Citation for published version:
Wood, SN 2000, 'Modelling and smoothing parameter estimation with multiple quadratic penalties', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 62, no. 2, pp. 413-428. https://doi.org/10.1111/1467-9868.00240

DOI:
10.1111/1467-9868.00240

Publication date:
2000

Document Version
Early version, also known as pre-print

Link to publication

This is the pre-peer reviewed version of the following article: Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 62(2), 413-428, which has been published in final form at http://dx.doi.org/10.1111/1467-9868.00240

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 22. Aug. 2022

J. R. Statist. Soc. B (2000)
62, *Part 2, pp.413-428*

Modelling and smoothing parameter estimation with multiple quadratic penalties

S.N.Wood†

University of St. Andrews, U.K.

Summary. Penalized likelihood methods provide a range of practical modelling tools, including spline smoothing, generalized additive models and variants of ridge regression. Selecting the correct weights for penalties is a critical part of using these methods and in the single penalty case the analyst has several well founded techniques to choose from. However, many modelling problems suggest a formulation employing multiple penalties, and here general methodology is lacking. A wide family of models with multiple penalties can be fitted to data by iterative solution of the generalized ridge regression problem: $minimise\ \|\mathbf{W}^{1/2}(\mathbf{Xp}-\mathbf{y})\|^2\rho + \sum_{i=1}^{m}\theta_i\mathbf{p}'\mathbf{S}_i\mathbf{p}$ ($\mathbf{p}$ is a parameter vector, $\mathbf{X}$ a design matrix, $\mathbf{S}_i$ a non-negative definite coefficient matrix defining the i^{th} penalty with associated smoothing parameter θ_i, $\mathbf{W}$ a diagonal weight matrix, $\mathbf{y}$ a vector of data or pseudodata and ρ an 'overall' smoothing parameter included for reasons of computational efficiency). This paper shows how smoothing parameter selection can be performed *efficiently* by applying generalized cross validation to this problem and how this allows non-linear, generalized linear and linear models to be fitted using multiple penalties, substantially increasing the scope of penalized modelling methods. Examples of non-linear modelling, generalized additive modelling and anisotropic smoothing are given.

Keywords: Generalized additive models; Generalized cross-validation; Generalized ridge regression; Model selection; Multiple smoothing Parameters; Non-linear modelling; Penalized likelihood; Penalized regression splines

1. Introduction

Penalized likelihood methods provide important modelling tools in applied statistics, with the books by Green and Silverman (1994), Hastie and Tibshirani (1990), and Wahba (1990) providing surveys of various approaches. A well known example of a penalized regression problem is the fitting of a cubic smoothing spline to data (y_i, x_i) by minimisation of:

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda\int[s''(x)]^2dx.$$

The spline s has one parameter per design point, x_i, but the family of models that can be obtained by solving the fitting problem ranges from a straight line ($\lambda \to \infty$) to an interpolating cubic spline ($\lambda \to 0$): λ controls which of a family of models of differing complexity is actually fitted. Hence the crucial step in the practical use of smoothing splines is the selection of the smoothing parameter, λ,

†*Address for correspondence:* Simon Wood, The Mathematical Institute, North Haugh, St. Andrews, Fife KY16 9SS U.K.
E-mail: snw@st-and.ac.uk

controlling the trade-off between fidelity to the data and smoothness of the fitted spline. Generalized cross validation (GCV, Craven and Wahba, 1979) is one method of smoothing parameter selection that has proven effective and has good theoretical properties, although there are other similar methods (AIC/unbiased risk estimation, and generalized maximum likelihood, see Wahba 1990). The availability of highly efficient implementations of GCV based smoothing parameter selection (e.g. Hutchinson and deHoog, 1985) is in large part responsible for the popularity of smoothing splines.

Spline models have been generalized. 'Thin plate' splines provide a way of modelling a single response to multiple covariates; data may be observations of linear functionals of the spline; models that are a mixture of parametric and spline terms can be produced, and all these extensions can be employed in a generalized linear modelling framework. Green and Silverman (1994) or Wahba (1990) provide more or less accessible accounts of these developments, again using GCV for smoothing parameter selection .

An obvious further generalization is to allow multiple smooth terms in a model, so that model fitting must balance goodness of fit against multiple penalties. Examples are provided by Wahba (1990), Wahba *et al.* (1995), Gu and Wahba (1993) and Hastie and Tibshirani (1990). With this generalization automatic smoothing parameter selection becomes much more difficult computationally, but also more important: with a single penalty it might be argued that an acceptable level of smoothing can be obtained by informal methods (examination of residual autocorrelation, for example), but this becomes difficult, time-consuming and ever less objective as the number of smoothing parameters to be chosen increases.

In principle, as in the single penalty case, multiple smoothing parameters can be chosen automatically by GCV, but to date the only practical method for doing this (Gu and Wahba, 1991) applies to reproducing kernel based spline models and not to other models with multiple penalties. Its operation count is of cubic order in the number of data to be fitted. In the context of generalized additive models (Hastie and Tibshirani, 1990) various approaches to multiple smoothing parameter selection have been taken which again use the special structure of the problem and do not generalize easily. The most defensible of these is based on iteratively estimating one smoothing parameter at a time, by minimising an approximation to the GCV objective function: but this approach is only effective if there is little covariance between the covariates with which the different penalties are associated and if the approximate objective can be minimised efficiently. In practice, efficient minimisation in this context requires that models be constructed solely from functions of single covariates.

While existing modelling approaches might be improved by the existence of a good multiple smoothing parameter selection method, substantial benefits should follow from a widening of the range of models that can be employed with multiple penalties. For example, there are many cases in which a non-linear model with multiple penalties is an obvious formulation, but an impractical one in the absence of a method of smoothing parameter selection. Section 5 provides one example. Other problems naturally produce linear or generalized linear models subject to multiple penalties (section 7 provides a case in point), but again lack of practical smoothing parameter selection methodology creates substantial practical difficulties.

By providing a fairly general multiple smoothing parameter selection method this paper aims to widen the range of models that can be employed with multiple penalties. Specifically, an efficient method is provided for models in which data, y_i, are treated as observations of random variables, Y_i, from some exponential family distribution and $E(Y_i) = f(\mathbf{p})$, where f is in general a non-linear function of a parameter vector $\mathbf{p}$. Estimates of the model parameters are found by minimisation of:

$$-l(\mathbf{p}, \mathbf{y})\rho + \frac{1}{2}\sum_{i=1}^{m} \theta_i \mathbf{p}'\mathbf{S}_i\mathbf{p} \tag{1}$$

(possibly subject to general linear equality constraints $\mathbf{Cp} = \mathbf{0}$) where l is the log likelihood of $\mathbf{p}$ given $\mathbf{y}$ and the matrices $\mathbf{S}_i$ contain coefficients defining the penalties applied to the parameter vector. $\boldsymbol{\theta}$ and ρ control the level of penalization. Attention will be restricted to the case in which the matrices $\mathbf{S}_i$ are non-negative definite (at least when projected into the null space of any constraints on the problem) and f is sufficiently well behaved for the problem to be solved by iterative least squares. Hence the methods are also applicable to penalized least squares problems which are not interpretable in terms of likelihood.

The remainder of the paper is structured as follows: firstly methods for fitting penalized models are reviewed to obtain a fitting framework helpful to smoothing parameter selection; then a GCV method is developed for multiple smoothing parameter selection in generalized ridge regression problems by building on Gu and Wahba's (1991) approach. To illustrate the fairly wide ranging utility of the results three brief examples are presented: non-linear modelling with multiple penalties, generalized additive modelling with penalized regression splines and anisotropic smoothing. The computer code implementing the method of section 4 can be obtained from:
`http://www.blackwellpublishers.co.uk/rss/`
or from
`http://www.ruwpa.st-and.ac.uk/simon/mgcv.html`.

