*J. R. Statist. Soc.* B (2003)
**65**, *Part* 1, *pp.* 95–114

# Thin plate regression splines

Simon N. Wood

*University of St Andrews, UK*

[Received October 2001. Final revision June 2002]

**Summary.** I discuss the production of low rank smoothers for $d \geqslant 1$ dimensional data, which can be fitted by regression or penalized regression methods. The smoothers are constructed by a simple transformation and truncation of the basis that arises from the solution of the thin plate spline smoothing problem and are optimal in the sense that the truncation is designed to result in the minimum possible perturbation of the thin plate spline smoothing problem given the dimension of the basis used to construct the smoother. By making use of Lanczos iteration the basis change and truncation are computationally efficient. The smoothers allow the use of approximate thin plate spline models with large data sets, avoid the problems that are associated with 'knot placement' that usually complicate modelling with regression splines or penalized regression splines, provide a sensible way of modelling interaction terms in generalized additive models, provide low rank approximations to generalized smoothing spline models, appropriate for use with large data sets, provide a means for incorporating smooth functions of more than one variable into non-linear models and improve the computational efficiency of penalized likelihood models incorporating thin plate splines. Given that the approach produces spline-like models with a sparse basis, it also provides a natural way of incorporating unpenalized spline-like terms in linear and generalized linear models, and these can be treated just like any other model terms from the point of view of model selection, inference and diagnostics.

*Keywords*: Generalized additive model; Regression spline; Thin plate spline

## 1. Introduction

Smoothing splines (Duchon, 1977; Wahba, 1990; Gu, 2002) provide an excellent means for estimation and inference with models like

$$y_i = f(x_i) + \varepsilon_i, \tag{1}$$

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i \tag{2}$$

or

$$y_i = f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + f_3(x_{4i}) + \ldots + \varepsilon_i \tag{3}$$

where in all cases $y$ is a response variable, the $x$s are covariates, the $f$s are smooth functions and the $\varepsilon$-terms are random variables (independent for different $i$).

For example, model (1) can be estimated by finding the function from an appropriate reproducing kernel Hilbert space which minimizes

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda \int f''(x)^2 \, \mathrm{d}x \tag{4}$$

*Address for correspondence*: Simon N. Wood, Mathematical Institute, University of St Andrews, North Haugh, St Andrews, Fife, KY16 9SS, UK.
E-mail: snw@st-and.ac.uk

 1369–7412/03/65095

where $\mathbf{y}$ is a vector of $y_i$s, $\mathbf{f}$ is the corresponding vector of $f(x_i)$-values and $\|\cdot\|$ is the Euclidean norm. $\lambda$ is a smoothing parameter, which must be chosen appropriately if the right balance is to be struck between minimizing model badness of fit as measured by the first term and model wiggliness as measured by the second. The result of this minimization turns out to be finite dimensional and is a cubic spline, which is a special case of a thin plate spline. In general these are obtained as the solution of the generalization of expression (4) to problems in which $f$ is a function of any finite number $d \geqslant 1$ of covariates and the order $m$ of differentiation in the wiggliness penalty can be any integer satisfying $2m > d$ (see Section 2). A further straightforward generalization of expression (4) is the replacement of the least squares term in the objective with a negative log-likelihood based on an exponential family distribution (see for example Green and Silverman (1994) and Gu (2002)).

There are two obstacles to the widespread adoption of thin plate spline smoothers in practical statistical work. The first is computational. To fit a thin plate spline to $n$ data requires the estimation of $n$ parameters and an additional smoothing parameter. Except in the case $d = 1$ this involves $O(n^3)$ operations, which is frequently prohibitive. Indeed, without the availability of efficient $O(n)$ algorithms for the $d = 1$ case (e.g. Hutchinson and de Hoog (1995)) it is doubtful that cubic smoothing splines would have achieved their current popularity. (Furthermore, although not generally critical, thin plate spline fitting problems can have condition numbers in excess of $10^9$, which has the potential to cause problems if a thin plate spline is embedded in a non-linear model, for example.) The second obstacle to a widespread adoption of these smoothers is the fact that their use requires a change in modelling methodology relative to conventional linear or generalized linear modelling: the flexibility of a fitted model must be selected by adjusting the smoothing parameter $\lambda$, rather than by adding or dropping model terms. This precludes many model building strategies that are ordinarily used for (generalized) linear models.

