Stable and efficient multiple smoothing parameter estimation for generalized additive models

Simon N. Wood

Department of Statistics, University of Glasgow

Glasgow G12 8QQ U.K.

March 30, 2004

Abstract

Representation of Generalized Additive Models using penalized regression splines allows GAMs to be employed in a straightforward manner using penalized regression methods. Not only is inference facilitated by this approach, but it is also possible to integrate model selection/ smoothing parameter selection into model fitting in a computationally efficient manner using well founded criteria such as GCV. The current fitting and smoothing parameter selection methods for such models are usually effective, but do not provide the level of numerical stability to which users of linear regression packages, for example, are accustomed. In particular the existing methods can not deal adequately with numerical rank deficiency of the GAM fitting problem, and it is not straightforward to produce methods which can do so, given that the degree of rank deficiency can be smoothing parameter dependent. In addition, models with the potential flexibility of GAMs can also present practical fitting difficulties as a result of indeterminacy in the model likelihood: data with many zeros fitted by a model with a log link are a good example. In this paper it is proposed that GAMs with a ridge penalty provide a practical solution in such circumstances, and a multiple smoothing parameter selection method suitable for use in the presence of such a penalty is developed. The method is based on the pivoted

QR decomposition and the Singular Value Decomposition, so that with or without a ridge penalty it has good error propagation properties and is capable of detecting and coping elegantly with numerical rank deficiency. The method also allows mixtures of user specified and estimated smoothing parameters and the setting of lower bounds on smoothing parameters. In terms of computational efficiency the method compares well with existing methods. A simulation study compares the method to existing methods, including treating GAMs as mixed models.

Keywords: Generalized additive model; GAM; Generalized additive mixed model; spline; SS-ANOVA; generalized cross validation; GCV; REML; ridge regression; regularization; stable computation; multiple smoothing parameter estimation; Penalized Quasi-Likelihood; PQL.

A generalized additive model (GAM, Hastie and Tibshirani, 1986, 1990) is a generalized linear model (GLM, McCullagh and Nelder, 1989) where the linear predictor is specified as a sum of smooth functions of some or all of the covariates. For example,

$$E(Y_i) \equiv \mu_i, \quad g(\mu_i) = \eta_i \equiv \mathbf{X}_i^* \boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \ldots$$

where response variable $Y_i \sim$ 'an exponential family distribution'; g is a monotonic link function; $\mathbf{X}_i^*$ is the ith row of $\mathbf{X}^*$, the model matrix for the strictly parametric part of the model, with corresponding parameter vector $\boldsymbol{\beta}^*$, and the f_j are smooth functions of the covariates $\mathbf{x}_j$ (an f_j may be a function of more than one covariate). To avoid over fitting, such models are estimated by *penalized* maximum likelihood estimation, for example by maximizing:

$$l(\boldsymbol{\eta}) - \frac{1}{2}\sum_j \theta_j \int [f_j''(x)]^2 dx$$

where l is the log likelihood of the linear predictor and the terms in the summation are measures of the wiggliness of the component functions of the GAM, serving to penalize models with overly complicated component functions (the particular form of the penalty is just an example, there are numerous alternative possibilities). The θ_i are smoothing parameters controlling the tradeoff between fit and smoothness. In practice the penalized likelihood is maximized by penalized iteratively reweighted least squares (P-IRLS). For example, given the likelihood above, at the k^{th} P-IRLS iteration the following penalized sum of squares would be minimized with respect to $\boldsymbol{\eta}$ to find the $(k+1)^{th}$ estimate of the linear predictor, $\boldsymbol{\eta}^{[k+1]}$:

$$\|\mathbf{W}^{[k]}(\mathbf{z}^{[k]} - \boldsymbol{\eta})\|^2 + \sum_j \theta_j \int [f_j''(x)]^2 dx. \tag{1}$$

$\mathbf{W}^{[k]}$ and $\mathbf{z}^{[k]}$ are iterative weights and pseudodata respectively, and are given by $W_{ii}^{[k]} = 1/\sqrt{g'(\mu_i^{[k]})^2 V_i^{[k]}}$ and $z_i^{[k]} = \eta_i^{[k]} + g'(\mu_i^{[k]})(y_i - \mu_i^{[k]})$, where $V_i^{[k]}$ is proportional to the variance of Y_i according to the current estimate $\mu_i^{[k]}$. The obvious extra difficulty introduced by the use of the penalized likelihood approach is that the smoothing parameters, θ_i, have to be estimated.

GAMs have become popular due largely to the work of Hastie and Tibshirani (1986, 1990) and the availability of well designed software implementing their approach in S-PLUS. Hastie and Tibshirani minimized (1) by modified backfitting (Buja *et al.*, 1989), in which the component functions can be estimated using no more than simple linear scatterplot smoothers and standard least squares methods. However, θ_i estimation and reliable confidence interval calculation is difficult to integrate into this approach. Parallel to the GAM work of Hastie and Tibshirani the Smoothing Spline ANOVA (SS-ANOVA) work of Gu and Wahba, in particular, (Gu and Wahba, 1993; Wahba *et al.* 1995; Wahba, 1990; Gu, 2002) has provided a mathematically elegant theory for function estimation, including GAMs. The SS-ANOVA approach operates by finding the *functions* minimizing (1) out of all reasonable candidate functions. Gu and Wahba (1991) developed well founded smoothing parameter selection methods for these models, as well as confidence intervals with good coverage probabilities (see e.g. Wahba, 1990, Gu, 2002), but this appealing generality comes at high computational cost — the methods generally require the estimation of one parameter per datum so that the computational cost scales as the cube of the number of data (but see Gu and Kim, 2002).

In response to the high computational cost of the SS-ANOVA approach, the problems with inference and smoothing parameter selection in the Hastie and Tibshirani GAM methods, and following early work by, for example, Wahba (1980) and Parker and Rice (1985), several authors have suggested representing GAMs using penalized regression splines. Essentially all that is required is to choose a relatively low rank basis for representing each component function of the GAM, so that it becomes a parametric model, with a single model matrix, and one quadratic penalty on the parameter vector for each wiggliness penalty on the original likelihood. Hastie and Tibshirani (1990), Hastie (1996), Marx and Eilers (1998) and Wood (2000, 2003) have all discussed using penalized regression splines for GAM modelling, with Wood (2000) also providing an efficient smoothing parameter selection method for these models, based on the approach used in Gu and Wahba's (1991) SS-ANOVA specific method. In the Generalized case the method applies Generalized Cross Validation (GCV, Craven and Wahba, 1979) or similar criteria to estimate the smoothing parameters for each problem (1) of the P-IRLS, an approach termed 'performance iteration' by Gu (1992) who introduced it in the generalized spline smoothing case.

Efficient smoothing parameter selection methods are important for practical GAM modelling. It is possible, for example, to perform smoothing parameter selection by direct grid search optimization of criteria such as GCV or AIC, but if the model has more than two or three terms this is usually so computationally expensive as to preclude the kind of careful model building and checking required in most applied contexts. Yet if GAMs are to be used for more than purely exploratory analysis then smoothing parameter selection is a key component of model selection.

Wood (2000) (and Gu and Wahba, 1991, in the SS-ANOVA context) provides an efficient method for smoothing parameter selection, but leaves some significant practical problems un-resolved. A key problem relates to the fitting of GAMs by P-IRLS. It is well known (e.g. McCullagh and Nelder, 1989) that IRLS can fail to converge in cases where the expected value of a response does not correspond to a finite value of the linear predictor in a GLM. Perhaps the simplest example is provided by Poisson regression with a log link. Consider the case of a response observation $y_i = 0$ of a Poisson random variable with mean μ_i and linear predictor $\mathbf{X}_i\boldsymbol{\beta}$, where $\mathbf{X}_i$ is the ith row of a model matrix and $\boldsymbol{\beta}$ a parameter vector. $\boldsymbol{\beta}$ is estimated by IRLS, but at any iterate of this algorithm, we have that the pseudodata for next iterate is:

$$z_i = \mathbf{X}_i\boldsymbol{\beta} + \frac{1}{\mu_i}(0 - \mu_i) = \mathbf{X}_i\boldsymbol{\beta} - 1$$

that is, the target value for the linear predictor at the next iterate is one less than its current value. This situation is not likely to encourage convergence. In the GLM case such inherently divergent tendencies are usually countered by the need for a relatively inflexible GLM model structure to fit the non-zero data points as well, but with a model structure as flexible as a GAM such stabilization often fails to occur. In most such cases, while the IRLS pseudodata is suggesting every wider divergence of the linear predictor, the IRLS weights are becoming progressively smaller. This feature suggests that for practical purposes either a simple ridge penalty or lower bounds on the smoothing parameters could stabilize the iteration. However the inclusion of such fixed penalties changes the GAM fitting problem so that the Wood (2000) and Gu and Wahba (1991) approaches are no longer applicable.

A second class of problems relate to the basic numerical stability of the GAM fitting and smoothing parameter estimation method, even supposing that the IRLS method is convergent. Very wide divergence in the magnitude of iterative weights, near co-incidence of covariate values, poor relative scaling of the covariates of a multi-dimensional smooth, an unfortunate choice of basis to represent a smooth term or simple colinearity or concurvity problems can all lead to numerical rank deficiency of the model matrix of the GAM: the methods of Wood (2000) and Gu and Wahba (1991), while functioning perfectly well on “well behaved” fitting problems do not deal well with these rank deficient cases. However, in penalized regression contexts, effective treatment of numerical rank deficiency, beyond simple co-linearity or concurvity, can be difficult when the smoothing parameters are not known in advance. This is because the penalties tend to act as regularization terms on the model fit, with the result that the degree of ill-conditioning (near rank deficiency) can be smoothing parameter dependent: this presents obvious difficulties when trying to produce an effective smoothing parameter estimation method.

Finally, during practical model building it is often desirable to be able either to put lower bounds on some smoothing parameters, or to supply the values for some smoothing parameters while estimating others. This is particularly important if fully automatic smoothing parameter selection has resulted in one or more model terms clearly over-fitting: it is always important to be able to over-ride automatic model selection. Another quite common reason for wanting to fix or bound smoothing parameters is when smooth terms are in some sense hierarchical: some terms being present in the model only in order to explain variability that can not be explained by the covariates that are really of interest (smooth functions of spatial location are the obvious example). In such circumstances it is often of interest to explore model fits in which these ‘nuisance’ terms are given a series of fixed smoothing parameters, while the ‘interesting’ terms are left with free smoothing parameters: in this way the ‘interesting’ covariates can be forced to do as much of the explanatory work as possible.