2. Model fitting

It can be shown that a wide range of models with multiple penalties can be fitted to data by (possibly) iterative solution of the generalized ridge regression problem:

$$minimise \;\; \|\mathbf{W}^{1/2}(\mathbf{Xp}-\mathbf{y})\|^2\rho + \sum_{i=1}^{m}\theta_i\mathbf{p}'\mathbf{S}_i\mathbf{p} \tag{2}$$

subject to the linear constraints:

$$\mathbf{Cp} = \mathbf{0}. \tag{3}$$

$\|.\|$ is the euclidean norm, $\mathbf{X}$ is an $n \times q$ design matrix, $\mathbf{p}$ the model parameter vector, $\mathbf{y}$ a vector of (pseudo)data to be fitted, and $\mathbf{W}$ a vector of weights (which may be iteratively re-calculated in practice). The $\mathbf{S}_i$'s are non-negative definite coefficient matrices defining the penalties, each with associated smoothing parameter θ_i. ρ is an overall smoothing parameter introduced for reasons of computational efficiency that will become clear later. $\mathbf{C}$ is a matrix of known coefficients defining the constraints.

The fitting objective is self evidently appropriate for generalized ridge regression problems, but slightly less obvious in other contexts. First, consider fitting a generalized linear model by penalized likelihood maximisation. Specifically, assume that the y_i's are observations of independent random variables from some exponential family distribution and that $\mu_i \equiv E\{Y_i\}$ is related to the model parameters by $g(\mu_i) = \mathbf{X}_i\mathbf{p}$, where $\mathbf{X}_i$ is the i^{th} row of a design matrix and $g()$ a monotonic 'link' function. Further suppose that m quadratic penalties are to be applied to the parameter vector. The model parameters are estimated by maximising the penalized log-likelihood:

$$l_p = \rho\sum_{i=1}^{n} l_i(y_i, \mathbf{X}_i\mathbf{p}) - \frac{1}{2}\sum_{i=1}^{m}\theta_i\mathbf{p}'\mathbf{S}_i\mathbf{p}.$$

$\mathbf{p}^{[k]}$ (the k^{th} estimate of $\mathbf{p}$) can be updated by Fisher scoring:

$$\mathbf{p}^{[k+1]} = \mathbf{p}^{[k]} + \left[E\left\{ -\frac{\partial^2 l_p}{\partial \mathbf{p}^{[k]} \partial \mathbf{p}^{[k]\prime}} \right\} \right]^{-1} \frac{\partial l_p}{\partial \mathbf{p}^{[k]}}$$

(where the notation $\partial l/\partial \mathbf{p}$, refers to the vector whose i^{th} element is $\partial l/\partial p_i$). Define $\mathbf{W}^{[k]}$ to be the diagonal matrix such that $W_{ii}^{[k]} = [g'(\mu_i^{[k]})^2 V_i^{[k]}]^{-1}$ where $V_i^{[k]}$ is the variance of Y_i according to the estimates, $\mu_i^{[k]}$, implied by $\mathbf{p}^{[k]}$, and let $\mathbf{\Gamma}^{[k]}$ be the diagonal matrix such that $\Gamma_{ii}^{[k]} = 1/g'(\mu_i^{[k]})$. The Fisher scoring update equations become:

$$\mathbf{p}^{[k+1]} = \mathbf{p}^{[k]} + \left[\mathbf{X}'\mathbf{W}^{[k]}\mathbf{X}\rho + \sum \theta_i \mathbf{S}_i \right]^{-1} \left[\mathbf{X}'\mathbf{W}^{[k]}\mathbf{\Gamma}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]}) - \sum \theta_i \mathbf{S}_i \mathbf{p}^{[k]} \right].$$

This is equivalent to finding $\mathbf{p}^{[k+1]}$ by solving the weighted penalized least squares problem:

$$minimise \;\; \rho \|\mathbf{W}^{1/2}(\mathbf{z}^{[k]} - \mathbf{X}\mathbf{p})\|^2 + \sum \theta_i \mathbf{p}'\mathbf{S}_i \mathbf{p}$$

where $\mathbf{z}^{[k]} = \mathbf{X}\mathbf{p}^{[k]} + \mathbf{\Gamma}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$ (and I've dropped the superscripts on $\mathbf{W}$ for no good reason). This kind of approach can be found in O'Sullivan *et al.* (1986), or Green and Silverman (1994).

Constraints are incorporated into such a scheme (Gill *et al.*, 1981) by finding a column basis, $\mathbf{Z}$, for the null space of the constraint matrix $\mathbf{C}$ (QR or QT factorisation will provide this). Solution proceeds as outlined above, but with $\mathbf{X}$ replaced by $\mathbf{XZ}$ and $\mathbf{S}_i$ replaced by $\mathbf{Z}'\mathbf{S}_i\mathbf{Z}$: the working parameter vector will be $\mathbf{p}_z$ where $\mathbf{p} = \mathbf{Z}\mathbf{p}_z$. Constraints are often required in the event of parameter aliasing (i.e. column rank deficiency in $\mathbf{X}$) when it is necessary to impose equality constraints on some of the parameters.

In the more general non-linear case it is often possible to maximise the penalized likelihood using a slightly modified Gauss-Newton method which results in an iterative least squares scheme similar to the one employed for generalized linear models. The Jacobian $\mathbf{J}^{[k]}$ takes the place of the design matrix $\mathbf{X}$, where $J_{ij}^{[k]} = \partial \mu_i^{[k]}/\partial p_j^{[k]}$. The 'pseudodata' at the k^{th} iteration is now given by: $\mathbf{z}^{[k]} = \mathbf{y} - \boldsymbol{\mu}^{[k]} + \mathbf{J}^{[k]}\mathbf{p}^{[k]}$ (assuming an identity link). Constraints are applied in the same manner as described above.

In all cases it is possible to impose inequality constraints on the model by treating the generalized ridge regression problem as a quadratic programming problem (see e.g. Gill *et al.*, 1981).

So, the generalized ridge regression problem (2, 3) is not only the basis for fitting penalized linear models, but also allows GLMs and non linear models to be fitted by penalized likelihood maximisation.

3. Multiple smoothing parameter selection

The methods of the last section are of little practical use unless the smoothing parameters, θ_i/ρ, can also be estimated. In principle this can be achieved for the problem (2) using GCV by minimising:

$$V = \frac{\|\mathbf{W}^{1/2}\{\mathbf{y} - \mathbf{A}(\rho, \boldsymbol{\theta})\mathbf{y}\}\|^2/n}{[1 - tr\{\mathbf{A}(\rho, \boldsymbol{\theta})\}/n]^2} \tag{4}$$

with respect to the θ_i/ρ's. $\mathbf{A}(\rho, \boldsymbol{\theta})$ is the influence (hat) matrix for the model: i.e. $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$. When solving (2,3) iteratively, as part of a penalized likelihood maximisation, the smoothing parameters can be estimated at each iteration using the pseudodata and iterative weights in place of $\mathbf{y}$ and $\mathbf{W}$.

This is the standard approach taken by Wahba *et al.* (1995) or Gu and Wahba (1993), for example, although O'Sullivan *et al.* (1986) adopt a slightly different method.

GCV is appealing both from a theoretical perspective (e.g. Craven and Wahba, 1979, Utreras 1981) and because of its successful use in a range of practical applications, although Generalized Maximum Likelihood (GML), Akaike's Information Criterion (AIC) or equivalent Unbiased Risk Estimators (UBRE) are well founded alternatives (see, e.g. Wahba 1990 and Akaike 1973). The practical difficulty in using GCV lies in minimisation of (4), which has the potential to be prohibitively expensive, numerically. The troublesome term is $tr(\mathbf{A})$, which would take $O(nq^2)$ operations to evaluate directly, for each trial set of smoothing parameters (q is the dimension of $\mathbf{p}$; AIC/UBRE and GML have similar problems). A simple grid search for m smoothing parameters using k gridpoints per parameter would require of order nq^2k^m floating point operations: an effective barrier to use of general models with multiple penalties, particularly if model fitting must be repeated as part of a computer intensive analysis. So, in order to allow practical use of models with multiple penalties, the next section reports an efficient method for multiple smoothing parameter selection that requires only a few times q^3 operations.