One approach to the problem of computational cost is to employ regression splines. The basis implied by solving the spline smoothing problem for a small representative data set is found and this small basis is used to construct a model for the full data set of interest. The model is typically fitted as a linear or generalized linear model without imposing a wiggliness penalty. The covariate points that are used to obtain the reduced basis are known as the 'knots' of the regression spline. The number of knots controls the flexibility of the model, but unfortunately their location also tends to have a marked effect on the fitted model (see for example Hastie and Tibshirani (1990), section 9.3). In principle, conventional hypothesis-testing-based model selection can be used to determine the appropriate flexibility for regression spline models, but in practice there are difficulties. If the knots of order $k$ and order $k - 1$ regression spline models for a data set are arranged to ensure the best performance of both models, then the two models will not generally be nested. Alternatively, if knots are not moved, but some knots are simply dropped during model selection, then nesting is maintained, but very uneven knot spacings can result: this has undesirable approximation theoretic consequences (see for example Wahba (1990), page ix). Another more subtle problem with the latter strategy is 'knot confounding' (Zhou and Shen, 2001). Finally, when $d > 1$, even deciding where to place knots so that they appear evenly spread through the covariates can become problematic.

Some of the problems with knot placement can be partially alleviated by abandoning pure regression splines in favour of penalized regression splines (e.g. Wahba (1980) and Parker and Rice (1985)), where the required penalty is that associated with the regression spline basis. But in this case model flexibility is again controlled by a smoothing parameter $\lambda$, rather than the basis dimension, so that some conventional (generalized) linear modelling methods are once again inapplicable.

The first aim of this paper is to find optimal approximations to the thin plate splines which will remove the computational obstacles to their use, while minimizing the deterioration in model performance that is entailed by the approximation (i.e. to find optimal penalized regression splines). The second aim is to remove the knot placement problem from regression spline modelling in a way that will allow model selection by the hypothesis testing methods that are usually employed in regression modelling. Two immediate results of achieving these aims are to provide a good way of incorporating smooth function terms into non-linear models and to provide a way of incorporating thin plate spline like terms into generalized additive models (GAMs).

## 2. Low rank thin plate spline like smoothers

This section begins with standard, but essential, background material on thin plate splines (Duchon, 1977) and then uses these standard results as the starting-point for the production of low rank smoothers with 'good' properties. Purely for simplicity of presentation, I shall ignore the possibility of tied covariate values for the moment and cover them later. Consider the problem of estimating the smooth function $f(\mathbf{x})$ where $\mathbf{x}$ is a $d$-vector, from $n$ ($\geqslant d$) observations $(y_i, \mathbf{x}_i)$ such that

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

where $\varepsilon_i$ is a random error term. Thin plate splines can be used to estimate $f$ by finding the function $g$ minimizing

$$\|\mathbf{y} - \mathbf{g}\|^2 + \lambda J_{md}(g) \tag{5}$$

where $\mathbf{y}$ is the vector of $y_i$ data, $\mathbf{g} = (g(\mathbf{x}_1), g(\mathbf{x}_2), \ldots, g(\mathbf{x}_n))'$, $J_{md}(g)$ is a penalty functional measuring the wiggliness of $g$ and $\lambda$ controls the trade-off between data fitting and smoothness of $g$. The wiggliness penalty is defined as

$$J_{md} = \int \ldots \int_{\Re^d} \sum_{\nu_1+\ldots+\nu_d=m} \frac{m!}{\nu_1! \ldots \nu_d!} \left( \frac{\partial^m g}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 \ldots \mathrm{d}x_d. \tag{6}$$

Provided that we impose the technical restriction $2m > d$, it can be shown that the function minimizing expression (5) has the form

$$g(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(\|\mathbf{x} - \mathbf{x}_i\|) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}) \tag{7}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are unknown parameter vectors subject to the constraint that $\mathbf{T}'\boldsymbol{\delta} = \mathbf{0}$ and $T_{ij} = \phi_j(\mathbf{x}_i)$. The $M = \binom{m+d-1}{d}$ functions $\phi_i$ are linearly independent polynomials spanning the space of polynomials in $\Re^d$ of degree less than $m$ (i.e. the space of polynomials for which $J_{md}$ is 0). Furthermore

$$\eta_{md}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} \log(r) & d \text{ even}, \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & d \text{ odd}. \end{cases}$$

