal, to get

$$\hat{\gamma} = \mathbf{V}(\mathbf{D}^2 + n\lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{F}_2^T\mathbf{W}^T\mathbf{y} .$$

The "hat" matrix is

$$\mathbf{A}(\lambda) = \mathbf{W}^{-T}\mathbf{F} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}\mathbf{D}^2(\mathbf{D}^2 + n\lambda\mathbf{I})^{-1}\mathbf{U}^T \end{pmatrix} \mathbf{F}^T\mathbf{W}^T .$$

One can choose λ to minimize the GCV criterion (Craven and Wahba, 1979)

$$V(\lambda) = \frac{n \| \mathbf{W}^T(\mathbf{I} - \mathbf{A}(\lambda))\mathbf{y} \|^2}{[tr(\mathbf{I} - \mathbf{A}(\lambda))]^2} , \tag{2.12}$$

or as some intermediate value if this is seen as being too "far" from the previous value.

2.2. Cholesky approach

If we choose to leave λ fixed, one can take the cheaper approach of a Cholesky decomposition of

$$\mathbf{M} = \mathbf{J}_2^T\mathbf{J}_2 + n\lambda\mathbf{I} = \mathbf{C}^T\mathbf{C} ,$$

leading to the estimate of γ by solving

$$\mathbf{C}^T\mathbf{C}\hat{\gamma} = \mathbf{J}_2^T\mathbf{F}_2^T\mathbf{W}^T\mathbf{y} .$$

The "hat" matrix becomes

$$\mathbf{A}(\lambda) = \mathbf{W}^{-T}\mathbf{F} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2\mathbf{C}^{-1}\mathbf{C}^{-T}\mathbf{J}_2^T \end{pmatrix} \mathbf{F}^T\mathbf{W}^T . \tag{2.13}$$

This route was followed by O'Sullivan, Yandell and Raynor (1986), iterating to a solution for fixed λ. The "optimal" λ was chosen by minimizing $V(\lambda)$ over a grid of $\log(\lambda)$.

3. Iteratively Reweighted Least Squares Models

Iteratively reweighted least squares (IRLS) models differ from semi-parametric GLMs in that only the penalty matrix remains fixed (Green, 1984). The log-likelihood parameter θ can be locally linearized, but the $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{K}$ matrices are no longer

fixed:

$$\mathbf{S} = \frac{\partial L}{\partial \alpha}\ ,\ \mathbf{T} = \frac{\partial L}{\partial \beta}\ ,\ \mathbf{K} = \frac{\partial L}{\partial \delta}\ .$$

We still only need form and decompose Σ as in (2.3) and (2.4) exactly once. However, the (unweighted) design (2.2) may change with each iteration. Hence, the remaining computations need to be done at each iteration. One could proceed in the same manner as for the generalized linear models, but reconstructing **X**, and hence **Z** and **J**, each time.

4. General Nonlinear Models

General nonlinear problems could proceed in the same manner as for IRLS, except that $\mathbf{K}_U$ changes each time. Thus most computations need to be redone. It may be possible for some nonlinear problems to reparameterize them as SGLM or IRLS problems to eliminate this difficulty.

5. Diagnostics

The diagonal elements of the "hat" matrix have been used for diagnostics in generalized linear models (Pregibon, 1981) as well as in smoothing spline models (Eubank 1984, 1985). It is natural to think of extending these uses to the present array of models (Green and Yandell, 1985; Green, 1985). The diagonal elements can be computed as

$$\{\mathbf{A}(\lambda)\}_{ii} = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{e}_i \|^2 + \| \mathbf{M}^{-\frac{1}{2}}\mathbf{F}_2^{\mathrm{T}}\mathbf{e}_i \|^2$$

in which $\mathbf{e}_i$ is the n-vector with a 1 in the i-*th* position and 0's elsewhere. For the SVD approach this is simply

$$\{\mathbf{A}(\lambda)\}_{ii} = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{e}_i \|^2 + \| \mathbf{D}(\mathbf{D}+n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{e}_i \|^2\ ,$$

and for the Cholesky approach (*cf.* O'Sullivan (1985)),

$$\{\mathbf{A}(\lambda)\}_{ii} = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{e}_i \|^2 + \| \mathbf{C}^{-\mathrm{T}}\mathbf{J}_2^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{e}_i \|^2\ .$$