This paper aims to address the above issues, by proposing an improved multiple smoothing parameter estimation method which can deal with fixed penalties (such as a ridge penalty, or the penalty resulting from fixing some smoothing parameters), which offers the maximum possible numerical stability and which can deal with rank deficiency even when the numerical rank depends on the smoothing parameters. At the same time the method is only slightly more computationally costly than the method of Wood (2000), and offers the benefit of only requiring standard numerical linear algebra methods readily available in public domain libraries (specifically LINPACK and LAPACK).

An alternative approach to GAM estimation treats the model as a generalized linear mixed model, and the smoothing parameters as the (reciprocals of) variance components. The appendix provides more details. This mixed model approach is compared to the method developed in this paper in simulation comparisons, which also compare the new method to the methods of Wood (2000) and Gu and Wahba (1991).

1 A method for multiple smoothing parameter estimation

As discussed in the introduction, GAM fitting is usually based on penalized iteratively re-weighted least squares (P-IRLS), and to fit a GAM and estimate its smoothing parameters requires the solution of, and smoothing parameter estimation for, a sequence of weighted penalized least squares problems (see, e.g. Wahba, 1990; Wood, 2000; Gu 2002, for a fuller discussion, or the appendix for a different approach). Usually some identifiability constraints are applied to the fitting problem, but for clarity of presentation I will neglect weights and constraints for the moment, and return to them at the end of this section.

The basic GAM fitting problem (typically nested within a P-IRLS loop in weighted, constrained form) is therefore:

$$\text{minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta} + \sum_{i=1}^{m} \theta_i \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_i\boldsymbol{\beta} \text{ w.r.t. } \boldsymbol{\beta}. \tag{2}$$

$\mathbf{X}$ is an $n \times q$ model matrix; $\boldsymbol{\beta}$ is a parameter vector; $\mathbf{y}$ is a response vector; $\mathbf{S}_i$ is the i^{th} (positive semi-definite) penalty matrix with unknown smoothing parameter θ_i; $\mathbf{H}$ is a fixed positive semi-definite penalty matrix: it allows several enhancements to existing GAM methods, for example, the imposition of ridge penalties, the imposition of lower bounds on smoothing parameters and the employment of mixtures of pre-specified and estimated smoothing parameters.

Given the smoothing parameters this problem is easily solved, but the smoothing parameters have to be estimated. Two possible methods are Generalized Cross Validation (GCV) or minimisation of an Un-

Biased Risk Estimator (UBRE) (see Craven and Wahba, 1979. UBRE can be viewed as an approximation to AIC for many GAMs.), where the parameters are chosen to minimize:

$$V_g = \frac{n\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2}{[\mathrm{tr}(\mathbf{I} - \gamma\mathbf{A})]^2}$$

or

$$V_u = \frac{1}{n}\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 - \frac{2}{n}\sigma^2\mathrm{tr}(\mathbf{I} - \gamma\mathbf{A}) + \sigma^2$$

respectively. $\mathbf{A}$ is the influence matrix or hat matrix of the fitting problem, and depends on the smoothing parameters, while $\gamma \geq 1$ is a parameter sometimes used to enforce smoother models than would otherwise occur (see e.g. Chambers and Hastie, 1993, discussion of `smooth.spline`), σ^2 is the variance of the y_i, or in a P-IRLS context, the scale parameter.

Neither criteria is easy to minimize with respect to multiple smoothing parameters, since direct evaluation of the $tr(\mathbf{A})$ term involves at least $nq^2/2$ operations plus $O(q^3)$ operations for each new set of smoothing parameter values: hence smoothing parameter estimation by direct search of the smoothing parameter space is often too computationally demanding for routine use (and prohibitive in the full SS-ANOVA case for which $q = n$). One pragmatic, but ad hoc, approach is to simply estimate the smoothing parameters using GCV applied to each of the single smoothing parameter problems that arises at each step of a backfitting algorithm (as in SAS PROC GAM, Xiang, 2001), but the theoretical properties of this method are unclear: some problems might be expected with correlated covariates. Hastie and Tibshirani (1990, section 9.4.3) took a different approach in the context of backfit GAMs, suggesting that a computationally efficient approximation to the GCV score for the whole model could be employed. However, a useful efficient approach based on the exact GCV score for the whole model was pioneered by Gu and Wahba (1991) who alternated efficient direct searches for an ‘overall smoothing parameter’ with Newton updates of $\eta_i = \log(\theta_i)$. Their method used the special structure of the general spline smoothing problem in which they were primarily interested (and in which $\mathbf{X}$ depends on the smoothing parameters and there is no $\mathbf{H}$). Wood (2000) adapted their approach to problems with the structure considered here, but without the fixed penalty. The Wood (2000) method provided an effective means of selecting the degree of smoothness for terms in GAM models, but has two drawbacks (in part shared with the Gu and Wahba, 1991, method). Firstly, because of the way the direct search for the overall smoothing parameter works, the method does not allow users to fix some smoothing parameters and estimate others, bound smoothing parameters from below or regularize the fit with a ridge penalty (i.e. no $\mathbf{H}$ term is possible). Secondly the method is not optimally stable numerically, and this can cause problems in practical application of the method in the GAM context.

Here an alternative to the Wood (2000) method is developed, which allows the fixed penalty term, is particularly robust numerically, and can deal elegantly with rank deficiency in the model, whether it occurs over all or only part of the smoothing parameter space. The difference in numerical robustness is similar to (actually slightly greater than) the difference between using pivoted QR or SVD methods in ordinary least squares, as opposed to solving the normal equations via a Choleski decomposition (see Golub and van Loan, 1996, for example), while the ability to impose ridge penalties is of practical use in GAM contexts where the models would otherwise be practically un-identifiable.

1.1 Stable and efficient score minimization

The basic approach is to perform Newton, or failing that steepest descent, updates of the log smoothing parameters. Hence stable and computationally efficient formation of the derivatives of the scores with respect to the log smoothing parameters is of primary concern. Working with log smoothing parameters has the advantage of ensuring that the smoothing parameter estimates are positive, and is also justified heuristically by the fact that plots of GCV and UBRE functions for one dimensional smooths appear more susceptible to quadratic approximation than do equivalent plots on the original scale.

Consider the influence matrix of the problem, since this is the term in the GCV or UBRE scores which is expensive to obtain:

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{H} + \sum_{i=1}^{m} \theta_i \mathbf{S}_i\right)^{-1} \mathbf{X}^{\mathrm{T}}.$$

The first step is to form the QR decomposition of $\mathbf{X}$,

$$\mathbf{X} = \mathbf{QR},$$

where $\mathbf{Q}$ is made up of columns of an orthogonal matrix and $\mathbf{R}$ is upper triangular. For maximum stability a pivoted decomposition should actually be used here (Golub and van Loan, 1996; LAPACK provides a suitable routine), which has the consequence that the parameter vector and $\mathbf{S}_i$ matrices have to be re-ordered before proceeding, and the parameter vector and covariance matrix put back into the original ordering at the end of the estimation procedure.

Defining $\mathbf{S} = \mathbf{H} + \sum_{i=1}^{m} \theta_i \mathbf{S}_i$ and $\mathbf{B}$ as any matrix square root of $\mathbf{S}$ such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{S}$, a singular value decomposition (Golub and van Loan 1996, LAPACK version used) can be formed:

$$\begin{bmatrix} \mathbf{R} \\ \mathbf{B} \end{bmatrix} = \mathbf{UDV}^{\mathrm{T}}.$$

$\mathbf{B}$ can be obtained efficiently by pivoted Choleski decomposition (available in LINPACK, for example) or by eigen-decomposition of the symmetric matrix $\mathbf{S}$ (see e.g. Golub and van Loan, 1996). The columns of $\mathbf{U}$ are columns of an orthogonal matrix and $\mathbf{V}$ is an orthogonal matrix. $\mathbf{D}$ is the diagonal matrix of singular values, and examination of these is the most reliable way of detecting numerical rank deficiency of the fitting problem (Golub and van Loan, 1996, Watkins, 1991). In particular, at this stage any singular values that are 'too small' should be removed along with the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$. This has the effect of recasting the problem into a reduced space in which the model parameters are identifiable. 'Too small' is usually judged with reference to the largest singular value. In the work reported here singular values less than the largest singular value multiplied by the square root of the machine precision were deleted. Note that, in addition to dealing with colinearity/concurvity directly identifiable from $\mathbf{X}$, this approach also deals effectively with the difficult problem of rank deficiency that may only occur over part of the smoothing parameter space.

Now defining the submatrix $\mathbf{U}_1$ of $\mathbf{U}$ such that $\mathbf{R} = \mathbf{U}_1\mathbf{DV}^{\mathrm{T}}$, we have that $\mathbf{X} = \mathbf{QU}_1\mathbf{DV}^{\mathrm{T}}$, while

$\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{S}=\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}$. Consequently

$$\begin{aligned}\mathbf{A} &= \mathbf{Q}\mathbf{U}_1\mathbf{D}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{D}\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}} \\ &= \mathbf{Q}\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\end{aligned}$$

Hence $\mathrm{tr}(\mathbf{A})=\mathrm{tr}(\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q})=\mathrm{tr}(\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}})$, which is relatively cheap to evaluate for new trial values of $\boldsymbol{\theta}$.

Turning to the derivatives, it is convenient to write the influence matrix as $\mathbf{A}=\mathbf{X}\mathbf{G}^{-1}\mathbf{X}^{\mathrm{T}}$ where $\mathbf{G}=\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{S}=\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}$ and hence $\mathbf{G}^{-1}=\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}$. Letting $\eta_i=\log\theta_i$ we then have that

$$\frac{\partial\mathbf{G}^{-1}}{\partial\eta_i}=-\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\eta_i}\mathbf{G}^{-1}=-\theta_i\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}\mathbf{S}_i\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}$$

and so

$$\frac{\partial\mathbf{A}}{\partial\eta_i}=\mathbf{X}\frac{\partial\mathbf{G}^{-1}}{\partial\eta_i}\mathbf{X}^{\mathrm{T}}=-\theta_i\mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{S}_i\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}.$$

For the second derivatives we have

$$\frac{\partial^2\mathbf{G}^{-1}}{\partial\eta_i\partial\eta_j}=\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\eta_j}\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\eta_i}\mathbf{G}^{-1}-\mathbf{G}^{-1}\frac{\partial^2\mathbf{G}}{\partial\eta_i\partial\eta_j}\mathbf{G}^{-1}+\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\eta_i}\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\eta_j}\mathbf{G}^{-1}$$

and, of course,

$$\frac{\partial^2\mathbf{A}}{\partial\eta_i\partial\eta_j}=\mathbf{X}\frac{\partial^2\mathbf{G}^{-1}}{\partial\eta_i\partial\eta_j}\mathbf{X}^{\mathrm{T}}.$$

This becomes

$$\frac{\partial^2\mathbf{A}}{\partial\eta_i\partial\eta_j}=\theta_i\theta_j\mathbf{Q}\mathbf{U}_1\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}[\mathbf{S}_j\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}\mathbf{S}_i]^{\ddagger}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}+\delta_j^i\frac{\partial\mathbf{A}}{\partial\eta_i}$$

where $\mathbf{B}^{\ddagger}\equiv\mathbf{B}+\mathbf{B}^{\mathrm{T}}$ and $\delta_j^i=1$ if $i=j$ and zero otherwise.