Now, defining matrix $\mathbf{E}$ by $E_{ij} \equiv \eta_{md}(\|\mathbf{x}_i - \mathbf{x}_j\|)$, the spline fitting problem becomes

$$\text{minimize } \|\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\delta}'\mathbf{E}\boldsymbol{\delta} \text{ subject to } \mathbf{T}'\boldsymbol{\delta} = \mathbf{0} \tag{8}$$

with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$. The function $g$ obtained by solving this system is something of an ideal smoother—it has been constructed by defining exactly what is meant by smoothness, exactly how much weight to give to the conflicting goals of matching the data and making $g$ smooth, and finding the function that best satisfies the resulting smoothing objective. The disadvantage is that the resulting smoother has a rather high rank—its computation requires that we estimate as many parameters as there are data: except for the case when $d = 1$ this means that the calculations require $O(n^3)$ operations. See Wahba (1990) or Green and Silverman (1994) for further information about thin plate splines.

### 2.1. Constructing an optimal approximating basis

In this section a family of low rank smoothers is constructed by starting from the ideal smoothing problem (8), finding the parameter space basis of a given rank that perturbs this problem as little as possible and solving the resulting low rank problem. The basis of the unpenalized functions is left unchanged—since these are the functions of zero wiggliness according to the measure used, it would make little sense to truncate their basis. I shall concentrate instead on the basis for the $\boldsymbol{\delta}$ parameter space. The ideal basis would be one that results in minimum change of both the goodness-of-fit term and the penalty term for any given $\boldsymbol{\delta}$, but of course no single basis can achieve this for all $\boldsymbol{\delta}$, so less ambitious criteria must be adopted.

To motivate the criteria for choosing a truncated basis, consider the rank $k$ matrix $\boldsymbol{\Gamma}_k$, the columns of which form a $k$-dimensional orthonormal basis for the $\boldsymbol{\delta}$ parameter space, so that $\boldsymbol{\delta} = \boldsymbol{\Gamma}_k\boldsymbol{\delta}_k$ where $\boldsymbol{\delta}_k$ is a $k$-vector. $k$ must be greater than $M$. Within the space spanned by $\boldsymbol{\Gamma}_k$ problem (8) becomes

$$\text{minimize } \|\mathbf{y} - \mathbf{E}\boldsymbol{\Gamma}_k\boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\delta}_k'\boldsymbol{\Gamma}_k'\mathbf{E}\boldsymbol{\Gamma}_k\boldsymbol{\delta}_k \text{ subject to } \mathbf{T}'\boldsymbol{\Gamma}_k\boldsymbol{\delta}_k = \mathbf{0}.$$

Defining the rank $k$ matrices $\tilde{\mathbf{E}}_k = \mathbf{E}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'$ and $\hat{\mathbf{E}}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\mathbf{E}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'$, problem (8) can be written as

$$\text{minimize } \|\mathbf{y} - \tilde{\mathbf{E}}_k\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\delta}'\hat{\mathbf{E}}_k\boldsymbol{\delta} \text{ subject to } \mathbf{T}'\boldsymbol{\delta} = \mathbf{0} \tag{9}$$

where $\boldsymbol{\delta} = \boldsymbol{\Gamma}_k\boldsymbol{\delta}_k$. An ideal $\boldsymbol{\Gamma}_k$ would induce a problem of the form (9) that is as close as possible to problem (8). To find such a basis requires a definition of what constitutes 'as close as possible'.

Considering the least squares term first, it is clear that the goal of minimizing the change in this term for all $\mathbf{y}$ and $\boldsymbol{\delta}$ (or even for those $\mathbf{y}$ and $\boldsymbol{\delta}$ that are consistent with being best fits according to the objective) cannot be achieved with a single basis selected independently of the data and parameter values. Instead, for a given $\boldsymbol{\delta}$, I focus on trying to minimize the change in fitted values ($\mathbf{E}\boldsymbol{\delta} + \mathbf{T}\boldsymbol{\alpha}$) caused by substituting $\mathbf{E}$ with $\tilde{\mathbf{E}}_k$, the rank deficient approximation induced by the change and truncation of basis. The basis change and truncation will cause a change $(\mathbf{E} - \tilde{\mathbf{E}}_k)\boldsymbol{\delta}$ in the fitted values. It is clearly not possible to find a single basis that will uniformly minimize this quantity for all $\boldsymbol{\delta}$, but a more feasible objective is obtained by weakening requirements further and seeking to minimize the 'worst' possible change:

$$\varepsilon_k = \max_{\boldsymbol{\delta} \neq \mathbf{0}} \left\{ \frac{\|(\mathbf{E} - \tilde{\mathbf{E}}_k)\boldsymbol{\delta}\|}{\|\boldsymbol{\delta}\|} \right\}$$

where $\|\cdot\|$ is the usual Euclidean norm. The scaling by $\|\boldsymbol{\delta}\|$ is necessary to ensure, for example, that the resulting smoothers do not have different behaviours when different measurement scales for $y_i$ are used. The intuitive idea is that the basis change and truncation should make

minimal change to the model fit, although the measure that is used for this necessarily weakens the intuitive criterion a little.