Covariance matrices can be computed by noting that $COV(\mathbf{y}) = \mathbf{W}^{-\mathrm{T}}\mathbf{W}^{-1}$. We find from (2.11) that

$$COV(\hat{\theta}) = \mathbf{W}^{-\mathrm{T}}\mathbf{F}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_2\mathbf{M}^{-1}\mathbf{J}_2^{\mathrm{T}}\mathbf{J}_2\mathbf{M}^{-\mathrm{T}}\mathbf{J}_2^{\mathrm{T}}\end{bmatrix}\mathbf{F}^{\mathrm{T}}\mathbf{W}^{-\mathrm{T}}\ .$$

Hence, the variances are

$$VAR(\theta_i) = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2 + \| \mathbf{J}_2\mathbf{M}^{-\mathrm{T}}\mathbf{J}_2^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2\ .$$

Noting the relation

$$\mathbf{M}^{-1}\mathbf{J}_2^{\mathrm{T}}\mathbf{J}_2\mathbf{M}^{-\mathrm{T}} = \mathbf{M}^{-1}(\mathbf{I} - n\lambda\mathbf{M}^{-1})\ ,$$

the variances can be written as

$$VAR(\theta_i) = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2 + \| \mathbf{C}^{-\mathrm{T}}\mathbf{J}_2^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2 - n\lambda \| \mathbf{C}^{-1}\mathbf{C}^{-\mathrm{T}}\mathbf{J}_2^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2$$

for the Cholesky approach. For the SVD we have

$$VAR(\theta_i) = \| \mathbf{F}_1^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2 + \| \mathbf{D}^2(\mathbf{D}^2+n\lambda\mathbf{I})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{F}_2^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{e}_i \|^2\ .$$

The covariance among the coefficients can be derived, using (2.9), (2.10) and (2.6), as

$$COV\begin{bmatrix}\beta\\ \alpha\\ \tilde{\mathbf{F}}_2^{\mathrm{T}}\delta\end{bmatrix} = \mathbf{E}\mathbf{Q}_2\mathbf{G}_1^{-1}\mathbf{G}_1^{-\mathrm{T}}\mathbf{Q}_2^{\mathrm{T}}\mathbf{E}^{\mathrm{T}} +$$

$$\mathbf{E}\mathbf{Q}\begin{bmatrix}\mathbf{R}_1^{-\mathrm{T}}\\ -\mathbf{G}_1^{-1}\mathbf{F}_1^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\end{bmatrix}\mathbf{M}^{-1}\mathbf{J}_2^{\mathrm{T}}\mathbf{J}_2\mathbf{M}^{-\mathrm{T}}\begin{bmatrix}\mathbf{R}_1^{-\mathrm{T}}\\ -\mathbf{G}_1^{-1}\mathbf{F}_1^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\ .$$

In many situations we may be only interested in $COV(\alpha)$. Further, if the penalty Σ is of the proper rank, then the QR decomposition of (2.4) should simply permute the indices for the coefficients. In other words, $\mathbf{EQ}_2$ often simply permutes the coefficients α (and β) into ω. In this case, let $\tilde{e}_i$ denote the permutation for α_i, $i = 1, \cdots, c$. For the SVD approach,

$$VAR(\alpha_i) = \| \mathbf{G}_1^{-\mathrm{T}}\tilde{e}_i \|^2 + \| \mathbf{D}(\mathbf{D}^2+n\lambda\mathbf{I})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{W}\mathbf{F}_1\mathbf{G}_1^{-\mathrm{T}}\tilde{e}_i \|^2\ .$$

For the Cholesky approach,

$$VAR(\alpha_i) = \| \mathbf{G}_1^{-\mathrm{T}}\tilde{e}_i \|^2 + \| \mathbf{C}^{-\mathrm{T}}\mathbf{W}\mathbf{F}_1\mathbf{G}_1^{-\mathrm{T}}\tilde{e}_i \|^2 - n\lambda \| \mathbf{C}^{-1}\mathbf{C}^{-\mathrm{T}}\mathbf{W}\mathbf{F}_1\mathbf{G}_1^{-\mathrm{T}}\tilde{e}_i \|^2\ .$$

Joint work is in progress with Peter J. Green (Green and Yandell, 1985) on analogues to diagnostic tools for generalized linear models along the lines of Pregibon (1981, 1982) and Nelder and Pregibon (1986).