4. Minimisation of the GCV score

This section builds on Gu and Wahba's (1991) method for reproducing kernel models to produce a smoothing parameter selection method that applies to all model fitting problems of the form (2,3). The approach is to alternate highly efficient, one dimensional, direct searches for ρ with Newton updates of $log(\boldsymbol{\theta})$, so that the 'overall' level of smoothing is optimized at each step while the relative weight given to each penalty is adjusted iteratively.

The problem (2,3) has the following influence matrix:

$$\mathbf{A} = \rho\mathbf{XZ}\left(\mathbf{Z}'\mathbf{X}'\mathbf{WXZ}\rho + \sum_{i=1}^{m}\theta_i\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}'\mathbf{W}$$

(where $\mathbf{Z}$ is a column basis for the null space of $\mathbf{C}$). $\mathbf{A}$ is not in a convenient form, so let $\mathbf{Q}'\tilde{\mathbf{R}} = \mathbf{W}^{1/2}\mathbf{XZ}$, where $\mathbf{Q}$ is an orthogonal matrix made up of Householder transformations, $\tilde{\mathbf{R}}' = (\mathbf{R}', \mathbf{0})$ and $\mathbf{R}$ is full rank (q) upper triangular with inverse $\mathbf{L}$ (note that $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{L}$ need be obtained only once, before iteration). Defining $\tilde{\mathbf{A}} = \mathbf{W}^{1/2}\mathbf{A}\mathbf{W}^{-1/2}$ gives:

$$\tilde{\mathbf{A}} = \rho\mathbf{Q}'\left(\begin{array}{c}\mathbf{I}\\ \mathbf{0}\end{array}\right)\left(\mathbf{I}\rho + \sum_{i=1}^{m}\theta_i\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{ZL}\right)^{-1}(\mathbf{I},\mathbf{0})\mathbf{Q},$$

$\mathbf{I}$ being the rank q identity matrix. Now form the decomposition:

$$\sum_{i=1}^{m}\theta_i\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{ZL} = \mathbf{UTU}'$$

where $\mathbf{U}$ is orthogonal (again a product of Householder transformations) and $\mathbf{T}$ is tridiagonal. Given the $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{ZL}$'s, which are formed before iteration begins, this step is $O(q^3)$. It is clear that:

$$\tilde{\mathbf{A}} = \rho\mathbf{Q}'\left(\begin{array}{c}\mathbf{I}\\ \mathbf{0}\end{array}\right)\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'(\mathbf{I},\mathbf{0})\mathbf{Q}.$$

So $tr(\mathbf{A}) = tr(\tilde{\mathbf{A}}) = \rho tr(\mathbf{I}\rho + \mathbf{T})^{-1}$, which can be evaluated in $O(q)$ operations by Elden's method (Elden, 1984, or see Wahba, 1990, p.139). To get the GCV score we also need the residual sum of squares. By defining $\mathbf{z} = \mathbf{U}'[\mathbf{I}, \mathbf{0}]\mathbf{Q}\mathbf{W}^{1/2}\mathbf{y}$ and $\mathbf{x}$ to be the vector containing the last $n-q$ rows of $\mathbf{Q}\mathbf{W}^{1/2}\mathbf{y}$, it is straightforward to show that:

$$\|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{A}\mathbf{y})\|^2 = \|\{\mathbf{I} - \rho(\mathbf{I}\rho + \mathbf{T})^{-1}\}\mathbf{z}\|^2 + \|\mathbf{x}\|^2.$$

So the GCV score for the problem is:

$$V = \frac{\|\{\mathbf{I} - \rho(\mathbf{I}\rho + \mathbf{T})^{-1}\}\mathbf{z}\|^2/n + \|\mathbf{x}\|^2/n}{\{1 - \rho tr(\mathbf{I}\rho + \mathbf{T})^{-1}/n\}^2}.$$

Of course, any product $\mathbf{B} = (\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{D}$ can be obtained efficiently by solving $(\mathbf{I}\rho + \mathbf{T})\mathbf{B} = \mathbf{D}$, after obtaining the bidiagonal Choleski factors of $(\mathbf{I}\rho + \mathbf{T})$ in $O(q)$ operations. Hence, once $\mathbf{T}$ has been obtained (and $\|\mathbf{x}\|^2$ evaluated), V can by evaluated in $O(q)$ operations for any ρ, allowing the optimal ρ to be found economically by direct search. Note that the estimated parameter vector is given by:

$$\hat{\mathbf{p}} = \rho\mathbf{Z}\mathbf{L}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{z}.$$

Updating the parameters θ_i is less straightforward. I will again follow Gu and Wahba (1991) and update the variables $\eta_i = \log\theta_i$ by Newton's method (see Gill *et al.* 1981). This requires the derivatives of $\tilde{\mathbf{A}}$ w.r.t. $\boldsymbol{\eta}$. First define $\mathbf{G} = \mathbf{U}(\mathbf{I}\rho + \mathbf{T})\mathbf{U}'$ $(= \mathbf{I}\rho + \sum_{i=1}^m \theta_i\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L})$, so that $\tilde{\mathbf{A}} = \rho\mathbf{Q}'(\mathbf{I}, \mathbf{0})'\mathbf{G}^{-1}(\mathbf{I}, \mathbf{0})\mathbf{Q}$. Clearly,

$$\frac{\partial\mathbf{G}}{\partial\eta_i} = \theta_i\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L} \quad \text{and} \quad \frac{\partial^2\mathbf{G}}{\partial\eta_i\partial\eta_j} = \left\{ \begin{array}{ll} 0 & i \ne j \\ \theta_i\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L} & i = j \end{array} \right. .$$

So, using $\partial\mathbf{G}^{-1}/\partial\eta_i = -\mathbf{G}^{-1}\partial\mathbf{G}/\partial\eta_i\mathbf{G}^{-1}$ gives,

$$\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_i} = -\rho\theta_i\mathbf{Q}'\begin{pmatrix}\mathbf{I}\\ \mathbf{0}\end{pmatrix}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'(\mathbf{I}, \mathbf{0})\mathbf{Q}$$

and

$$\frac{\partial^2\tilde{\mathbf{A}}}{\partial\eta_i\partial\eta_j} = \rho\theta_i\theta_j \times$$

$$\left[\mathbf{Q}'\begin{pmatrix}\mathbf{I}\\ \mathbf{0}\end{pmatrix}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_j\mathbf{Z}\mathbf{L}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'(\mathbf{I}, \mathbf{0})\mathbf{Q}\right]^{\ddagger}$$

$$-\rho\theta_i\mathbf{Q}'\begin{pmatrix}\mathbf{I}\\ \mathbf{0}\end{pmatrix}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{U}(\mathbf{I}\rho + \mathbf{T})^{-1}\mathbf{U}'(\mathbf{I}, \mathbf{0})\mathbf{Q}\delta_{ij}$$

where $\delta_{ij} = 0$ if $i \ne j$, $\delta_{ii} = 1$ and I have used $\mathbf{M}^{\ddagger}$ to mean $\mathbf{M} + \mathbf{M}'$.