Writing $\alpha=\|\mathbf{y}-\mathbf{A}\mathbf{y}\|^2$, we are in a position to obtain the derivatives needed to find the derivatives of the scores. Some rather tedious manipulation leads to the following.

Define: (i) $\mathbf{y}_1=\mathbf{U}_1^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{y}$; (ii) $\mathbf{M}_i=\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{S}_i\mathbf{V}\mathbf{D}^{-1}$ and (iii) $\mathbf{K}_i=\mathbf{M}_i\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_1$. Then

$$\mathrm{tr}\left(\frac{\partial\mathbf{A}}{\partial\eta_i}\right)=-\theta_i\mathrm{tr}(\mathbf{K}_i)$$

$$\mathrm{tr}\left(\frac{\partial^2\mathbf{A}}{\partial\eta_i\partial\eta_j}\right)=2\theta_i\theta_j\mathrm{tr}\left(\mathbf{M}_j\mathbf{K}_i\right)-\delta_j^i\theta_i\mathrm{tr}\left(\mathbf{K}_i\right)$$

while

$$\frac{\partial \alpha}{\partial \eta_i} = 2\theta_i \left[\mathbf{y}_1^{\mathrm{T}} \mathbf{M}_i \mathbf{y}_1 - \mathbf{y}_1^{\mathrm{T}} \mathbf{K}_i \mathbf{y}_1 \right]$$

and

$$\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} = 2\theta_i\theta_j \mathbf{y}_1^{\mathrm{T}} \left[\mathbf{M}_i\mathbf{K}_j + \mathbf{M}_j\mathbf{K}_i - \mathbf{M}_i\mathbf{M}_j - \mathbf{M}_j\mathbf{M}_i + \mathbf{K}_i\mathbf{M}_j \right] \mathbf{y}_1 + \delta_j^i \frac{\partial \alpha}{\partial \eta_i}$$

The above derivatives can be used to find the derivatives of V_g or V_u w.r.t. the η_i. First let $\delta = n - \gamma \mathrm{tr}(\mathbf{A})$, so that

$$V_g = \frac{n\alpha}{\delta^2} \quad \text{and} \quad V_u = \frac{1}{n}\alpha - \frac{2}{n}\delta\sigma^2 + \sigma^2.$$

Then

$$\frac{\partial V_g}{\partial \eta_i} = \frac{n}{\delta^2}\frac{\partial \alpha}{\partial \eta_i} - \frac{2n\alpha}{\delta^3}\frac{\partial \delta}{\partial \eta_i}$$

and

$$\frac{\partial^2 V_g}{\partial \eta_i \partial \eta_j} = -\frac{2n}{\delta^3}\frac{\partial \delta}{\partial \eta_j}\frac{\partial \alpha}{\partial \eta_i} + \frac{n}{\delta^2}\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} - \frac{2n}{\delta^3}\frac{\partial \alpha}{\partial \eta_j}\frac{\partial \delta}{\partial \eta_i} + \frac{6n\alpha}{\delta^4}\frac{\partial \delta}{\partial \eta_j}\frac{\partial \delta}{\partial \eta_i} - \frac{2n\alpha}{\delta^3}\frac{\partial^2 \delta}{\partial \eta_i \partial \eta_j}.$$

Similarly

$$\frac{\partial V_u}{\partial \eta_i} = \frac{1}{n}\frac{\partial \alpha}{\partial \eta_i} - 2\frac{\partial \delta}{\partial \eta_i}\frac{\sigma^2}{n}$$

and

$$\frac{\partial^2 V_u}{\partial \eta_i \partial \eta_j} = \frac{1}{n}\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} - 2\frac{\partial^2 \delta}{\partial \eta_i \partial \eta_j}\frac{\sigma^2}{n}.$$

These derivatives can be obtained quite efficiently for each new $\boldsymbol{\theta}$, so that Newton's method can be used to find the optimum $\boldsymbol{\theta}$ fairly efficiently. If the score is not locally concave (i.e. if some eigenvalues of the Hessian of the score with respect to the smoothing parameters are not positive) then steepest descent steps can be substituted for Newton steps, and in either case, if the direction fails to decrease the score then repeated step length halving can be applied until either the score decreases, or the direction is deemed not to lead to decrease. Some care is required to get good starting values for the θ_i. In iteratively re-weighted least squares contexts it is usual to carry forward the previous $\boldsymbol{\theta}$ estimate as the starting $\boldsymbol{\theta}$ at the next iteration: if this is done then it is important to reset to a default starting value any θ_i for which $\partial V./\partial \eta_i$ is of too small a magnitude, to avoid becoming stuck on flat portions of the score. In a similar vein, optimal η_i at ±∞ can be problematic, since false early convergence of the Newton method is quite easily triggered in such cases: a sensible precaution is to check that a fairly large step for each smoothing parameter (in the direction suggested by the score gradient) would not improve the score, once an apparent optimum has been achieved.

The best fit GAM parameters are:

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{D}^{-1}\mathbf{y}_1$$

with corresponding estimated Bayesian parameter covariance matrix (see e.g. Wood 2000)

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}\hat{\sigma}^2 \;\; \text{where} \;\; \hat{\sigma}^2 = \alpha/(n - \text{tr}(\mathbf{A})).$$

1.2 The more general problem

In the previous section weights and constraints were neglected for clarity. In general the problems of interest will in fact be of the form:

$$\text{minimize } \|\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta} + \sum_{i=1}^{m} \theta_i \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_i\boldsymbol{\beta} \;\; \text{w.r.t. } \boldsymbol{\beta} \;\; \text{subject to} \;\; \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$$

where $\mathbf{W}$ is typically a square root of the inverse of the covariance matrix of $\mathbf{y}$ (i.e. $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{V}_y^{-1}$), or the iterative weights in a GLM weighted least squares iteration. The linear constraints typically impose identifiability constraints of some sort: in the GAM context usually that each smooth should sum to zero over its covariate values.

The constraints can be found by forming the QR decomposition of $\mathbf{C}^{\mathrm{T}}$. The final $q - c$ columns of the resulting orthogonal factor, $\mathbf{Q}^*$, (where c is number of constraints) gives the null space of $\mathbf{C}$: $\mathbf{Z}$, say. Writing $\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\beta}_z$ ensures that the constraints are met. Letting $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}\mathbf{Z}$, $\tilde{\mathbf{H}} = \mathbf{Z}^{\mathrm{T}}\mathbf{H}\mathbf{Z}$ and $\tilde{\mathbf{S}}_i = \mathbf{Z}^{\mathrm{T}}\mathbf{S}_i\mathbf{Z}$ the problem becomes

$$\text{minimize } \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}_z\|^2 + \boldsymbol{\beta}_z^{\mathrm{T}}\tilde{\mathbf{H}}\boldsymbol{\beta}_z + \sum_{i=1}^{m} \theta_i \boldsymbol{\beta}_z^{\mathrm{T}}\tilde{\mathbf{S}}_i\boldsymbol{\beta}_z \;\; \text{w.r.t. } \boldsymbol{\beta}_z$$

which is in exactly the form required for the method described in section 1.1.

Transforming the parameters and their variances back to the original parameter space after fitting is straightforward:

$$\hat{\boldsymbol{\beta}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_z \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Z}\mathbf{V}_{\boldsymbol{\beta}_z}\mathbf{Z}^{\mathrm{T}}$$

while the degrees of freedom per parameter in the unconstrained space are given by the leading diagonal of

$$\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{X}/\hat{\sigma}^2.$$

1.3 Convergence issues

This section discusses some "convergence" issues relating to both the smoothing parameter estimation procedure itself and the GAM fitting methods within which it is usually embedded. Firstly consider the convergence conditions for the smoothing parameter estimation algorithm. Writing V for the GCV or UBRE score functions and $\hat{\boldsymbol{\theta}}_k$ for the estimated smoothing parameter values at iteration k, convergence is deemed to have occurred if

$$V(\hat{\boldsymbol{\theta}}_{k-1}) - V(\hat{\boldsymbol{\theta}}_k) \le \epsilon_0(1 + |V(\hat{\boldsymbol{\theta}}_k)|) \;\; \text{and} \;\; \left(\frac{1}{m} \sum_i \left. \frac{\partial V}{\partial \theta_i} \right|_{\hat{\boldsymbol{\theta}}_k}^2 \right)^{1/2} \le \epsilon_0^{1/3}(1 + |V(\hat{\boldsymbol{\theta}}_k)|)$$

(where ϵ_0 is a small constant) or (as a pragmatic 'fail-safe') there is no decrease in V along the steepest descent direction after a specified number of step-halvings. The former conditions are taken from Gill *et al.* (1981). The conditions do not involve the values of the smoothing parameter estimates directly, since these may legitimately not converge if a 'true' value is at infinity.

When $\gamma = 1$ the asymptotic mean square error minimization properties of V_g and V_u (see e.g. Gu, 2002 for a clear exposition) mean that as the sample size tends to infinity we expect the scores to have global minima increasingly close to what is optimal in MSE terms. However, the scores may in practice sometimes display local minima, and it is not possible to guarantee that these will always be avoided by the optimization method, although simple precautions, such as examining transects through the score functions can help to identify such problems if they do occur.

In general contexts in which the penalized likelihood is maximized iteratively, convergence is not guaranteed theoretically. Since iteratively re-weighted least squares is not guaranteed to converge in all circumstances then neither is a penalized version, particularly when smoothing parameters must also be estimated. In fact the smoothing parameter estimation makes a theoretical treatment of convergence a difficult and currently open issue, at least when the computationally efficient 'performance iteration' (Gu, 1992) is used, so that smoothing parameters are estimated by GCV or UBRE for each penalized regression problem generated by the iteratively re-weighted least squares procedure. Gu (2002, section 5.2) provides a fuller discussion of the issues. However it is worth noting that this paper was motivated by investigation of a number of practical convergence problems, which turned out to relate not to non-existence of a stationary point in the performance iteration, but rather to the stability issues addressed by the new method.