6. Numerical Comparisons

We focus our investigations upon the Poisson and binomial special cases of the semi-parametric generalized linear model as these are potentially of wide interest and easy to formulate. We allowed up to c initial iterations of the Cholesky decomposition (CD) for $\lambda = \infty$ (perfectly smooth case), and up to c CDs following each SVD, where c was 1, 2, or 10. No case required more than 7 CD following an SVD, or more than 7 SVD overall.

We examined some real data on leafhopper oviposition and potato pathogen in a field, both Poisson, and data on rat survival, which was binomial. In addition we simulated data which we thought might be "cumbersome" for the numerical algorithms. The simulations were Poisson with a normal shaped curve of θ = log(mean value), with peak height of between $\theta = 1.5$ and 20. Binomial simulations used a similar normal shaped curve for θ = logit(mean value), with peak height of between θ = logit(.01) and logit(.3). Simulations were conducted for $n = 50$ and 100.

The Cholesky steps in the real examples increased the run time by 20-35%, including one-time costs and construction of the diagonals of the "hat" matrix (see Tables 1-3). This occurred because the number of SVDs was not reduced by more intermediate CDs, nor were the sequences of optimal λ's for the linearized problems markedly altered by the CDs. In addition, each CD took about 10% of the time for an SVD. In these examples, the signal was fairly apparent, indicating that the linear approximation was adequate using the SVD iterations alone.

Table 1. Poisson Oviposition Data (n=27)

task	c=0	c=1	c=2	c=10
one-time	4.40	4.40	4.43	4.50
cholesky	0.78	4.22	7.78	11.92
svd	24.93	25.02	24.73	24.78
hat	2.20	2.22	2.23	2.22
total	31.07	34.57	37.85	42.10
no. svd	5	5	5	5
no. chol	1	6	11	19

Table 2. Binomial Rats Data (n=127)

task	c=0	c=2	c=10
one-time	34.6	33.9	35.0
cholesky	7.5	58.7	74.2
svd	245.0	245.6	243.9
hat	34.6	34.8	35.2
total	312.8	364.2	379.3
no. svd	5	5	5
no. chol	1	9	12

Table 3. 2-D Poisson Fungi (n=400, k=100)			
task	c=0	c=2	c=10
one-time	279	279	283
cholesky	140	1004	1475
svd	4486	4413	4425
hat	594	598	598
total	5354	6150	6637
no. svd	7	7	7
no. chol	2	16	26

The simulations showed that when the "signal" is small relative to the "noise", the CDs seem to stabilize the minimization problem, reducing the number of SVDs required and cutting the run time. Table 4(a-b) present the combined CD and SVD run times, while Table 4(c-d) present the numbers of SVDs and CDs. As the height of the Poisson peak rises, the CD iterations have a reduced impact on convergence. However, note that on several occassions iteration with only one CD increased the number of SVDs required. Allowing more than 2 CD steps only seemed to increase the overall run time; the number of SVDs was reduced in only a few instances. In addition, a few simulations, not shown here, converged when up to 2 CDs per SVD were allowed, but did not converge when 0 or up to 10 were allowed. Similar statements can be made about the binomial simulations (Table 5(a-b)).