The gradient vector and Hessian of V w.r.t. the η_i's are needed. For notational compactness let $V = \alpha/\beta$ (by definition of α and β) so that the gradients are given by $\partial V/\partial\eta_i = \partial\alpha/\partial\eta_i/\beta - \partial\beta/\partial\eta_i\alpha/\beta^2$, for example. To obtain the gradient and Hessian requires the following:

$$\alpha = \frac{1}{n}\left(\mathbf{y}'\mathbf{W}\mathbf{y} - 2\mathbf{y}'\mathbf{W}^{1/2}\tilde{\mathbf{A}}\mathbf{W}^{1/2}\mathbf{y} + \mathbf{y}'\mathbf{W}^{1/2}\tilde{\mathbf{A}}'\tilde{\mathbf{A}}\mathbf{W}^{1/2}\mathbf{y}\right)$$

so,

$$\frac{\partial\alpha}{\partial\eta_i} = \frac{2}{n}\left(-\mathbf{y}'\mathbf{W}^{1/2}\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_i}\mathbf{W}^{1/2}\mathbf{y} + \mathbf{y}'\mathbf{W}^{1/2}\frac{\partial\tilde{\mathbf{A}}'}{\partial\eta_i}\tilde{\mathbf{A}}\mathbf{W}^{1/2}\mathbf{y}\right)$$

and

$$\frac{\partial^2\alpha}{\partial\eta_i\partial\eta_j} = \frac{2}{n}\left(-\mathbf{y}'\mathbf{W}^{1/2}\frac{\partial^2\tilde{\mathbf{A}}}{\partial\eta_i\partial\eta_j}\mathbf{W}^{1/2}\mathbf{y} + \mathbf{y}'\mathbf{W}^{1/2}\frac{\partial^2\tilde{\mathbf{A}}'}{\partial\eta_i\partial\eta_j}\tilde{\mathbf{A}}\mathbf{W}^{1/2}\mathbf{y} + \mathbf{y}'\mathbf{W}^{1/2}\frac{\partial\tilde{\mathbf{A}}'}{\partial\eta_i}\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_j}\mathbf{W}^{1/2}\mathbf{y}\right).$$

Turning to β:

$$\beta = \left\{1 - \frac{1}{n}tr(\tilde{\mathbf{A}})\right\}^2$$

from whence:

$$\frac{\partial\beta}{\partial\eta_i} = \frac{2}{n}\left\{\frac{1}{n}tr(\tilde{\mathbf{A}}) - 1\right\}tr\left(\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_i}\right)$$

and

$$\frac{\partial^2\beta}{\partial\eta_i\partial\eta_j} = \frac{2}{n^2}tr\left(\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_j}\right)tr\left(\frac{\partial\tilde{\mathbf{A}}}{\partial\eta_i}\right) - \frac{2}{n}\left\{1 - \frac{1}{n}tr(\tilde{\mathbf{A}})\right\}tr\left(\frac{\partial^2\tilde{\mathbf{A}}}{\partial\eta_i\partial\eta_j}\right).$$

Given $(\mathbf{I},\mathbf{0})\mathbf{Q}\mathbf{W}^{1/2}\mathbf{y}$, evaluation of the various derivatives of α requires only $O(q^2)$ operations, but the derivatives of β are less straightforward, because of the terms $tr(\partial\tilde{\mathbf{A}}/\partial\eta_i)$ and $tr(\partial^2\tilde{\mathbf{A}}/\partial\eta_i\partial\eta_j)$. The following approach minimises the computational burden. In many circumstances each penalty term only applies to a subset of the model parameters, so that the penalty matrices $\mathbf{S}_i$ only contain a relatively small block of non-zero elements. For this reason it is worth finding square roots of the $\mathbf{S}_i$'s each having as few columns as possible (see below). The matrices $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i^{1/2}$ are then obtained once only. For each new $\boldsymbol{\eta}$, the matrices $\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i^{1/2}$ are formed (in the common situation that the penalties apply to non-overlapping subsets of the parameters, this step has a total operation count of $q^3/2$). If $\mathbf{K}$ is a (bidiagonal) Choleski factor of $\mathbf{I}\rho + \mathbf{T}$ then the next step is to solve $\mathbf{K}\mathbf{K}'\mathbf{F}_i = \mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i^{1/2}$ for $\mathbf{F}_i$ (operation count $O(q^2)$) before evaluating $tr(\partial\tilde{\mathbf{A}}/\partial\eta_i) = \rho\theta_i tr(\mathbf{F}_i\mathbf{F}_i')$ in $\le q^2$ operations.

The second derivatives require formation of the matrices $\mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{U}$: since square roots of these are already available, this is usually achieved in a total of q^3 operations (for the non-overlapping penalty case). It's then a matter of solving the m equations $\mathbf{K}\mathbf{K}'\mathbf{D}_i\mathbf{K}' = \mathbf{U}'\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{U}$ for $\mathbf{D}_i$ (again $O(q^2)$). The first term in $tr(\partial^2\tilde{\mathbf{A}}/\partial\eta_i\partial\eta_j)$ is now given in q^2 operations by $2\rho\theta_i\theta_j tr(\mathbf{D}_i\mathbf{D}_j')$ (since $tr(\mathbf{M}) = tr(\mathbf{M}')$, in general) and the second term is already available, when needed.

Now consider forming the $\mathbf{S}_i^{1/2}$'s efficiently. Let $\tilde{\mathbf{S}}_i$ be the smallest sub-matrix containing all the non-zero elements of $\mathbf{S}_i$. I obtained $\mathbf{H}_i$'s such that $\tilde{\mathbf{S}}_i = \mathbf{H}_i\mathbf{H}_i'$ as follows: (i) Perform a Householder tridiagonalisation: $\tilde{\mathbf{S}}_i = \mathbf{P}\mathbf{N}\mathbf{P}'$ (where $\mathbf{P}$ is orthogonal and $\mathbf{N}$ is tri-diagonal). (ii) Obtain the Choleski decomposition of the non-zero part of $\mathbf{N}$, so that $\mathbf{N} = \mathbf{B}\mathbf{B}'$. (iii) Set $\mathbf{H}_i = \mathbf{P}\mathbf{B}$. Now typically: $\mathbf{S}_i^{1/2} = (\mathbf{0},\mathbf{H}_i',\mathbf{0})'$. (It might be considered more straightforward to use an algorithm based on truncated singular value decomposition, although this would be less efficient.)

Minimisation of V with respect to the θ_i/ρ's is done iteratively. Exact searches for ρ given $\boldsymbol{\eta}$ are alternated with Newton updates of $\boldsymbol{\eta}$ given ρ (with $\boldsymbol{\eta}$ normalised to avoid overflow or underflow, which is legitimate since only the relative sizes of the η_i's matter). Under some circumstances the Hessian may not be positive definite, and some regularization will be required.

Reasonable starting values (see Wahba 1990) are:

$$\theta_i^{(0)} = \frac{1}{n}\{tr(\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L})\}^{-1} \quad \theta_i^{(1)} = n(\theta_i^{(0)})^2\mathbf{p}'\mathbf{Z}\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}\mathbf{Z}'\mathbf{p}.$$

The operation count for the algorithm falls into two parts: start up costs and costs per iteration. Start up operations consist of an initial QR decomposition and the formation of matrices of the form $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i^{1/2}$ and $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}$: clearly the $O(nq^2)$ QR decomposition dominates these costs in the usual case in which n is substantially greater than q. The operation count per iteration depends on the nature of the quadratic penalties. The most costly steps are the formation of the $2m$ matrices $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i^{1/2}$ and $\mathbf{L}'\mathbf{Z}'\mathbf{S}_i\mathbf{Z}\mathbf{L}$: the total operation count for these steps varies between $O(q^3)$ and $O(mq^3)$, depending on the extent to which penalties overlap in the parameters which they penalize.

Clearly GCV is not the only sensible criterion for choosing multiple smoothing parameters, but the method given is easily modified to employ other criteria where the awkward term is the trace of the influence matrix .