Turning to the penalty term, similar reasoning suggests that a suitable measure of the worst possible change introduced by the basis truncation is

$$e_k = \max_{\boldsymbol{\delta} \neq \mathbf{0}} \left\{ \frac{\boldsymbol{\delta}'(\mathbf{E} - \hat{\mathbf{E}}_k)\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2} \right\}$$

and again the aim is to choose the basis minimizing this quantity, for a given $k$. The intuitive idea is that the basis change and truncation should make minimal change to the shape of the smooth function as measured by the penalty functional.

Given the goal of simultaneously minimizing $e_k$ and $\varepsilon_k$, the appropriate basis to use turns out to be a truncated eigenbasis of $\mathbf{E}$. Specifically let $\mathbf{E} = \mathbf{UDU}'$ where $\mathbf{D}$ is a diagonal matrix of eigenvalues of $\mathbf{E}$ arranged so that $|D_{i,i}| \geqslant |D_{i+1,i+1}|$, $i = 1, \ldots, n-1$, and $\mathbf{U}$ is a matrix whose $i$th column is the eigenvector corresponding to $D_{i,i}$. The best basis of rank $k$ is given by $\mathbf{U}_k$ (i.e. by setting $\boldsymbol{\Gamma}_k = \mathbf{U}_k$): the first $k$ columns of $\mathbf{U}$, which implies that $\hat{\mathbf{E}}_k = \tilde{\mathbf{E}}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k'$ $(= \mathbf{E}_k$, say).

It is easy to demonstrate that the basis chosen minimizes $\varepsilon_k$. $\varepsilon_k = \|\mathbf{E} - \tilde{\mathbf{E}}_k\|_2$, the spectral norm of $\mathbf{E} - \tilde{\mathbf{E}}_k$ (see for example Watkins (1991)). But it is well known that of all rank $k$ matrices $\mathbf{F}_k$, the matrix $\mathbf{E}_k$ based on truncating the eigenvalues of smallest magnitude in the spectral decomposition of $\mathbf{E}$, minimizes $\|\mathbf{E} - \mathbf{F}_k\|_2$ (see for example Watkins (1991), page 413).

Demonstrating that the basis also minimizes $e_k$ is only slightly more involved. First define the (symmetric) matrix $\boldsymbol{\Delta}_k = \mathbf{E} - \hat{\mathbf{E}}_k$. It is straightforward to produce a square root of $\boldsymbol{\Delta}_k$, $\boldsymbol{\Delta}_k^{1/2}$, by taking square roots of the eigenvalues of $\boldsymbol{\Delta}_k$ in its spectral decomposition. This means that

$$e_k = \max_{\boldsymbol{\delta} \neq \mathbf{0}} \left\{ \frac{\|\boldsymbol{\Delta}_k^{1/2}\boldsymbol{\delta}\|^2}{\|\boldsymbol{\delta}\|^2} \right\}$$

so that $e_k = \|\boldsymbol{\Delta}_k^{1/2}\|_2^2$, the squared spectral norm of $\boldsymbol{\Delta}_k^{1/2}$. Since the spectral norm of a matrix is given by its largest singular value (which here corresponds to the magnitude of its largest (magnitude) eigenvalue), it is clear from the construction of $\boldsymbol{\Delta}_k^{1/2}$ that $\|\boldsymbol{\Delta}_k^{1/2}\|_2^2 = \|\boldsymbol{\Delta}_k\|_2 = \|\mathbf{E} - \hat{\mathbf{E}}_k\|_2$, and hence that $e_k$ is minimized by the same basis that minimizes $\varepsilon_k$. (This somewhat remarkable fact is clearly rather special to splines, as is easily verified by considering more general penalized regression problems.)