Table 4(a). Poisson Run Times (n=50)				
peak	c=0	c=1	c=2	c=10
1.5	134	120	94	103
2	163	150	130	141
2.5	134	148	126	134
3	132	148	125	138
4	159	178	155	142
5	158	180	157	144
6	131	173	155	120
7	133	159	127	161
8	131	175	157	141
9	135	178	158	144
10	157	204	188	174
15	134	180	187	181
20	158	207	189	175

Table 4(b). Poisson Run Times (n=100)				
peak	c=0	c=1	c=2	c=10
1.5	974	848	885	904
2	950	834	880	933
2.5	1149	1051	1098	932
3	759	824	659	718
4	956	1048	882	967
5	955	1069	1100	988
6	970	1244	915	1006
7	938	1038	873	970
8	939	1053	1105	1043
9	955	1280	1138	1026
10	1129	1245	1106	1371
15	941	1252	1109	762
20	962	1276	1131	1143

Table 4(c). Poisson Runs (n=50)				
	no. SVD / no. CD iterations			
peak	c=0	c=1	c=2	c=10
1.5	5/0	4/4	3/5	3/10
2	6/1	5/6	4/8	4/12
2.5	5/0	5/5	4/7	4/12
3	5/0	5/5	4/7	4/13
4	6/0	6/6	5/8	4/15
5	6/0	6/7	5/9	4/15
6	5/0	6/6	5/9	3/16
7	5/0	5/5	4/7	4/19
8	5/0	6/6	5/9	4/14
9	5/1	6/6	5/9	4/15
10	6/0	7/7	6/10	5/16
15	5/1	6/7	6/10	5/18
20	6/0	7/7	6/10	5/16

Table 4(d). Poisson Runs (n=100)				
	no. SVD / no. CD iterations			
peak	c=0	c=1	c=2	c=10
1.5	5/0	4/4	4/6	4/10
2	5/0	4/4	4/7	4/12
2.5	6/0	5/5	5/8	4/11
3	4/0	4/4	3/6	3/11
4	5/0	5/5	4/7	4/13
5	5/0	5/6	5/8	4/14
6	5/1	6/6	4/9	4/15
7	5/0	5/5	4/7	4/14
8	5/0	5/6	5/9	4/17
9	5/0	6/7	5/9	4/16
10	6/0	6/6	5/9	5/23
15	5/0	6/6	5/9	3/13
20	5/0	6/6	5/9	4/19

Table 5(a). Binomial Run Times (n=100)					
size	prob	c=0	c=1	c=2	c=10
10	.3	108	87	90	91
	.2	106	118	125	131
	.1	133	118	92	97
	.05	135	148	130	135
20	.3	109	91	92	96
	.2	137	119	123	127
	.1	109	120	124	129
	.05	165	151	159	168

Table 5(b). Binomial Run Times (n=100)					
size	prob	c=0	c=1	c=2	c=10
10	.3	943	827	671	692
	.2	970	829	858	882
	.1	968	860	885	937
	.05	1171	1064	898	977
	.01	1166	1046	1097	935
20	.3	743	604	632	635
	.2	760	617	636	645
	.1	780	838	650	680
	.05	795	849	681	742
	.01	1351	1261	1103	1225
	.005	1513	1676	1536	1683

Table 5(c). Binomial Runs (n=50)					
		no. SVD / no. CD iterations			
size	prob	c=0	c=1	c=2	c=10
10	.3	5/0	4/4	3/5	3/8
	.2	4/0	4/4	4/6	4/9
	.1	4/0	3/3	3/5	3/6
	.05	5/0	5/5	4/8	4/11
20	.3	4/0	4/4	4/6	4/9
	.2	5/1	4/4	4/6	4/8
	.1	4/1	3/4	3/5	3/7
	.05	6/0	5/5	5/8	5/12

Table 5(d). Binomial Runs (n=100)					
		no. SVD / no. CD iterations			
size	prob	c=0	c=1	c=2	c=10
10	.3	5/0	4/4	3/6	3/8
	.2	5/0	4/4	4/6	4/8
	.1	5/0	4/5	4/7	4/11
	.05	6/0	5/5	4/7	4/12
	.01	6/0	5/5	5/8	4/11
20	.3	4/0	3/3	3/5	3/6
	.2	4/1	3/3	3/5	3/7
	.1	4/1	4/4	3/5	3/8
	.05	4/0	4/4	3/6	3/10
	.01	7/0	6/6	5/8	5/14
	.005	8/0	8/8	7/11	7/20

Since we know that the estimates converge for fixed λ (O'Sullivan, Yandell and Raynor, Jr., 1986), a few iterations for fixed λ may guard against nonlinearity in the penalized likelihood. It is not known at this time what conditions are required on the penalized likelihood, as a function of λ, to insure convergence in the SVD-only approach.