5. Example: a non-linear model

The first example concerns a non-linear model with multiple penalties, where the non- linearity arises from attempting to incorporate basic biology into the model to be fitted. Marine copepods are crustacea that form an important link in many marine food chains. Typically they eat phytoplankton and are eaten by fish larvae. Mortality rates can be high for such organisms, but are impossible to observe directly: there is a sizeable literature devoted to methods for estimating these rates from time series of population data (see Asknes *et al.* 1997). As copepods age they pass through a series of clearly identifiable stages of fixed duration separated by moults. There are typically 11 stages before adulthood, but the major physiological changes are between the 6^{th} and 7^{th}. From the 7^{th} stage the animals swim much more strongly and are better able to avoid predators. A reasonable model of a stage structured copepod population can be constructed using standard delay differential equation methodology (see Gurney and Nisbet, 1998, for the basic model structure). The equation for the population in stage i is:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = R_i(t) - M_i(t) - \mu_i(t)n_i(t)$$

where the maturation rate M_i can be expressed in terms of the recruitment rate, R_i, the mortality rate, μ_i, and the stage duration, τ_i, using $M_i = R_i(t-\tau_i)exp\left(-\int_{t-\tau_i}^{t}\mu_i(x)dx\right)$. For all but the first stage $R_{i+1}(t) = M_i(t)$. The model neglects demographic stochasticity, which is reasonable for large populations over short timescales, particularly when sampling error is large. Given the known biology, it is reasonable to set μ_1 to μ_6 to the same unknown function of time and similarly to set μ_7 to μ_{11} equal to another unknown function of time. $R_1(t)$ is also an unknown function of time and the initial conditions at $t = 0$ were set to be consistent with $\mu_i(t) = \mu_i(0)$ and $R_i(t) = R_i(0)$ for all $t < 0$.

To fit this model in practice the three unknown functions were each represented using a cubic spline basis, each was penalized with a standard cubic spline wiggliness penalty, and each had natural end conditions (see section 2.2 of Green and Silverman, 1994, for example). The parameterization was set up so that the free parameters for each spline were the values of the spline at a set of design points (knots) in its domain: 15 evenly spaced knots were employed per spline. Given the values of the free parameters of the splines the model can be integrated numerically to give predicted populations in any stage at any relevant time: adaptive time-stepping with an embedded Runge-Kutta 2(3) scheme was used, along with cubic Hermite interpolation of lagged variables (see Highman, 1993, for a description of how to integrate delay differential equations accurately). The model was fitted by penalized least squares using the iterative least squares method for penalized non-linear models given in Section 2, with smoothing parameters estimated by the method of sections 3 and 4 (and $W_{ii} = 1, \forall i$). The Jacobian matrices, $\mathbf{J}^{[k]}$, required for this, were estimated by finite differencing (although it is possible to derive a system of delay differential equations to be

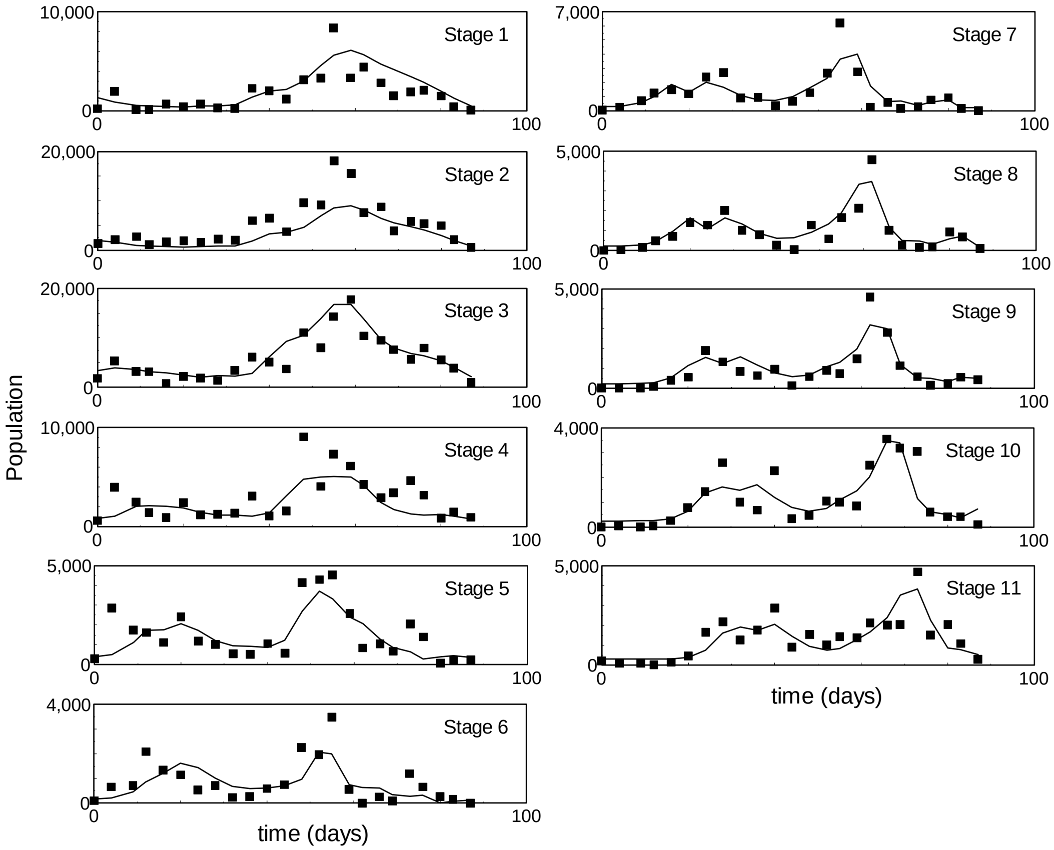

Fig. 1. Model predictions (joined by solid line) of data (squares), for a non-linear population dynamic model fitted to marine copepod data. The model has three unknown functions: birth rate, mortality rate in first 6 stages and mortality rate in remaining stages. These functions were modelled using penalized regression splines. The model was fitted by iterative least squares, with smoothing parameters chosen as described in the text.

solved for the elements of $\mathbf{J}^{[k]}$): differences must be selected carefully (see e.g. Gill *et al.*, 1981, section 8.6) and the same integration mesh should be used for both model integrations used for each finite difference estimate (different meshes can lead to severe loss of accuracy and precision).

The model was fitted to data for a mesocosm population of *Pseudocalanus* taken from Hay *et al.* (1988, 1991). The sampling protocol is not straightforward, but the data are counts of individuals in each life history stage caught in net hauls multiplied by the volume of the mesocosm divided by the volume of water filtered by a net haul. Figure 1 shows the model fit and data, stage by stage, while figure 2 shows best fit functions and approximate '95% confidence bands' calculated using the approximating linear model at convergence. There are a number of ways of calculating such bands. The ones shown are similar to the Bayesian intervals developed by Wahba (1983), and can be justified by making the large sample approximation that $\mathbf{z}|\mathbf{p} \sim N(\mathbf{J}\mathbf{p}, \mathbf{I}\sigma^2)$ and assuming a multivariate normal prior on $\mathbf{p}$ with mean $\mathbf{0}$ and covariance matrix proportional to an appropriate pseudoinverse of $\sum_i \theta_i \mathbf{S}_i$: see Hastie and Tibshirani, (1990) section 3.6, based on Silverman (1985) (they assume