The final convergence issue concerns convergence of the model parameter estimators, $\hat{\boldsymbol{\beta}}$ as sample size $n \to \infty$. The presence of the penalty terms in (2) ensures that these will generally be biased, however the fixed dimension of $\boldsymbol{\beta}$ ensures (slightly artificially) that consistency is straightforward to demonstrate. In the Gaussian case it helps to re-parameterize a little so that (2) becomes

$$\text{minimize } \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta} + \sum_{i=1}^{m} \theta_i \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_i\boldsymbol{\beta} \quad \text{w.r.t. } \boldsymbol{\beta}.$$

If $\mathbf{H} \to 0$ and $\theta_i \to 0$ as $n \to \infty$ for all θ_i (except possibly those relating to penalties on subsets of $\boldsymbol{\beta}$ with true values in the null space of the penalty concerned) then consistency of $\hat{\boldsymbol{\beta}}$ follows from the fact that the fitting objective becomes entirely dominated by its least squares component as $n \to \infty$. Similarly in the generalized case the P-IRLS iterations maximize a penalized likelihood and again if $\mathbf{H} \to 0$ and $\theta_i \to 0$ as $n \to \infty$ the likelihood part dominates as sample size increases, yielding consistency of $\hat{\boldsymbol{\beta}}$ from the consistency of maximum likelihood estimation. $\mathbf{H} \to 0$ is entirely in the hands of the modeller. As $n \to \infty$ the shapes of V_g and V_u become dominated by $\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2/n$ which is minimized when $\theta_i \to 0$ (for all θ_i excepting those excepted previously) as required for consistency of $\hat{\boldsymbol{\beta}}$. Of course if the dimension of $\boldsymbol{\beta}$ increases with n then a much more careful argument is required.

2 Comparison of the new and older methods

The proposed method improves in several ways on the method of Wood (2000), which was in turn developed from the method of Gu and Wahba (1991). In this (slightly technical) section the new method is compared with the Wood (2000) method.

To understand the differences between the methods it is helpful to consider the expression for the influence matrix, $\mathbf{A}$ used in the two methods and the calculation of the key term $\mathrm{tr}(\mathbf{A})$. The Wood (2000) method used the expression

$$\mathbf{A} = \rho\mathbf{Q}\left(\mathbf{I}\rho + \sum_i \theta_i \mathbf{R}^{-\mathrm{T}}\mathbf{S}_i\mathbf{R}^{-1}\right)^{-1}\mathbf{Q}^{\mathrm{T}}$$

where ρ is an extra 'overall' smoothing parameter included in order to allow use of existing single smoothing parameter methods as part of smoothing parameter estimation (there is no $\mathbf{H}$ term possible for this method). The method then forms the decomposition

$$\sum_i \theta_i \mathbf{R}^{-\mathrm{T}}\mathbf{S}_i\mathbf{R}^{-1} = \mathbf{P}\mathbf{T}\mathbf{P}^{\mathrm{T}}$$

where $\mathbf{P}$ is orthogonal and $\mathbf{T}$ tri-diagonal, so that

$$\mathbf{A} = \rho\mathbf{Q}\mathbf{P}\left(\mathbf{I}\rho + \mathbf{T}\right)^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}$$

There are three potential sources of numerical instability here.

1. $\mathbf{X}$ must be of full rank or $\mathbf{R}$ will be rank deficient and the terms involving $\mathbf{R}^{-1}$ will not be identifiable. Near rank deficiency of $\mathbf{X}$ (for example as a result of colinearity or concurvity) will also cause numerical difficulties through the resulting near rank deficiency (high condition number) of $\mathbf{R}$.

2. Since the $\mathbf{S}_i$ are usually rank deficient, it is only the $\mathbf{I}\rho$ term which gives $\mathbf{I}\rho + \mathbf{T}$ the necessary full rank. Hence as $\rho \to 0$ or any $\theta_i \to \infty$, $(\mathbf{I}\rho + \mathbf{T})$ tends to singularity (condition number tends to infinity) and $(\mathbf{I}\rho + \mathbf{T})^{-1}$ becomes numerically problematic.

3. The Wood (2000) method uses a Choleski decomposition of $\mathbf{I}\rho + \mathbf{T}$ in order to solve for terms involving $(\mathbf{I}\rho+\mathbf{T})^{-1}$, but the Choleski method will not be reliable if the condition number of $\mathbf{I}\rho+\mathbf{T}$ is greater than approximately $1/\sqrt{\epsilon}$ where ϵ is the machine precision (see Golub and van Loan, 1996, section 6.3.4, page 240). There is clearly scope for this to happen either through near rank deficiency of $\mathbf{X}/\mathbf{R}$ or if smoothing parameters θ_i become very large.

By careful scaling of $\mathbf{X}$ and guarding of the range of allowed smoothing parameters (as in R package mgcv) it is possible to avoid these issues leading to actual numerical problems for most 'well behaved' models, but they can not be eliminated altogether, particularly if the appropriate value for some smoothing parameters → ∞, or if the model really is close to unidentifiable, at least for some values of the smoothing parameters. (Note that pivoting the Choleski decomposition is problematic for this method if the efficiency benefits accruing from $\mathbf{T}$ being tri-diagonal are to be maintained.)

The reader will notice that none of these problems occur with the new method for which the expression for the influence matrix is $\mathbf{A} = \mathbf{Q}\mathbf{U}_1\mathbf{U}_1^{\rm T}\mathbf{Q}^{\rm T}$. Indeed it is difficult to see how this expression could be improved on with regard to numerical stability. The only potential cause of difficulty is if rank deficiency of $\mathbf{X}$ is under-estimated so that $\mathbf{U}$ is insufficiently truncated. In this case the fitted values are unaffected, but the trace of the influence matrix will be over-estimated by an amount bounded above by the underestimation of the rank of the problem. Since SVD is the most reliable way of estimating problem rank, it is not easy to see how this minor potential problem can be avoided, and in any case the new method is doing much better than the Wood (2000) method, which simply assumed $\mathbf{X}$ to be of full rank and could fail badly if it was not.

The issues are very similar when it comes to comparison of the derivative calculations. For example, the Wood (2000) method uses the expression

$$\frac{\partial^2 \mathbf{A}}{\partial \eta_i \partial \eta_j} = \rho\theta_i\theta_j \left[\mathbf{Q}\mathbf{P}(\mathbf{I}\rho+\mathbf{T})^{-1}\mathbf{P}^{\rm T}\mathbf{R}^{-\rm T}\mathbf{S}_i\mathbf{R}^{-1}\mathbf{P}(\mathbf{I}\rho+\mathbf{T})^{-1}\mathbf{P}^{\rm T}\mathbf{R}^{-\rm T}\mathbf{S}_j\mathbf{R}^{-1}\mathbf{P}(\mathbf{I}\rho+\mathbf{T})^{-1}\mathbf{P}^{\rm T}\mathbf{Q}^{\rm T}\right]^{\ddagger} + \delta^i_j \frac{\partial \mathbf{A}}{\partial \eta_i}.$$

Clearly the same issues apply to this expression as applied to the expression for $\mathbf{A}$, only more so, given that the potentially problematic terms now recur. In comparison the expression used in the new method is substantially better: systems involving $\mathbf{D}^{-1}$ do have to be solved, but $\mathbf{D}$ will have been truncated if rank deficiency was detected so that its condition number will always be safely bounded above. Again it is worth emphasizing here that the method used to detect numerical rank deficiency is the best known (Golub and van Loan, 1996), and again it is difficult to see how this calculation could be further improved in terms of numerical stability.

So, the new method eliminates the potential sources of poor numerical performance in the Wood (2000) method, in addition to extending the class of problems that can be addressed via the introduction of $\mathbf{H}$ and removal of the full rank condition on $\mathbf{X}$. Furthermore there are no other sources of numerical difficulty evident in the new method, and while it is always possible that more extensive experience with will reveal unthought of problems, I am fairly confident that further major improvements in this direction are unlikely to be possible (several alternative methods of 'intermediate stability' were discarded en route to obtaining the method reported here).

The new method has two possible weaknesses relative to Wood (2000). Firstly, although the leading order QR decomposition of $\mathbf{X}$ is the same for both methods, the sub leading order calculations are slightly more costly for the new method. However this issue is complicated in poorly conditioned cases by the fact that the more accurate calculations can lead to faster convergence. The second potential weakness is that the new method does not include the global search for an overall smoothing parameter that was a feature of the Wood (2000) method. In theory such a search might help the algorithm to escape local minima in the GCV score, but in practice it seems to lead to local minima at infinite smoothing parameters as often as it helps to escape from local minima. Hence careful choice of initial values and step length guarding during optimization are usually more helpful at avoiding spurious local minima in practice.

3 Performance of the new method

The development of the new method arose from attempts to diagnose the cause of convergence and fitting problems in a number of rather complex practical modelling problems. Typically such failures have more than one cause, while the models themselves are too complicated to serve as good illustrations of the basic types of convergence problems. In this section I therefore illustrate the effectiveness of the new method using synthetic data designed to simply exemplify the kinds of problem that can cause difficulty in real modelling situations. In each case I compared the performance of Gu and Wahba's (1991) method as implemented in the R package `gss` (version 0.8-2) and Wood's (2000) method as implemented in R package `mgcv` (version 0.8-7) to the new method. All tests were carried out on a Pentium IV 1.7Ghz PC running Windows XP and R 1.7.0 (R Core development team, 2003). For the new method the convergence tolerance ϵ_0 has set to 1×10^{-6}.

Another possibility for model and multiple smoothing parameter estimation is to write the model as a mixed model and estimate using REML in the additive model case and Penalized Quasi Likelihood (PQL) in the generalized additive model case. The appendix gives details of how this can be done in a straightforward manner suitable for use with the R routines `lme` (`nlme` version 3.1-39) or `glmmPQL` (Venables and Ripley, 2002), and comparisons were also made with this mixed model approach. For both the mixed model and penalized regression based GAMs the number of parameters to use to represent each term must be chosen in advance, although the actual effective degrees of freedom will be controlled by the smoothing parameters which are estimated automatically. Provided that an overly restrictive choice is not made, the decision should not have much effect on the final estimates, but it will have some effect and the choice is to some extent arbitrary.