If one follows Elden (1984) to stop the singular value decomposition after the bidiagonalization, considerable time can be saved since the effort to diagonalize is magnified by the number of iterations. Earlier work on GCVPACK (Bates et al., 1985) indicated that half of the singular value decomposition time may be spent on bidiagonalization. Of course, once convergence is reached, one could complete the diagonalization, doing this only once, to easily derive the diagonal of the "hat" matrix. Such a savings in computation would further reduce the advantage of iterating via Cholesky with fixed λ.

Acknowledgements

This research has been supported in part by United States Department of Agriculture CSRS grant 511-100, and National Sciences Foundation grant DMS-8404970. Computing was performed on the UW-Madison Statistics VAX 11/750 Research Computer.

References

Bates, D. M., Lindstrom, M. J., Wahba, G. and Yandell, B. S. (1985) GCVPACK – Routines for Generalized Cross Validation. Technical Report#775, Dept. of Statistics, U. of Wisconsin.

Cox, D. D. (1983) Gaussian approximation of smoothing splines. Technical Report#743, Dept. of Statistics, U. of Wisconsin.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377-403.

Elden, L. (1984) A note on the computation of the generalized cross-validation function for ill-conditioned least squares problems. *BIT*, **24**, 467-472.

Eubank, R. L. (1984) The hat matrix for smoothing splines. *Statist. and Prob. Letters*, **2**, 9-14.

Eubank, R. L. (1985) Diagnostics for smoothing splines. *J. Roy. Statist. Soc. Ser. B*, **47**. (to appear)

Golub, G. H., Heath, M. and Wahba, G. (1979) Generalised cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215-224.

Good, I. J. and Gaskins, R. A. (1971) Non-parametric roughness penalties for probability densities. *Biometrika*, **58**, 255-277.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J. Roy. Statist. Soc. Ser. B*, **46**, 149-192.

Green, P. J. (1985) Penalized likelihood for general semi-parametric regression models. Technical Report#2819, Math. Research Center, U. of Wisconsin.

Green, P. J., Jennison, C. and Seheult, A. H. (1983) Contribution to the discussion of the paper by Wilkinson et al. *J. Roy. Statist. Soc. Ser. B*, **45**, 193-195.

Green, P. J. and Yandell, B. S. (1985) Semi-parametric generalized linear models. In *GLIM85: Proceedings of the International Conference on Generalized Linear Models, September 1985* (R. Gilchrist, ed.) Lecture Notes in Statistics, Springer-Verlag. (Technical Report#2847, Math. Res. Cen., U. of Wisconsin)

Leonard, T. (1982) An empirical Bayesian approach to the smooth estimation of unknown functions. Technical Report#2339, Math. Research Center, U. of Wisconsin.

Nelder, J. A. and Pregibon, D. (1986) An extended quasi-likelihood function. Unpublished manuscript.

O'Sullivan, F. (1983) The analysis of some penalized likelihood schemes. Technical Report#726, Dept. of Statistics, U. of Wisconsin.

O'Sullivan, F. (1985) Contribution to the discussion of the paper by Silverman. *J. Roy. Statist. Soc. Ser. B*, **47**, 39-40.

O'Sullivan, F., Yandell, B. S. and Raynor, Jr., W. J. (1986) Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, **81**, 000-000. (Technical Report#734, Statistics Dept., U. of Wisconsin)

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705-724.

Pregibon, D. (1982) Score tests in GLIM. In *Proc. GLIM82 Conf.* (R. Gilchrist, ed.) New York: Springer-Verlag.

Speckman, P. (1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13**, 970-983.