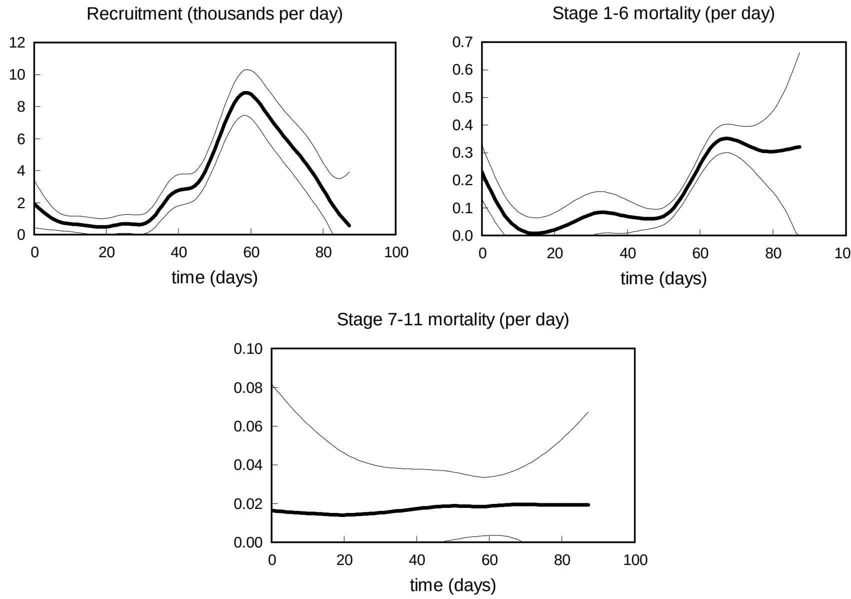

Fig. 2. Estimated birth (recruitment) and mortality rate functions for the marine copepod population. Thick line is estimate, thin lines are upper and lower 95% 'confidence limits'.

square $\mathbf{J}$, but this is not necessary). Neglecting constraints and assuming uniform weights, the estimated posterior covariance matrix for the parameters is given by $\hat{\sigma}^2(\mathbf{J}'\mathbf{J} + \sum_i \hat{\theta}_i \mathbf{S}_i/\hat{\rho})^{-1}$, where $\hat{\sigma}^2 = \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2/[tr(\mathbf{I} - \mathbf{A})]$ and all quantities are estimated at convergence. The approximate posterior distribution of $\hat{\mathbf{p}}$ is multivariate normal, so approximate confidence intervals for the parameters are easily obtained. Since the splines have been parameterised in terms of function values at discrete times, it is straightforward to use approximate confidence intervals for these parameters to construct confidence intervals for the splines: for figure 2 I plotted the upper 95% confidence limit for each parameter at the appropriate time and interpolated these points with a cubic spline to get a smooth curve; the same procedure was used for the lower limit. Other approximations can be used to obtain confidence intervals (e.g. next section) and there are also strong arguments for obtaining intervals by bootstrapping (Wang and Wahba, 1995).

This example is fairly specialised, but the general approach is not and should be widely applicable.

6. Example: Generalized additive models with penalized regression splines

As a second example, consider the problem of finding $f(\mathbf{x})$ in the model:

$$g(E[Y_i]) = f(\mathbf{x}_i) \quad Y_i \sim \text{Distribution from exponential family}$$

given a vector of observations $\mathbf{y}$ on the random variables $\mathbf{Y}$ and for each y_i a vector of covariates $\mathbf{x}_i$. g is a monotonic C^2 link function. Generalized additive modelling (Hastie and Tibshirani, 1990) consists of simplifying this problem with a decomposition something like:

$$f(\mathbf{x}) = a + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4) + \dots$$

i.e. into the sum of a parameter and low dimensional smooth functions of the covariates (often functions of only one covariate are used, and strictly parametric terms can be included without difficulty). This decomposition is sometimes justified by prior knowledge. As written the model is underdetermined in general: constraints are needed on all functions to avoid degeneracy. Suitable constraints are that the functions should integrate to zero (which can easily be written as linear equality constraints as required).

It is straightforward to represent the functions in this decomposition using penalized polynomial regression splines in the one dimensional case (e.g. Eilers and Marx, 1996) and, for functions of more than one covariate, to use penalized regression splines based on the basis functions for thin plate splines or tensor product splines (e.g. Green and Silverman, 1994). Model fitting is then done by penalized maximum likelihood, using the iterative least squares scheme given in Section 2, with smoothing parameters re-estimated at each iteration by the method of section 4 (using the pseudodata in place of original data, and the iteratively calculated weights for $\mathbf{W}$).

It is frequently desirable to remove spurious covariates, and some means are needed of selecting terms for removal. In the identity link case $\hat{\mathbf{p}} = \mathbf{B}\mathbf{y}$ where $\mathbf{B} = \rho\mathbf{Z}\ (\mathbf{Z}'\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{Z}\rho + \sum_{i=1}^{m} \theta_i \mathbf{Z}'\mathbf{S}_i\mathbf{Z})^{-1}\ \mathbf{Z}'\mathbf{X}'\mathbf{W}$, so if $\mathbf{p}_i$ is the subset of parameters relating to the i^{th} covariate, then a submatrix of $\mathbf{B}$, $\mathbf{B}_i$ gives $\hat{\mathbf{p}}_i = \mathbf{B}_i\mathbf{y}$. An approximate covariance matrix $\mathbf{V}_i$ for $\hat{\mathbf{p}}_i$ is easily obtained: $\mathbf{V}_i = \hat{\sigma}^2\mathbf{B}_i\mathbf{W}^{-1}\mathbf{B}_i'$, where $\hat{\sigma}^2 = \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{A}\mathbf{y})\|^2/tr(\mathbf{I} - \mathbf{A})$ (For some error models, if overdispersion is not a problem, $\hat{\sigma}^2$ can be set to one). If the covariate has no explanatory power then a high smoothing parameter is appropriate and its covariance matrix should be close to rank d, where d is the minimum possible degrees of freedom of the model term. In this circumstance it is expected that approximately: $\hat{\mathbf{p}}_i'\mathbf{V}_i^{d+}\hat{\mathbf{p}}_i \sim \chi^2_d$, where $\mathbf{V}_i^{d+}$ is the rank d pseudoinverse of $\mathbf{V}_i$ (calculated by truncating all but the largest d singular values when inverting the matrix using its singular value decomposition).

Model selection can now proceed by deletion of the least significant of any covariates having p values greater than some predetermined level. Limited Monte-Carlo experimentation with models having 3 real and 3 nuisance covariates suggests that the approach is effective. If large numbers of covariates are being ‘screened’, then it may be efficient to tolerate increased bias in the early stages of model selection, by using few knots per spline, increasing the knots per spline as the model becomes more parsimonious.

In some circumstances interest may focus on testing for interaction terms. This requires some care. Consider an interaction between x_1 and x_2. An appropriate model structure might be:

$$f(\mathbf{x}) = \alpha + f_1(x_1) + f_2(x_2) + f_3(x_1, x_2) + \ldots$$

clearly the space spanned by two univariate functions is a subspace of the space of bivariate functions, so that this model is not identifiable in general, but this is not necessarily a property of the approximating spaces defined by any particular representation in terms of basis functions. For practical modelling it may be possible to avoid these difficulties by setting the model up in possibly degenerate form, and then applying appropriate equality constraints to parameters that are aliased as a result.