It should be stressed that several of the comparisons have been made on deliberately extreme cases — for most models of most datasets the alternative methods are perfectly adequate, but where the data/model combination shows some features of the examples given below the new method is expected to perform better (and certainly to be 'safer'). The next section provides one practical example for which the new method gives a useful improvement on the alternatives.

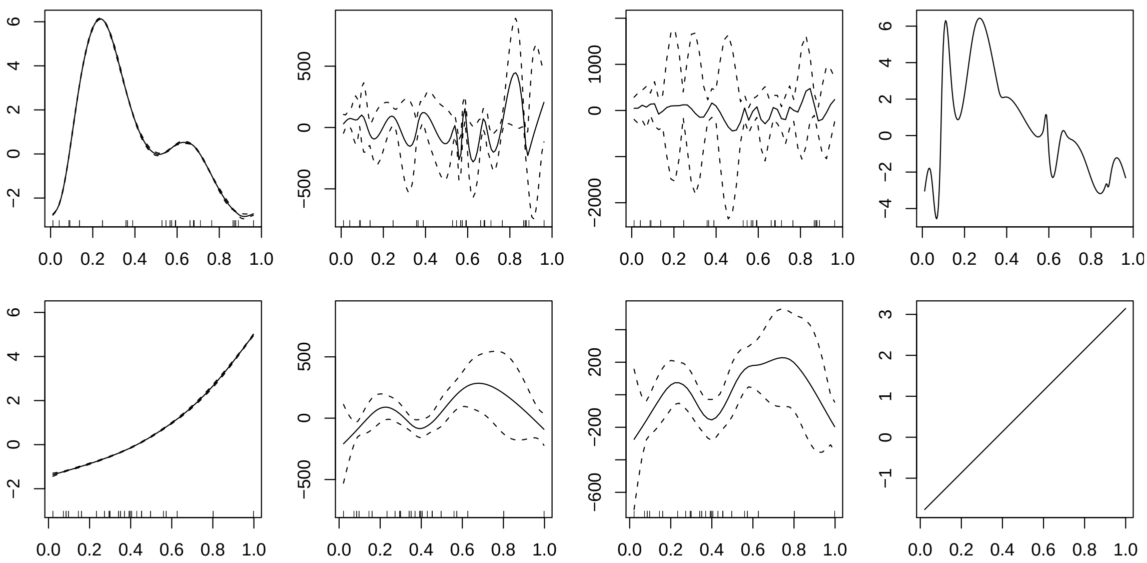

Figure 1: Comparison of reconstructions of f_1 and f_2 from section 3.1 using different GAM methods with integrated smoothing parameter selection. The upper row are reconstructions of $f_1(x)$: for each plot the horizontal axes is x and the vertical axis $\hat{f}_1$. The lower row are reconstructions of $f_2(z)$: the horizontal and vertical axes are z and $\hat{f}_2$ respectively. The first column (left most) shows reconstructions using the new method — these are excellent reconstructions of the original functions. The second column contains reconstructions using the method of Wood (2000). The third column uses the method of Gu and Wahba (1991), while the fourth column uses REML via `lme`. The methods have been used on exactly the same data, which is deliberately constructed to have the potential to cause numerical difficulty. In each panel the solid line is the estimate, and the dashed lines are at plus and minus two standard errors, except for the final column where only point estimates are shown. The rug plots show the covariate values.

3.1 Almost co-incident covariates

Covariate values that are very close together can cause near rank deficiency for spline based GAMs since they can lead to almost identical basis functions occuring in the GAM. As a fairly extreme example of the effect, covariates x and z were simulated, such that x_i were i.i.d. $U(0,1)$ for $i = 1, \ldots, 25$ and $x_{i+25} = x_i + w_i$ for $i = 1, \ldots, 25$ where the w_i are i.i.d. $U(0, \epsilon)$. The z_i were generated independently in the same way. So each set of covariates consist of 25 pairs of covariate values, each pair separated by no more than ϵ.

Data were simulated from,

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$, $f_1(x) = x^{11}[10(1-x)]^6 + 10(10x)^3(1-x)^{10} - 1.396$ and $f_2(z) = e^{2z} - 3.75887$.

The data were fit by a two term GAM, using the R SS-ANOVA package `gss`, the R GAM package `mgcv`, the R GAM package `mgcv` modified by replacing the call to Wood's (2000) method in `gam.fit()` by a call to an implementation the new method and finally by REML using `lme` as outlined in the appendix. For a range of σ values the estimates of f_1 and f_2 are in very close agreement for all the methods, provided $\epsilon > 10^{-6}$, but for smaller ϵ values the Gu and Wahba (1991) and Wood (2000) methods start to display numerical problems. As ϵ is reduced the REML approach generates increasingly frequent numerical warnings or failures from `lme` and occasionally very poor or nonsensical estimates without warning: however it also continues to produce sensible estimates for a substantial proportion of replicates. By contrast the new method performs well all the way down to $\epsilon = 0$. Figure 1 shows a comparison of results when $\sigma = 0.01$ and $\epsilon = 10^{-6}$. All computations were performed in double precision, with the machine precision being $\approx 2.2 \times 10^{-16}$. For the new method, Wood (2000) method and Gu and Wahba (1991) methods GCV was used. In the cases using `mgcv` style GAMs the model terms were each allowed a maximum of 24 degrees of freedom.

The reconstructions shown in the figures are unusually bad for $\epsilon = 1 \times 10^{-6}$ for all methods except the new one, however they are typical of what is produced for $\epsilon \leq 1 \times 10^{-7}$ for the Wood (2000) and Gu and

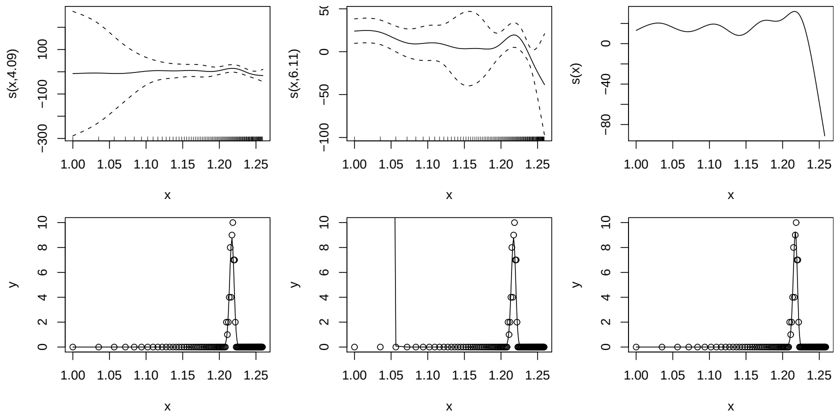

Figure 2: Comparison of model estimates and fits for the example in section 3.2. The upper row shows estimates of the linear predictor (the smooth): the rug plot shows the covariate values, the solid curves are the estimates of the smooth and the dashed curves (where present) give 95% ‘confidence limits’. The figures given in the vertical axis caption are estimated degrees of freedom, when these are available. The lower row shows corresponding predictions on the response scale: symbols are data, continuous lines are model predictions. The left column shows the results using the new method, with a small ridge penalty on the GAM fit. The middle column shows the equivalent using the method of Wood (2000) for which no ridge penalty is possible. The right column shows a fit using PQL.

Wahba (1991) methods (when the methods produce results rather than failing). REML is less consistent in that very good fits can be achieved for a substantial (albeit declining) proportion of replicates as $\epsilon \to 0$. In no replicate tried did the new method fail to perform well. Of the other two methods Gu and Wahba’s (1991) method tended to either fail or produce very wide confidence intervals, which at least means that the user is unlikely to be misled. The Wood (2000) method performed worst, in that it produced obviously poor results more often than Gu and Wahba’s (1991) method, and in such cases invariably produced confidence intervals that were too narrow.

3.2 A Poisson model with a log link and too many zeros

In this example the single covariate x took the values $1^{1/20}, 2^{1/20}, 3^{1/20}, \ldots, 100^{1/20}$. The expectation of the response, y, was set to 0 for all except y_{45} to y_{50} where the expectations were given by the integers 1 to 6, and y_{51} to y_{55} where the expectations were set to 6. The actual data were then simulated from Poisson distributions of the appropriate means. The purpose of these data is to give a clear example of the problems that can occur when using a log link and a flexible model for data with large areas of zeroes. As before, the data were fitted using the methods of Wood (2000), Gu and Wahba (1991), the new method and via Penalized Quasi Likelihood with the model treated as mixed model. The new method employed a small ridge penalty in the GAM fitting process (ridge parameter 10^{-3}, i.e. $\mathbf{H} = 10^{-3}\mathbf{I}$). In each case the model was:

$$\log(\mu_i) = f(x_i)$$

where $\mu_i \equiv E(y_i)$, $y_i \sim$ Poisson and f is a smooth function. For the Wood (2000), PQL and new methods f was represented by a rank 10 (penalized) thin plate regression spline (Wood, 2003). UBRE was used for smoothing parameter estimation in all cases except PQL, of course. Using the Wood (2000) method the GAM fitting iterations failed to converge, although an estimate was produced. The Gu and Wahba (1991) method failed altogether on this example. By contrast the new method with a small ridge penalty produced a quite reasonable fit on the response scale and PQL produced a good fit without difficulty.

This example illustrates the point that although the model is not formally identifiable, regularization using a ridge penalty does offer a pragmatic way of dealing with indeterminacy, if interest lies primarily in prediction on the response scale.

3.3 Logistic Regression

This example is another case in which indeterminacy of the model causes problems. In this case 500 covariates x_i, z_i were simulated on the unit square. A response variable was created which was zero for all except those points for which $x_i > 0.9$ and $z_i > 0.9$, for which the response was set to 1. These

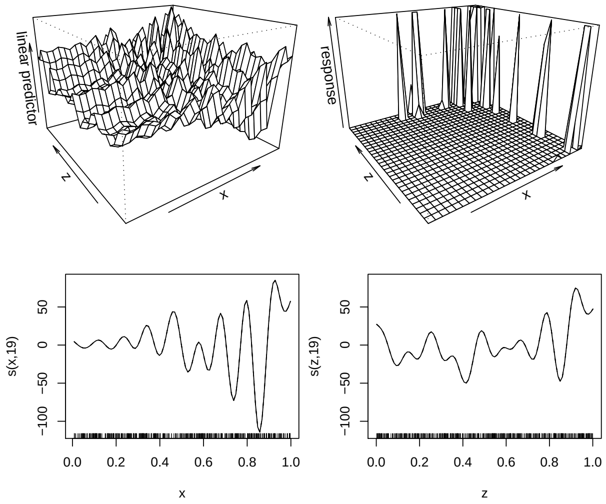

Figure 3: Logistic regression fit to example in section 3.3 using the Wood (2000) method. The top left panel is the model fit on the scale of the linear predictor, and top right panel is the fitted model on the response scale. The lower two panels show the estimates of the model terms.