So, given the choice of basis, $\boldsymbol{\delta} = \mathbf{U}_k\boldsymbol{\delta}_k$ (in which case $\boldsymbol{\delta}_k = \mathbf{U}_k'\boldsymbol{\delta}$) and the approximation to problem (8) becomes

$$\text{minimize } \|\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\delta}_k'\mathbf{D}_k\boldsymbol{\delta}_k \text{ subject to } \mathbf{T}'\mathbf{U}_k\boldsymbol{\delta}_k = \mathbf{0}.$$

Now find any orthogonal column basis $\mathbf{Z}_k$ such that $\mathbf{T}'\mathbf{U}_k\mathbf{Z}_k = \mathbf{0}$ ($QR$- or $QT$-factorization will provide this easily; see for example Gill *et al.* (1981)). Restricting $\boldsymbol{\delta}_k$ to this space by writing $\boldsymbol{\delta}_k = \mathbf{Z}_k\tilde{\boldsymbol{\delta}}$ yields the unconstrained problem that must be solved to fit this 'best' rank $k$ approximation to the smoothing spline:

$$\text{minimize } \|\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\tilde{\boldsymbol{\delta}}'\mathbf{Z}_k'\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}}$$

with respect to $\tilde{\boldsymbol{\delta}}$ and $\boldsymbol{\alpha}$. Having fitted the model, evaluation of the spline at any point is easy: just evaluate $\boldsymbol{\delta} = \mathbf{U}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}}$ and use equation (7). In the rest of this paper I shall often refer to the parameter vector and design matrix of a thin plate regression spline as $\boldsymbol{\beta}' \equiv (\tilde{\boldsymbol{\delta}}', \boldsymbol{\alpha}')$ and $\mathbf{X} \equiv (\mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k, \mathbf{T})$ respectively and the wiggliness penalty matrix as $\mathbf{S}$, which will have $\mathbf{Z}_k'\mathbf{D}_k\mathbf{Z}_k$ in its top left-hand corner and 0s elsewhere, so that the wiggliness penalty is $\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}$.

These thin plate regression splines can be treated as pure regression splines by setting $\lambda$ to 0. In this case model selection becomes a matter of choosing $k$, which can be performed either by criteria like generalized cross-validation (GCV) and the Akaike information criterion (AIC) or by conventional hypothesis-testing-based model selection, since the columns of $\mathbf{U}_{k-1}\mathbf{D}_{k-1}\mathbf{Z}_{k-1}$ clearly span a subspace of the space spanned by the columns of $\mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k$ (the subspace of the latter is the intersection of the null space of the constraints and the subspace spanned by $\mathbf{U}_k\mathbf{D}_k$ whereas the subspace of the former is the intersection of the same null space of the constraints with a subspace of $\mathbf{U}_k\mathbf{D}_k$). Of course the model terms are not orthogonal but, since there is a natural order in which to consider their deletion from the model, this does not matter for practical purposes.

Alternatively thin plate regression splines can be treated as penalized regression splines, in which case the value chosen for $k$ will not be critical (see Eilers and Marx (1996) for illustration of this in the case of 'P-splines'), but should be somewhat larger than the degrees of freedom believed to be required for the modelling situation concerned. The actual model degrees of freedom will be controlled by $\lambda$, which must be selected by some criterion like GCV, generalized maximum likelihood or AIC (see for example Craven and Wahba (1979), Wahba (1990) and Akaike (1973)) or by considering the spline as a random effect. In practice $k$ should probably be increased if the estimated $\lambda$ is too close to 0: one pragmatic approach would be to increase $k$ if the estimated degrees of freedom (see Section 3.1) for a thin plate regression spline exceeds some specified proportion (e.g. 0.8–0.9) of the basis dimension. Note that in the penalized case the penalty (6) has not been replaced by an approximate penalty: $\tilde{\boldsymbol{\delta}}'\mathbf{Z}_k'\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}}$ is exactly penalty (6) for any function in the truncated space.

Tied covariate values are dealt with by simply reducing the data set to one involving only unique covariate combinations, setting up the thin plate regression spline for this reduced data set and then duplicating rows of the resulting $\mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k$- and $\mathbf{T}$-matrices as necessary to model the full set of data.

This section has presented a way of obtaining approximate thin plate splines, which are suitable for incorporation into a wide range of model structures. The approximations are 'optimal', but in a slightly weak sense: the criteria are not minimized subject to the linear restrictions $\mathbf{T}'\boldsymbol{\delta} = \mathbf{0}$ that are applied for model fitting (it is not possible to minimize both criteria simultaneously under that restriction). None-the-less, given the good performance of the approximation reported below, it is useful to know the sense in which the approximation is optimal (and in practice it was the search for some sort of optimality that led to the approach reported here, rather than more obvious approaches).