As an example I fitted a simple additive model to the mackerel egg abundance data discussed in much greater detail in Augustin *et al.* (1998) and Borchers *et al.* (1997). The data result from a fish egg survey in which survey vessels sampled fish eggs by net tows over a large but irregular grid in the Eastern Atlantic. At each of 1165 sampling stations there are a number of covariates available: date, latitude, longitude, time of day, sea-bed depth, temperature at 20 metres depth, surface temperature, sampling depth and distance from the 200m sea-bed depth contour. Samples were modelled with an overdispersed Poisson error structure, and a parametric covariate proportional to effective

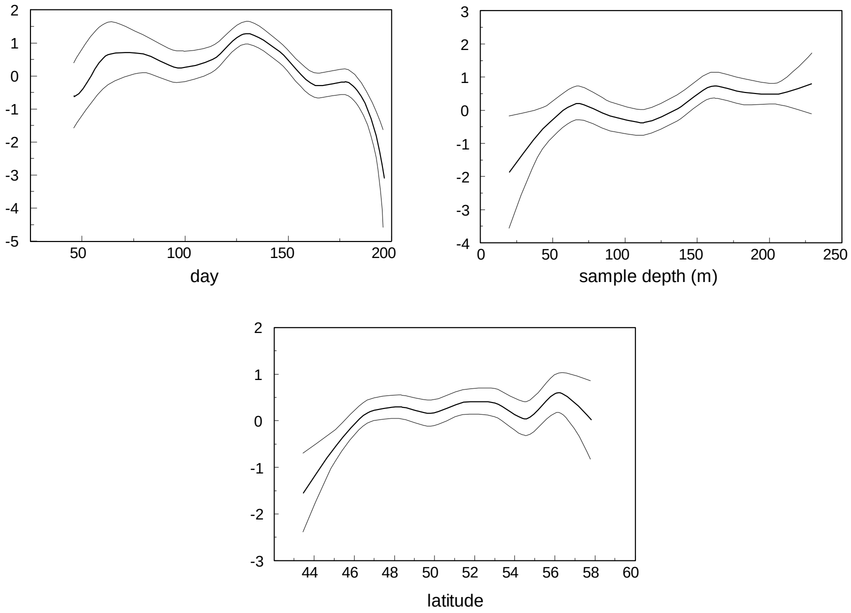

Fig. 3. The selected fitted smooths for the mackerel egg example of section 6. The thick lines are the final selected regression splines, with smoothing parameter chosen by GCV. The thin lines show corresponding 95% confidence intervals.

sea surface area sampled. Response to the 9 other covariates was modelled using penalized cubic regression splines with 10 knots each (i.e. a maximum of 10 degrees of freedom each): the splines were parameterised in terms of the function values at the knots. A log link function was employed so that the covariates were combined multiplicitively. For justification of this model structure and the initial set of covariates, see Borchers *et al.* (1997) and Augustin *et al.* (1998). Since the primary purpose of this paper is not the analysis of these data, I ignored interaction terms.

Model selection proceeded sequentially in the manner suggested above, with the final model containing date, sample depth and latitude. All removed terms would have been rejected using p values up to 0.5, except for the last term removed, sea bed depth, for which p was just below 0.1. Figure 3 shows plots of the selected model terms. Approximate 95% confidence limits on the parameters were obtained as follows: conditional on the estimates of the θ_i/ρ's, the parameter estimators for the i^{th} spline are approximately distributed as $N(\mathbf{p}_i, \mathbf{V}_i)$, so confidence intervals for each parameter can be obtained; the parameters give the values of the spline at a set of covariate values, so it is possible to plot the upper and lower confidence limits for each parameter at the appropriate covariate value and interpolate to obtain approximate continuous pointwise confidence intervals. This was the approach used for figure 3.

7. Example: Anisotropic smoothing

Consider the problem of estimating a function g of two covariates, x and t, say, given noisy observations:

$$y_i = g(x_i, t_i) + \epsilon_i,$$

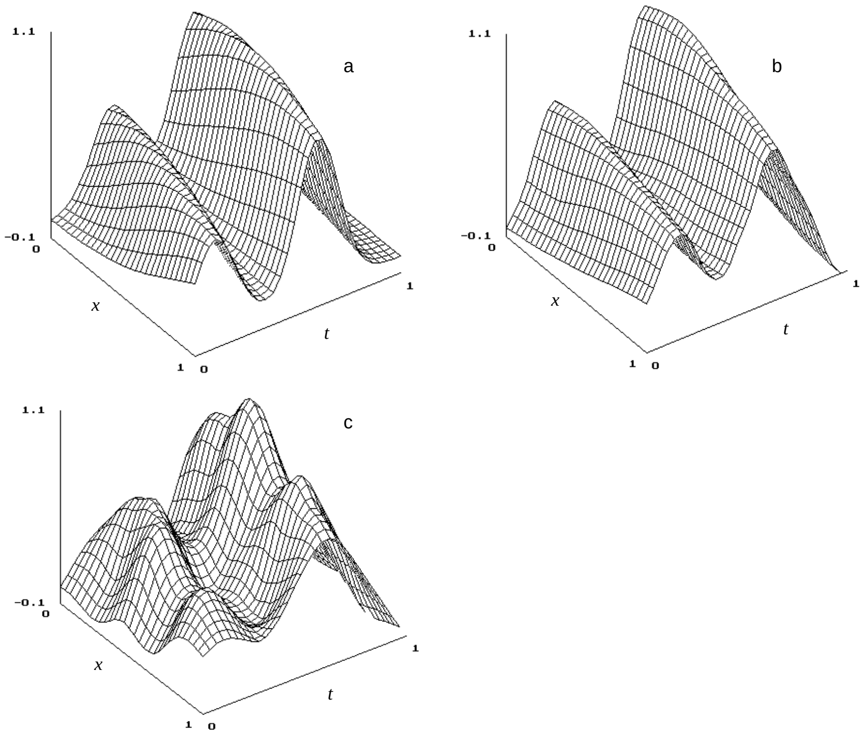

Fig. 4. Attempt to reconstruct the bivariate test function, $f(x,t) = e^{-(x-0.5)^2}\ [0.6e^{-(t-0.3+0.2x)^2} + e^{-(t-0.7+0.2x)^2}]$ (a) by rescaling method described in the text (b) and by conventional thin plate spline smoothing with smoothing parameter selected by GCV (c). The reconstructions are from datasets created from 100 evaluations with x and t co-ordinates each randomly selected from a uniform distribution on $[0,1]$. Independent guassian noise ($\sigma = 0.2$) was added to each evaluation. In 100 replicate reconstructions the mean square deviation of the rescaled surface from the true surface was on average 0.6 of the equivalent deviation for the thin plate spline without rescaling.

where the ϵ_i's are observations on independent zero mean random variables. A common example of this sort of problem is the estimation of interaction terms in generalized additive models. In the absence of a parametric model g can be estimated by smoothing. Whatever smoothing method is chosen, it is not possible to avoid choosing both the bandwidth of the smoother and the relative scaling of x and t, although it is common practice to pretend that the latter issue can be avoided by ignoring it (but see e.g. Hutchinson and Bischof, 1983, for a counter example). In some cases both covariates represent quantities with identical physical interpretation (for example x and t may be position data) and there may be strong arguments for setting the scaling to unity. In cases in which the covariates are not even measured in the same units their relative scaling should usually be estimated.

This problem is quite straightforward to address with the multiple smoothing parameter methodology developed above by using a thin plate spline smoother. In the two dimensional case the thin plate spline f is fitted by minimisation of:

$$\|\mathbf{y}-\mathbf{f}(x,t)\|^2+\lambda\int\int_{\Omega}\left(\frac{\partial^2 f}{\partial x^2}\right)^2+2\left(\frac{\partial^2 f}{\partial x\partial t}\right)^2+\left(\frac{\partial^2 f}{\partial t^2}\right)^2 dxdt$$

where Ω is the region of the $x-t$ plane that is of interest (see Wahba 1990 or Green and Silverman 1994 for full details). The magnitude of the penalty in this case is dependent on the relative scaling of x and t. For example if the t_i co-ordinate of each observation y_i is replaced with new co-ordinate $t_i' = t_i/k$ then $f(x,t'k)$ will give the same fit to the data as $f(x,t)$ gave under the old co-ordinates, but the co-ordinate change will alter the penalty to:

$$\lambda\int\int_{\Omega}\left(\frac{\partial^2 f}{\partial x^2}\right)^2+2k\left(\frac{\partial^2 f}{\partial x\partial t}\right)^2+k^3\left(\frac{\partial^2 f}{\partial t^2}\right)^2 dxdt.$$

Clearly k could be selected along with λ and this is facilitated by splitting the single penalty into 3 penalty terms with associated smoothing parameters λ, λk and λk^3. Given a suitable basis, these penalties can each be written in the form $\mathbf{p}'\mathbf{S}_i\mathbf{p}$ so that the thin plate spline problem with scaling becomes a generalized ridge regression problem similar to (2). In the work reported here I used the thin plate spline basis given, for example, in section 7.4 of Green and Silverman (1994) and the $\mathbf{S}_i$'s were obtained numerically using this basis. The additional twist in this case is that the smoothing parameters k and λ enter the ridge regression problem in a non-linear way. However the strategy of section 4 can be applied unaltered provided that the gradient vector and Hessian matrix of the GCV score, V, w.r.t. $\log(\lambda)$ and $\log(k)$ can be obtained: this is routine, since the method of section 4 already yields the gradient and Hessian of V w.r.t. $\log(\lambda)$, $\log(\lambda k)$ and $\log(\lambda k^3)$ so that simple transformations yield the required vector and matrix.