'data' were modelled using a logistic regression (binomial errors and logit link.) The linear predictor was given by the sum of two univariate smooth functions of x_i and z_i respectively. The smooth terms were represented by rank 20 thin plate regression splines (Wood, 2003) in the Wood (2000), new method and PQL examples.

Gu and Wahba's (1991) method failed on this example, while Wood's (2000) method 'converged' to obvious nonsense (see figure 3). PQL failed to converge. Again using a small ridge penalty with the new method gave a satisfactory fit on the response scale (figure 4). In all cases GCV was used as the fitting criteria (although UBRE was also tried with Gu and Wahba's method in an attempt to get results for comparison).

Again the purpose of this example is not to encourage the fitting of GAMs to inappropriate data, but rather to illustrate how difficult data can cause problems if methods are not carefully stabilized. While

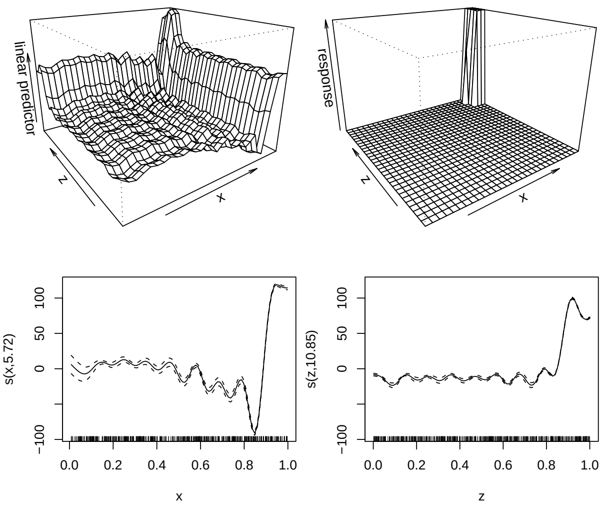

Figure 4: Logistic regression fit to example in section 3.3 using the new method and a ridge parameter of 10^{-9} (i.e. $\mathbf{H} = 10^{-9}\mathbf{I}$). The top left panel is the model fit on the scale of the linear predictor, and top right panel is the fitted model on the response scale. The lower two panels show the estimates of the model terms.

it is to be hoped that users of GAMs would not apply them to data as extreme as this example, the type of indeterminacy illustrated has the potential to cause problems in less extreme situations, if there are other contributing problems.

3.4 Comparisons on a 'well behaved' example

In the interests of consistency checking of the method, some comparisons were also performed on a simulated example taken from Wahba (1990) which does not have obvious numerical difficulties. 300 independent values for each of covariates x_1 to x_4 were simulated from $U(0, 1)$. Response data y_i were simulated from the model:

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \epsilon_i$$

where the ϵ_i are i.i.d. $N(0, 2^2)$. $f_1(x) = 2\sin(\pi x)$, $f_2(x) = e^{2x} - 3.75887$, $f_3(x) = x^{11}[10(1-x)]^6 + 10(10x)^3(1-x)^{10} - 1.396$ and $f_4(x) = 0$. 500 replicate data sets were produced, with new covariates and responses simulated for each replicate. The model was estimated using the GAM approach implemented in `mgcv` in which each smooth is represented by a rank 10 thin plate regression spline (Wood, 2003), using the Wood (2000) method and the new method. The model was also estimated using the SS-ANOVA approach and Gu and Wahba's (1991) method as implemented in `gss`. GCV was used in these three cases. As a final comparison the model estimated by `mgcv` and the new method was also estimated as a mixed model by REML using `lme` (see appendix for technical details).

The root mean square error in reconstructing the true simulated $E(y_i)$ was assessed for each method, and for the comparison of the Wood (2000) method and the new method, the mean absolute difference in fitted values between the fitted values was also assessed. In 75 of the 500 replicates the new method and Wood's (2000) method differed in mean absolute fitted value by more than 10^{-3}. Five of these 75 replicates were investigated in detail and in each case it appears that the GCV score is quite uninformative, in the sense of being almost flat with respect to some smoothing parameters in the vicinity of its minimum. However, in only 12 of the 75 cases in which the two methods differed was the Wood (2000) fit closer to the truth than the new method, and in addition the Wood (2000) method only improved on the new method by small amounts relative to the improvements of the new method on the Wood (2000) method. In the substantial majority of replicates the fits were essentially indistinguishable.

The SS-ANOVA approach is faced with a much more difficult task than the penalized regression spline based methods, simply because it relies *entirely* on GCV to decide how smooth the model terms should be, whilst the penalized regression spline method has effectively restricted the amount of flexibility allowed by using a low rank representation of the model. Hence improvements by the new method over the SS-ANOVA approach are not based on an entirely fair comparison. The results of this comparison are included only because the new method being worse than the SS-ANOVA approach for this example would clearly indicate a serious deficiency in the new method.

The REML method failed reporting numerical difficulties in 18 of the 500 replicates. Further investigation showed that this problem increased in frequency if the sample size was increased and reduced in frequency at smaller sample sizes. In the successful replicates REML achieved a modest but consistent improvement on the GCV used by the new method. The difference between REML and GCV can be reduced but not altogether eliminated by increasing γ in the GCV score to around 1.2, suggesting that part of the performance difference may result from undersmoothing by GCV.

Figure 5 shows the results of the comparisons, indicating that the new method tends to perform better than the older GCV methods. If the differences in fit are mostly caused by flatness of the GCV score, then it might be expected that the greater reliability of the derivatives of the new more stable method, would result in improved results in comparison to the less stable methods. However, although likely, it would be very difficult to *prove* that this explanation is correct.

Figures 6 and 7 illustrate the ability of the GAMs estimated by the new method to recapture the true functions used in the simulations. Figure 6 shows 5 estimates around the median of the mean square error distribution, while figure 7 shows 5 worse estimates, around the 90th percentile of the mean square error distribution.

Timing comparisons between the new method, the Wood (2000) method and the REML method

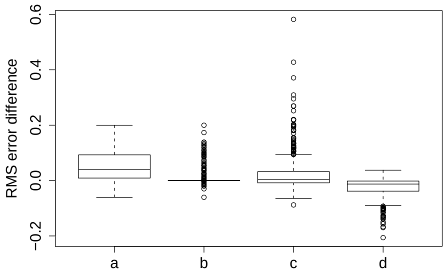

Figure 5: Differences in root mean square reconstruction error between competing methods and the new method for the example in section 3.4. Positive values indicate that the new method achieved a better reconstruction than the competing method. The mean RMS error for the new method was approximately 0.50.(a) shows the distribution of differences between the new method and the Wood (2000) method for the 75 cases in which the difference in fit was substantial. (b) is the equivalent box plot for all replicates, showing that in a substantial majority of cases the difference in fit was very small. (c) Shows the difference between the SS-ANOVA method and the new method: the apparently poorer performance here is the result of a tendancy to over-fit, brought about by the fact that the SS-ANOVA method is relying entirely on GCV for smoothing parameter selection, while the other approaches restrict model complexity a priori. (d) Shows the difference between the new method and REML estimation using `lme`, excluding the 18 cases in which this method failed altogether.

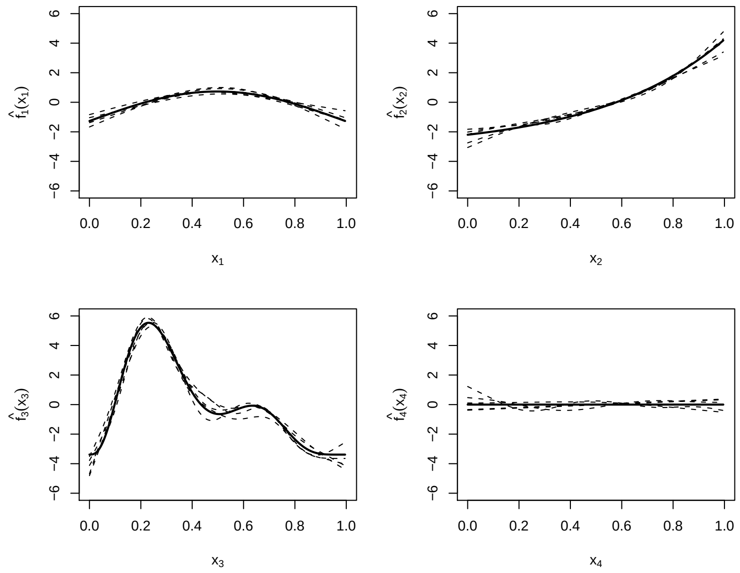

Figure 6: The estimated component functions of the 5 fits surrounding the median of the RMS reconstruction error distribution for the example from section 3.4 estimated by the new method. Heavy lines are the truth, dashed lines the reconstructions.

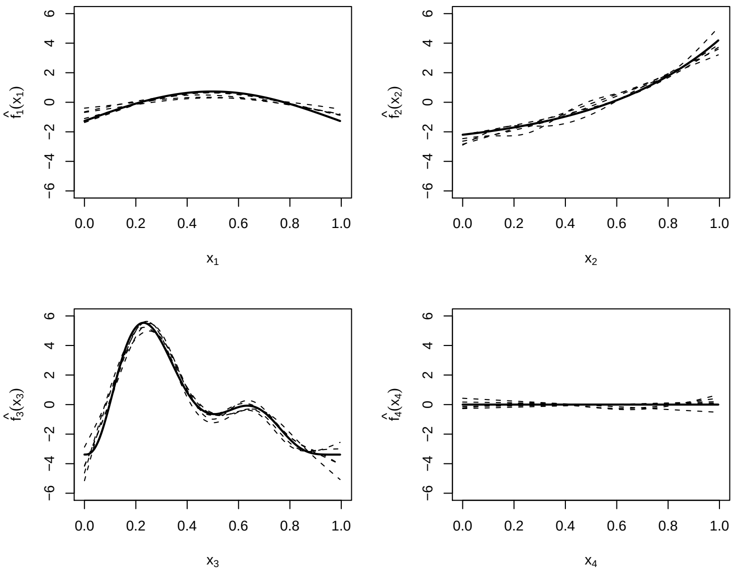

Figure 7: The estimated component functions of the 5 fits surrounding the 90th percentile of the RMS reconstruction error distribution for the example from section 3.4, estimated by the new method. Heavy lines are the truth, dashed lines the reconstructions.