### 2.2. Computational issues

Discarding the small magnitude eigenvectors of $\mathbf{E}$ can only improve the numerical conditioning of the thin plate regression splines relative to full thin plate splines, but in addition they have two further computational advantages:

(a) for small data sets they can be implemented very easily using linear algebra routines that are readily available in standard statistical packages;
(b) for large data sets it is possible to obtain thin plate regression spline bases very efficiently by using Lanczos iteration.

The first point is best appreciated by examining the steps, given in Appendix A, for implementing thin plate regression splines by using standard software. Such an approach might be appropriate

for incorporating smooth functions into a linear, generalized linear or non-linear model of a relatively small set of data.

For larger data sets the potential *computational* benefits of the thin plate regression spline approach relative to full thin plate spline models will only be fully realized if the truncated eigendecomposition of $\mathbf{E}$ can be calculated in substantially fewer than the $O(n^3)$ operations required for a full eigendecomposition. Lanczos iteration (see for example Demmel (1997)) is a method which obtains the truncated eigendecomposition in $O(kn^2)$ operations, by iteratively building up a tridiagonal matrix, the eigenvalues of which converge (in order of decreasing magnitude) to those required, as iteration proceeds. Appendix B gives details of an implementation that is suitable for use with $\mathbf{E}$. Note, for example, that fitting a thin plate regression spline to 5000 data using $k = 50$ will be of the order of 100 times faster using an $O(kn^2)$ algorithm as opposed to a standard $O(n^3)$ algorithm.

In the context of very large data sets, even greater computational efficiency could be achieved by using an approximate eigendecomposition calculated using the Nyström methods described, for example, in Williams and Seeger (2001) (see also Smola and Schölkopf (2000) for a related approach—I am grateful to a referee for pointing this out). However, in such cases it is probably more straightforward to subsample the data (e.g. to select randomly 1000 data points), to produce a thin plate regression spline basis for this subsample and to use this basis for the model of the whole data set.

Finally, note two computational tricks for avoiding poor numerical conditioning. Firstly, when $m > 2$, collinearity in the columns of $\mathbf{T}$ can be avoided by subtracting the mean from each covariate, so that each is centred near 0. Secondly, it is worth linearly transforming the model parameters to ensure that the columns of $\mathbf{X}$ have broadly similar average element sizes; otherwise 'poor scaling' of $\mathbf{X}$ can sometimes detract from numerical stability.

## 3. Practical properties: some simulation results

This section provides some straightforward illustrations of the advantages of the thin plate regression spline approach, relative to the more obvious 'knot placement' approaches and to full spline smoothing.

### 3.1. Comparison with 'knot-based' regression splines

The suggested thin plate regression spline basis allows model selection by using conventional hypothesis testing approaches in a way that is difficult by using traditional regression splines. Furthermore, given its theoretical motivation, a thin plate regression spline should be better able to represent most smooth functions than a smoother based on selecting a small set of 'knots' with which to construct a basis.

To illustrate the latter point, 100 random $(x, z)$ points in the unit square were chosen and the test function

$$f(x,z) = \frac{0.75}{\pi\sigma_x\sigma_z}\exp\{-(x-0.2)^2/\sigma_x^2 - (z-0.3)^2/\sigma_z^2\} + \frac{0.45}{\pi\sigma_x\sigma_z}\exp\{-(x-0.7)^2/\sigma_x^2 - (z-0.8)^2/\sigma_z^2\} \quad (10)$$

was evaluated at each ($\sigma_x = 0.3$ and $\sigma_z = 0.4$). The function was then reconstructed by fitting these data using a rank 16 thin plate regression spline and a more traditional rank 16 regression spline. The traditional regression spline was constructed by placing 16 points on a regular

lattice across the unit $(x, z)$ square and obtaining the basis that would have resulted by fitting a thin plate spline to response data at these 16 points. This basis was then used to represent the function to be fitted to the 100 data points from the test function. 16-dimensional bases were chosen because they allow a favourable regular lattice to be used for the knot-based spline but are also close to the basis dimension selected by hypothesis testing in the next example in this subsection. The results are shown in Fig. 1. For this rank of smoother in this example the thin plate regression spline improves considerably on the knot-based spline. Such differences become less marked for much higher or lower ranks but, for pure regression spline modelling (as opposed to penalized regression spline modelling), intermediate ranks are of most interest (rank 15 is selected for this test function in the next example). The ability of the thin plate regression spline to represent underlying functions using relatively few parameters should reduce estimator variances relative to approaches requiring more parameters.