Figure 4 shows the results of reconstructing a test surface using a conventional single smoothing parameter thin plate spline fitted by GCV and using the scaling method suggested above. k and λ were estimated using a single application of the method and then the re-scaling of t implied by the estimate of k was applied to the t_i's before re-fitting a conventional thin plate spline: this last step is advantageous since it allows the shape of the basis functions to change with k. It should be clear that the approach is easy to generalize beyond two dimensional smoothing.

8. Discussion

The aim of the work reported here was to develop general methodology which would enable the use of models with multiple penalties. This was driven by the applied needs to work with non-linear models with multiple penalties and to find a defensible way of dealing with interaction terms in models with a GAM structure, as well as a desire to find a less heuristic approach to GAM model selection. As the examples show, multiple smoothing parameter selection is the key to solving these modelling problems. The method presented makes the use of non-linear models with multiple penalties feasible (and in many cases straightforward) for the first time. Similarly, the same methodology suggests a significant advance in non-parametric modelling with multidimensional smoothers. GAMs (Hastie and Tibshirani, 1990) should benefit from the removal of the computationally driven need to approximate GCV for model selection, an efficient means for choosing the complexity of multidimensional terms and an objective method for modelling interactions by explicit estimation of the appropriate relative scaling of variables. Similarly the modelling approach based on formulating models as inhabitants of reproducing kernel Hilbert spaces (e.g. Wahba 1990) is improved by the availability of a smoothing parameter selection method that is applicable when there is no choice but to use an approximating basis function representation for the model: for example when a closed form expression for the reproducing kernel can not be found, or when there are too many data to consider modelling with one parameter per datapoint.

Techno- optimists, brazenly extrapolating a physics free exponential growth model for computer speed, might argue that, given a decade or so, direct evaluation of model selection criteria, like GCV and AIC, will be perfectly feasible without the need for the methodology presented here. While superficially appealing this argument seems less compelling after consideration of a simple example: imagine a dataset of 1000 observations modelled using 100 parameters and 4 penalties. A gridsearch for the smoothing parameters allowing each to take one of 10 values yields an operation count of order 10^{11}, whereas the method presented here would have operation count of order 10^7. As the number of penalties increases this difference becomes even more extreme. The corollary is that, for any level of computer technology, the methods presented here will always allow the practical use of more sophisticated models and analyses than can be achieved by the brute force approach (and in any case are a good deal more elegant).

In terms of applications, the most interesting uses of models with multiple penalties probably relate to models which combine the non-parametric with the mechanistic in the manner of the zooplankton example given above. Scientific models are frequently a combination of rather well known mechanisms acting on processes about which there is much less information: in the zooplankton example the basic demography can be described quite precisely by some non-linear balance equations, but the essential demographic processes of birth and death are almost impossible to measure directly, and are certainly not sufficiently well known for simple parameter sparse description. In that example the model is a fair reflection of the knowledge of the workings of the population. Hence the precision of estimates of the unobserved population process rates benefits from the model, while model mis-specification bias is minimised. Such models should be useful in both physical and biological sciences: especially so in environmental disciplines where data are sparse and expensive and knowledge of mechanism is rarely so complete as to completely specify the structure of a fully parametric model.

Acknowledgements

Thanks to those responsible for collecting the Mackerel data used in the GAM example: A. Eltink, P. Lucio, J. Masse, L. Motos, J. Molloy, J. Nichols, C. Porteiro, and M. Walsh. Thanks also to Steve Hay who originally provided the copepod problem as well as Steve Buckland and most of the anonymous referees for very helpful comments on the manuscript.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory* (eds B.N. Petrov and F. Csaki), pp.267-281. Budapest, Hungary: Akademiai Kiado.

Asknes, D.L., Miller, C.B. Ohman, M.D. and Wood, S.N. (1997) Estimation techniques used in studies of copepod population dynamics - a review of underlying assumptions. *Sarsia* **82**, 279-296.

Augustin, N.H., Borchers, D.L., Clarke, E.D., Buckland, S.T. and Walsh, M. (1998) Spatio-temporal modelling of annual egg production of fish stocks using generalized additive models. *Can. J. Fish. Aq. Sci* **55**, 2608-2621.

Borchers, D.L., Buckland, S.T., Priede, I.G. and Ahmadi, S. (1997) Improving the precision of the daily egg production method using generalized additive models. *Can. J. Fish. Aq. Sci* **54**, 2727-2742.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377-403.

Eilers, P.H.C. and Marx, B.D. (1996) Flexible Smoothing with B-splines and Penalties *Statistical Science* **11**, 89-121.

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

Gill, P.E., Murray, W. and Wright, M.H. (1981) *Practical Optimization.* London: Academic Press.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric regression and generalized linear models.* London: Chapman and Hall.

Gu, C and Wahba, G. (1991) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comp.*, **12(2)**, 383-398.

Gu, C and Wahba, G. (1993) Semiparametric analysis of variance with tensor product thin plate splines. *J. R. Statist. Soc. B* **55(2)**, 353-368.

Gurney, W.S.C. and Nisbet, R.M. (1998) *Ecological Dynamics.* New York: Oxford University Press.

Hastie. T.J. and Tibshirani, R.J. (1990) *Generalized additive models.* London: Chapman and Hall.

Hay, S.J., Evans, G.T. and Gamble, J.C. (1988) Birth, death and growth rates for enclosed populations of calanoid copepods. *J. Plank. Res.* **10**, 431-454.

Hay, S.J., Wood, S.N. and Nisbet, R.M. (1991) Loch Ewe Copepods: Some Speculation. In *Estimation of Mortality Rates in Stage-Structured Populations* pp 77-91, Wood, S.N. and Nisbet, R.M. Berlin: Springer-Verlag.

Highman, D.J. (1993) Error control for initial value problems with discontinuities and delays. *Applied Numerical Analysis* **12**, 315-330.

Hutchinson, M.F. and Bischof, R.J. (1983) A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New South Wales *Aust. Met. Mag.* **31**,179-184.

Hutchinson, M.F. and deHoog, F.R. (1985) Smoothing noisy data with spline functions. *Numer. Math.* **47**, 99-106.

O'Sullivan, F. Yandell, B.S. and Raynor, W.J. (1986) Automatic smoothing of regression functions in generalized linear models *J. Am. Statist. Ass.* **81**, 96-103.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J.R. Statist. Soc. B* **47**,1-52.

Utreras, F. (1981) Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comput.* **2**, 349-362.

Wahba, G. (1983) Bayesian confidence intervals for the cross validated smoothing spline. *J. R. Statist. Soc. B* **45**, 133-150.

Wahba (1990) *Spline models for observational data. CBMS-NSF Reg. Conf. Ser. Appl. Math.*: **59**.

Wahba, G., Wang, Y., Gu, C., Klein, R. And Klein, B. (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23(6)**, 1865-1895.

Wang, Y, and Wahba, G. (1995) Bootstrap Confidence Intervals for Smoothing Splines and their Comparison to Bayesian 'Confidence Intervals'. *J. Statist. Comput. & Simul.* **51**, 263-297.