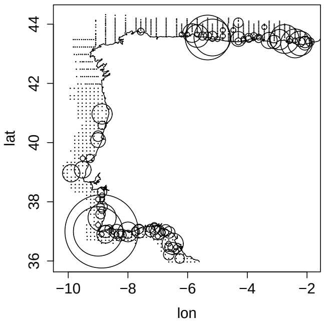

Figure 8: Stage I-III sardine eggs surveyed off the Iberian peninsula in 1997. The circle areas are proportional to the number of eggs found.

were also produced from this simulation study (using the computational setup described at the start of section 4). The per replicate timings were 0.11 seconds for the Wood (2000) method, 0.15 seconds for the new method and 2.24 seconds for the REML method. These timings exclude 0.27 seconds per replicate for setting up the model using thin plate regression splines. The SS-ANOVA based models required on average 7.43 seconds per replicate to set up and estimate (these models being much more parameter rich than the others).

4 A practical example: Sardine eggs off the Iberian peninsula

Fish stock management is a pressing problem worldwide and particularly so in the heavily overexploited coastal waters of Europe. Management is complicated by the difficulty of obtaining reliable estimates of abundance: data gathered from commercial fisheries operations are subject to a number of difficult biasing factors, while attempting to survey adult fish directly by fishing for them is subject to the obvious difficulty in converting catches to abundance estimates when fish are avoiding the fishing gear. One way of circumventing the problems with direct estimation of adult abundance is to count fish eggs instead, and then to convert from the abundance of eggs to the abundance (or more likely mass) of adults required to produce this egg abundance. The advantage of this approach is that eggs are comparatively easy to sample — they do not actively avoid the sampling gear.

However, egg data can be difficult to model. Typically the geographical distribution of eggs is not well known in advance in any given year, and this tends to mean that a well designed survey will yield a very high proportion of zero egg counts, at least in some years. The difficulties are reasonably well exemplified by Sardine egg survey data from the Iberian peninsula in 1997. Here I consider the total egg count in the first 3 identifiable egg stages at each of 888 survey stations. Egg counts were obtained by drawing a fine meshed net through the water column, from a pre-determined depth below the surface. For the analysis here, I use temperature, T_i, depth, d_i, longitude o_i and latitude, a_i, as covariates, and treat the egg counts, y_i, as Poisson distributed with mean μ_i. One model of interest is then:

$$\log(\mu_i) = f_1(a_i, o_i) + f_2(d_i) + f_3(T_i)$$

where the f_j are smooth functions. The data are shown in figure 6. Note that in this survey there are only 91 stations with a non-zero egg count. In addition, the arrangement of covariates is not easy, with many depth measurements very close together, despite a wide depth range, and a fairly unhelpful configuration of points in the longitude — latitude plane.

Attempts were made to estimate the model, including smoothing parameter estimation by the new method, the Wood (2000) method and penalized Quasi- Likelihood, in each case representing the smooth functions using a rank 50 thin plate regression spline for f_1 and rank 10 thin plate regression splines for f_2 and f_3. I also attempted to estimate the model using the method of Gu and Wahba (1991). UBRE was the estimation criterion for all methods except PQL.

The Wood (2000) method failed to converge. In this case, the failure appears to result from a loss of numerical stability in the solution of the underlying penalized least squares problem, as the weights and pseudodata of the P-IRLS become progressively more extreme. Similarly the Gu and Wahba (1991) method failed with an error message relating to a loss of numerical rank in the underlying spline fitting method. The mixed model/ PQL method fared no better, also failing with an error message relating to singularity. In this case simply substituting the new algorithm in place of the Wood (2000) algorithm

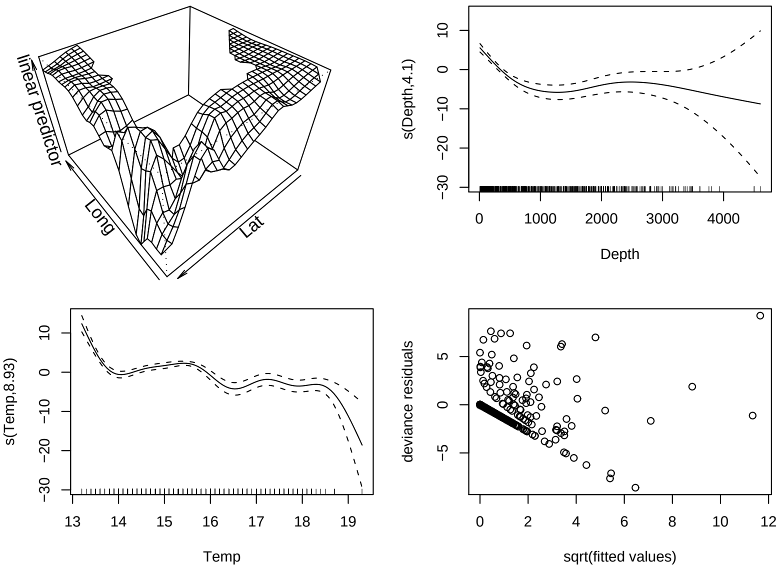

Figure 9: The sardine egg model term estimates and residual plot with the fit performed by the new method (without the need for a ridge penalty, in this case). The methods of Wood (2000), Gu and Wahba (1991) and penalized Quasi Likelihood all failed on this example. The top left plot shows the estimated smooth of longitude and latitude (only in the immediate vicinity of data). The top right panel and lower left panels are the estimated smooths of depth and temperature, respectively. The lower right panel is a residual plot, indicating some over-dispersion, relative to Poisson.

leads to convergence, without the need for extra regularization, although it is to be expected that in many similar cases some regularization would also be required.

Figure 7 shows the estimated effects and a residual plot for the fit obtained by the new method. Clearly in this case there is evidence for over-dispersion, requiring further modelling work to resolve, but since the purpose of this paper is not the analysis of these data, this will not be reported here.

Although not strictly comparable, since the smoothing parameters are not estimated, I also tried fitting the model using Hastie and Tibshirani's (1990) method as implemented in `gam()` in S-PLUS. In this case a loess smooth was employed to estimate the multi-dimensional smooth f_1, and splines for the other terms. Span (0.08) and degrees of freedom (9) were selected to give broadly similar flexibility to the models fitted by the other methods. With convergence tolerances set to be similar to those used for the other methods, convergence did not occur, unless the span of the loess smooth was increased so that the model was substantially less flexible than the models fitted by the alternative approaches. 'Convergence' did occur if the convergence tolerances were increased substantially (although whether this really constitutes 'convergence' is open to debate).

5 Discussion

The methodology developed in this paper offers some substantial advantages over existing GAM methods. The first is enhanced numerical stability of the basic penalized least squares method. By making use entirely of orthogonal matrix factorizations the method offers optimal numerical stability, while also allowing rank deficiency of the fitting problem to be reliably identified and effectively dealt with. In this respect the difference between the new method and the previous GCV methods is similar to the difference between solving ordinary least squares problems by pivoted QR or singular value decomposition methods and solving such problems by Choleski factorization based solution of the normal equations (see e.g. Golub and Van Loan, 1996). Note also that the structure of the method allows rank deficiency to be dealt with effectively even when it occurs only over a portion of the smoothing parameter space, or when the degree of numerical rank deficiency depends on the values of the smoothing parameters. On the basis of of the examples in sections 3 and 4 the new method appears to be somewhat more robust and faster than the mixed model approach using standard mixed modelling software, but section 3.4 also suggests that REML consistently does a slightly better job at estimating the models than GCV based methods.

The second advantage of the new method lies in the extensions to existing GAM methodology facilitated by the ability to incorporate a fixed penalty in the model. The fixed penalty allows ridge regression type regularization of other wise un-identifiable GAMs, an appealing solution in situations where in some sense the model is perfectly identifiable on the *response* scale. Further, the ability to set lower bounds on smoothing parameters and to fix some penalties while estimating others is also of considerable use in applications, both for basic model checking, and in situations where there are a priori reasons to believe that some functions should be smoother than others, and it is useful to explore the consequences of forcing them to be so (for example when a spatial term has been included simply to account for otherwise un-modelled trend).

I believe that these benefits are important if GAM methods are to achieve their full potential. GAMs offer the modeller great flexibility, but such flexibility inevitably expands the scope for difficulties with identifiability and numerical stability. If models are fit using methods designed for maximum stability, able to cope with rank deficiency and as a last resort allowing direct regularization, then such problems can at least be minimized, and GAM methods can nearer approach the reliability of GLMs or linear models.

A third advantage of the method suggested here is the fact that it is relatively simple and uses only matrix factorizations available in public domain linear algebra libraries such as LAPACK and LINPACK. Indeed the method could be implemented in a quite straightforward manner entirely within a high level language such as R or Matlab which give direct access to the required linear algebra routines.

One possible criticism of the use of the type of methods developed in this paper is that well specified models that are really appropriate for a set of data do not usually cause numerical difficulty in fitting, so that the slight extra computational cost of using orthogonal matrix factorizations is unjustified. Although there is some truth in the assertion that convergence problems are often caused by poor models, it is equally true that it is much more satisfactory to be rejecting flawed models on the basis of their poor fit to the data, rather than on the basis that it was not computationally possible to fit them.

If computational speed is really critical, then the method developed here could be made more computationally efficient by using a second pivoted QR decomposition in place of the Singular Value Decomposition. The broad outline of the approach is unchanged by doing this, although the detail is of course different. The pivoted QR decomposition is substantially cheaper to obtain than the full SVD, but rank estimation with the pivoted QR is not as straightforward or reliable as it is with the SVD (see Golub and Van Loan, 1996), and some extra work is required if a minimum norm solution is required in rank deficient cases. However, given that the first QR decomposition is actually the leading order step in terms of cost ($O(nq^2)$ operations), the replacement of the SVD step ($O(q^3)$ operations) is unlikely to offer great computational savings.