The example shown in Fig. 1 was also repeated with noisy data. 100 parameter sets were generated at each of seven noise levels. For each replicate, 100 $(x_i, z_i)$ points were generated randomly from a uniform distribution on the unit square. Function (10) was evaluated at each point and perturbed by additive Gaussian noise with standard deviation $\sigma$. The test function was reconstructed by fitting a thin plate regression spline and a knot-based spline to each replicate, with the mean-square error (MSE) of reconstruction calculated for both methods (means taken over the 100 $(x_i, z_i)$ points). Table 1 summarizes the results in terms of

$$\Delta_{\mathrm{MSE}} = \{\mathrm{MSE(knot\ \ based)} - \mathrm{MSE(TPRS)}\}/\mathrm{MSE(TPRS)}.$$

The final column in Table 1 gives the number of replicates, out of 100, in which the thin plate regression spline had a lower MSE than the knot-based spline. This comparison is of performance in a pure regression context: Section 3.2 also reports comparisons of thin plate regression and knot-based splines in the context of penalized regression modelling.

Turning to model selection: one approach is to start with a basis that is overparameterized and to truncate it until the truncated model differs significantly from the overparameterized model according to a conventional hypothesis test. As an example, data were simulated from function (10), by randomly choosing 100 $(x_i, z_i)$ locations in the unit square, and then forming data:

$$y_i = f(x_i, z_i) + \varepsilon_i$$

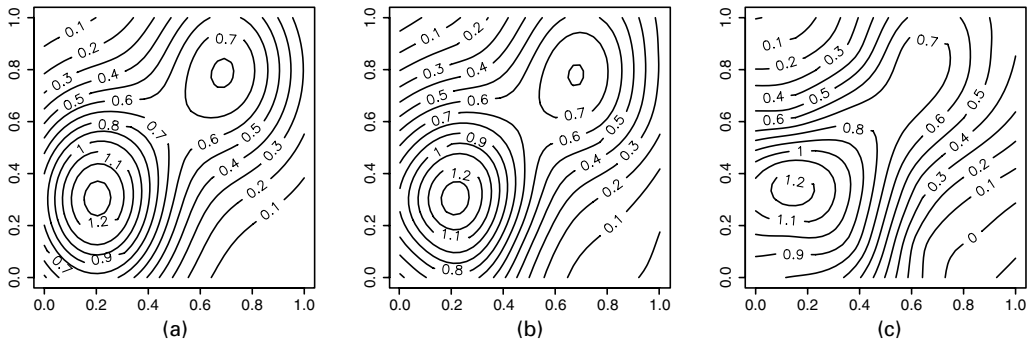


**Fig. 1.** Comparison of alternative reconstructions of the smooth function, using different low rank spline-based approaches (100 noise-free sample points were randomly placed over the function domain shown): (a) true function; (b) reconstruction from using a rank 16 thin plate regression spline basis of the type proposed; (c) reconstruction from using a rank 16 basis constructed by dividing the domain into 16 equal squares and placing a point at the centre of each—the thin plate spline basis that would result from fitting a thin plate spline to data at just these 16 points was then used as the model basis (e.g. Wahba (1980))

**Table 1.** Results of comparing MSE performance of thin plate and knot-based regression splines†

| $\sigma$ | *Mean* $\Delta_{\mathrm{MSE}}$ | *Minimum* $\Delta_{\mathrm{MSE}}$ | *Maximum* $\Delta_{\mathrm{MSE}}$ | $\Delta_{\mathrm{MSE}} > 0$ |
|---|---|---|---|---|
| 0 | 4.85 | 0.31 | 22.77 | 100 |
| 0.01 | 4.68 | 0.40 | 16.00 | 100 |
| 0.02 | 3.34 | 0.44 | 20.62 | 100 |
| 0.05 | 1.32 | 0.01 | 5.38 | 100 |
| 0.1 | 0.42 | −0.07 | 1.76 | 96 |
| 0.2 | 0.12 | −0.31 | 0.83 | 71 |
| 0.5 | 0.02 | −0.49 | 0.70 | 52 |

†See Section 3.1 for details.