Another interesting issue relates to whether smoothing parameter estimation is inner or outer to the P-IRLS loop. For maximal computational efficiency the "performance-iteration" method of Gu (e.g. Gu 2002) is appealing: smoothing parameter selection is performed on the weighted penalized least squares problem produced at each iteration of the P-IRLS algorithm. This method is very fast and a perfectly legitimate way to estimate smoothing parameters, but it is usually possible to find smoothing parameters yielding slightly lower GCV or UBRE scores than the smoothing parameters estimated using the power iteration. The reason for this is that the dependence of the iterative weights on the smoothing parameters is effectively neglected in the performance iteration method. Hence it may sometimes be desirable to use the approach originally suggested by O'Sullivan *et al.* (1986) of only evaluating the GCV/UBRE score at convergence of the P-IRLS method, and therefore making the minimisation of the GCV/UBRE score 'outer' to the P-IRLS loop. This can easily be done by using a quasi-Newton method with finite differencing of the GCV/UBRE score evaluated at P-IRLS convergence (the default for the R general purpose minimizer routine `nlm`), but of course such an approach is much slower than the performance iteration and still requires the performance iteration in order to find starting values. In the cases that I have looked at, such an approach leads to only small reductions in score and small changes in fit relative to the performance iteration, and hence its statistical benefits are questionable.

A final interesting question is whether an approach like the one developed here could be used in the case of low rank approximations to SS-ANOVA models in which the smoothing parameters occur in both the overall penalty matrix and the model matrix. In this case the expressions for the derivatives will be more complicated, but if the model matrix can be written in the form $\mathbf{X} = \sum_j \theta_j \mathbf{X}_j$ then an efficient method could be developed on the basis of first forming a QR factorization of the matrix $[\mathbf{X}_1, \mathbf{X}_2, \ldots]$.

The method reported in this paper is available in the free open source R package `mgcv` (versions 0.9 and above, see `cran.r-project.org`).

Acknowledgements

About half of this work was undertaken as part of my paid employment, for which I gratefully acknowledge the Universities of St. Andrews and Glasgow as well as NERC grants NER/T/S/1999/00072 and NER/T/S/2002/00142. I am also grateful to Mark Bravington and Sharon Hedley for particularly useful discussions of parts of this work during a CSIRO funded visit to CSIRO Tasmania, and to Stefan Lang for help with the mixed model approach to GAMs on a visit to the University of Munich. Thanks also to an associate editor for making a number of very helpful suggestions including the mixed model comparisons; an almost anonymous referee for further helpful comments; Miguel Bernal and Yorgos Stratoudaki for discussion, patience and the examples that broke the existing methods, as well as their institutions, IPIMAR and IEO, for permission to use the egg data. Finally to Charles Paxton whose systematic approach was a great help in un-covering the convergence issues that this paper tries to fix, and to numerous other people at St Andrews for stimulating discussions on GAMs in general.

Appendix: GAMs via mixed models

GAMs admit a straightforward mixed model representation in which components of the smooths in the null space of the penalties and any strictly parametric model terms are treated as fixed effects, while the "wiggly" components are treated as random effects. This allows smoothing parameters to be estimated via REML (see e.g. Wahba, 1985; Wang, 1998; Lin and Zhang, 1999).

The approach is most easily explained with reference to a single smooth of Gaussian data, but subject to the sort of centering condition required of components of a GAM to ensure identifiability (e.g. that the smooth should sum to zero over the covariate values). Note that in order to keep it straightforward, the notation in this appendix does not always correspond to that used in the main body of the paper.

Consider a smooth term with parameter vector $\boldsymbol{\beta}$, model matrix $\mathbf{X}$, penalty matrix $\mathbf{S}$ and constraint matrix $\mathbf{C}$, which could be estimated by minimization of the penalized regression objective

$$s(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta} \;\text{ subject to }\; \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$$

with respect to $\boldsymbol{\beta}$. The constraint can easily be absorbed by forming the QR decomposition $\mathbf{QR} = \mathbf{C}^{\mathrm{T}}$, setting $\mathbf{Z}$ to be $\mathbf{Q}$ less its first n_c columns, where n_c is the number of rows of $\mathbf{C}$, and writing $\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\beta}_z$ so the fitting objective becomes

$$s = \|\mathbf{y} - \mathbf{XZ}\boldsymbol{\beta}_z\|^2 + \lambda\boldsymbol{\beta}_z^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{SZ}\boldsymbol{\beta}_z.$$

Forming the eigen-decomposition $\mathbf{Z}^{\mathrm{T}}\mathbf{SZ} = \mathbf{UDU}^{\mathrm{T}}$ where $\mathbf{U}$ is orthonormal and $\mathbf{D}$ is diagonal (with eigen-values arranged in order of decreasing magnitude down the leading diagonal), we can write

$$s = \|\mathbf{y} - \mathbf{XZU}\boldsymbol{\beta}_u\|^2 + \lambda\boldsymbol{\beta}_u^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}_u$$

where $\boldsymbol{\beta}_u = \mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}_z$.

Now $\mathbf{S}$ is generally rank deficient so that the last few elements on the leading diagonal of $\mathbf{D}$ will be zero. Let $\mathbf{D}^+$ be the sub matrix of $\mathbf{D}$ with non-zero elements on the leading diagonal and partition $\boldsymbol{\beta}_u$ so that $\boldsymbol{\beta}_u^{\mathrm{T}} = [\mathbf{b}_u^{\mathrm{T}}, \boldsymbol{\beta}_F^{\mathrm{T}}]$ and $\boldsymbol{\beta}_u^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}_u = \mathbf{b}_u^{\mathrm{T}}\mathbf{D}^+\mathbf{b}_u$. Partitioning the columns of $\mathbf{XZU}$ into $[\mathbf{X}_u, \mathbf{X}_F]$ in a corresponding manner to the partitioning of $\boldsymbol{\beta}_u$, while letting $\mathbf{b} = \sqrt{\mathbf{D}^+}\mathbf{b}_u$ and $\mathbf{X}_R = \mathbf{X}_u(\sqrt{\mathbf{D}^+})^{-1}$ the objective becomes

$$s = \|\mathbf{y} - \mathbf{X}_F\boldsymbol{\beta}_F - \mathbf{X}_R\mathbf{b}\|^2 + \lambda\mathbf{b}^{\mathrm{T}}\mathbf{b}.$$

Now it is easy to show that, given λ, the estimates of $\mathbf{b}$ and $\boldsymbol{\beta}_F$ that result from minimizing s correspond to the expected values of $\mathbf{b}$ and the estimates of $\boldsymbol{\beta}_F$ given $\mathbf{y}$ under the mixed model

$$\mathbf{y} = \mathbf{X}_F\boldsymbol{\beta}_F + \mathbf{X}_R\mathbf{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2), \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{I}\tau)$$

where $\lambda = 1/\tau$ (by a simple generalization of Silverman's, 1985, Bayesian model for cubic smoothing splines). This immediately suggests estimating $1/\lambda$ by REML, which is straightforward using standard software such as `lme` in S.

In the additive model context, one produces an $\mathbf{X}_F$ and $\mathbf{X}_R$ for each smooth term in the model. The columns of the $\mathbf{X}_F$'s are combined into one fixed effects model matrix, $\mathbf{X}_\Sigma$ (usually with an additional column for the model intercept and possibly some extra columns for the strictly parametric part of the model), so that the model becomes something like

$$\mathbf{y} = \mathbf{X}_\Sigma\boldsymbol{\beta}_\Sigma + \mathbf{X}_{R1}\mathbf{b}_1 + \mathbf{X}_{R2}\mathbf{b}_2 + \ldots + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2), \quad \mathbf{b}_1 \sim N(\mathbf{0}, \mathbf{I}\tau_1) \quad \mathbf{b}_2 \sim N(\mathbf{0}, \mathbf{I}\tau_2), \ \ldots$$

Simultaneous estimation of the variance components/ smoothing parameters is as straightforward in this case as in the single term case.

In the work reported in the body of the paper the model terms were represented as thin plate regression splines (Wood, 2003) and estimation was performed using `lme` (Pinheiro and Bates, 2000) in R. For the generalized case the setup is identical except that the error model for $\mathbf{y}|\boldsymbol{\beta}_\Sigma, \mathbf{b}_1, \mathbf{b}_2, \ldots$ is different and estimation can be performed by Penalized Quasi Likelihood (Breslow and Clayton, 1993). The R routine `glmmPQL` (Venables and Ripley 2002, Section 10.4) was used in the work reported here.

References

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association **88**: 9- 25.

Buja, A., T. Hastie and R. Tibshirani (1989) Linear smoothers and additive models. *The Annals of Statistics.* **17**: 453-510.

Chambers, J.M. and Hastie, T.J. (eds. 1993) *Statistical Models in S* Chapman and Hall, New York.

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

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

Golub, G.H, and van Loan, C. F. (1996) *Matrix Computations.* Oxford University Press.

Gu, C. (1992) Cross-validating non-Gaussian data. *J. Comput. Graph. Statist.* **1**, 169-179

Gu, C. (2002) *Smoothing spline ANOVA models* Springer-Verlag, New York

Gu, C. and Kim, Y.J. (2002) Penalized Likelihood Regression: General Formulation and Efficient Approximation *Can. J. Stat.* **30 (4)**, 619-628

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.

Hastie, T. (1996) Pseudosplines. *J.R. Statist. Soc. B* **58(2)**, 379-396.

Hastie, T. and Tibshirani, R.J. (1986) Generalized additive models. (with discussion). *Statist. Sci.* **1**, 297-318

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

Lin X. and Zhang, D. (1999) Inference in generalized additive mixed models using smoothing splines. *J.R. Statist. Soc. B* **61**, 381-400.

Marx B.D. and Eilers, P.H.C. (1998) Direct generalized additive modeling with penalized likelihood. *Comp. Statist. Data. Anal.* **28**, 193-209

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models (2nd ed.)* Chapman and Hall, London.

Parker, R.L. and Rice, J.A. (1985) Discussion of Dr Silverman's paper. *J.R. Statist. Soc. B* **47**, 40-42.

Pinheiro, J.C. and Bates, D.M. (2000) *Mixed- Effects Models in S and S-PLUS* Springer-Verlag, New York.

R Development Core Team (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.

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

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.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics in S 4th ed.* , Springer-Verlag, New York.

Wahba, G, (1980) Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In *Approximation Theory III* W. Cheney, ed., Academic Press, New York, pp.905-912

Wahba, G, (1985) A Comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* 13:1378-1402

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.

Watkins, D.S. (1991) Fundamentals of Matrix Computation. *John Wiley and Sons, New York*

Wood, S.N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc. B* **62**, 413-428.

Wood, S.N. (2003) Thin plate regression splines. *J. R. Statist. Soc. B* **65**, 95-114.

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance *J.R. Statist. Soc. B* **60**, 159-174.

Xiang, D. (2001) Fitting Generalized Additive Models with the GAM Procedure. SAS Institute Paper P256-26.