where the $\varepsilon_i$ were independent and identically distributed $N(0, 0.1^2)$. Let $\mathrm{rss}_k$ denote the residual sum of squares for the rank $k$ thin plate regression spline model of these data. Testing the null hypothesis that the rank $k_0$ basis describes the model generating the data against the larger alternative that the rank $k_1$ basis is appropriate uses the standard result that under the null hypothesis

$$\frac{(\mathrm{rss}_{k_0} - \mathrm{rss}_{k_1})/(k_1 - k_0)}{\mathrm{rss}_{k_1}/(n - k_1)} \sim F_{k_1 - k_0, n - k_1}$$

where $n$ is the number of data. This result depends on the nested nature of the thin plate regression spline bases of different ranks. In practice, starting from a $k_1$ that is larger than is needed, $k_0$ is reduced until the above $F$-ratio is significant at the investigator's favourite level. Fig. 2 shows the results of applying this approach with a significance level of 5% to select a pure thin plate regression spline model for simulated data and compares this with a penalized thin plate regression spline model of the same data where smoothing parameter ($\lambda$) selection was by GCV. $k_1$ was set to 40, as was $k$ for the penalized model. GCV selected 22 effective degrees of freedom for the model (for penalized models effective degrees of freedom are defined as $\mathrm{tr}\{\mathbf{X}(\mathbf{X}'\mathbf{X} + \hat{\lambda}\mathbf{S})^{-1}\mathbf{X}\}$, where $\hat{\lambda}$ is the estimated smoothing parameter; see for example Wahba (1990)). Significance testing at the 5% level selected 15 degrees of freedom. This sort of difference

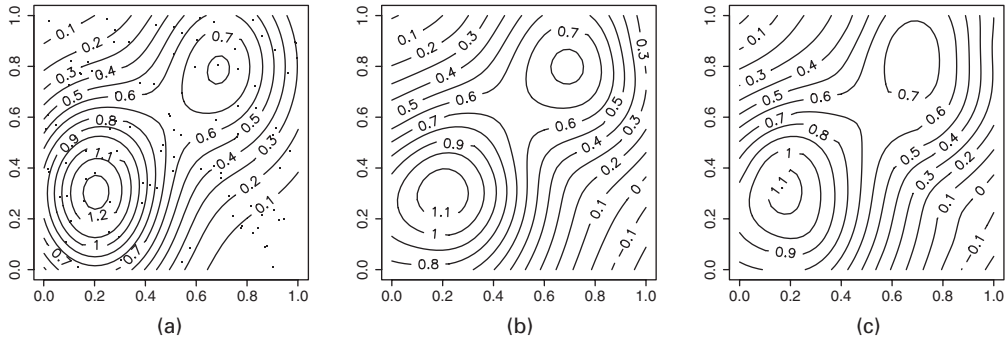


**Fig. 2.** Comparison of a pure thin plate regression spline selected by analysis of variance and a penalized thin plate regression spline: (a) true underlying function from which data have been sampled with Gaussian error ($\sigma$ = 0.1), at the $n$ = 100 randomly chosen sample locations shown; (b) pure regression spline fitted to the data with the rank (15) of the basis chosen by using conventional $F$-ratio testing; (c) penalized regression spline fit to the same data, with the smoothing parameter $\lambda$ chosen by GCV (effective degrees of freedom, 22)

is expected, since GCV is a mean-square prediction error criterion, whereas hypothesis testing addresses the question of how simple a model is plausible for a set of data.

### 3.2. Comparison with 'full' spline models

What is gained and what is lost by using thin plate regression splines rather than full thin plate splines?

One expected gain is computational speed. To illustrate this some timing experiments were carried out using full generalized smoothing spline models as implemented in R package `gss`, and using thin plate regression splines as implemented (by the author) in R package `mgcv`. The results are for estimating the GAM (see Section 4)

$$\log(\mu_i) = f_1(x_i, z_i) + f_2(w_i)$$

where the response variable $y_i \sim \text{Poi}(\mu_i)$, $i = 1, \ldots, n$, and the $f_1$ and $f_2$ used for simulation were both quadratics. Timings given are central processor unit seconds (on a PII 400 MHz computer running Linux). The thin plate penalized regression spline based model used a total of 50 parameters and all smoothing parameters were estimated by GCV. The results are given in Table 2.

The 0.25 Gbytes of memory that were available on the test computer were insufficient for `gss` above $n = 600$, so the final `gss` timing is estimated from the cubic dependence on $n$. As expected, the thin plate regression splines produce quite large reductions in the computational effort required. Partly, this is because the thin plate regression spline calculations are at most $O(kn^2)$, rather than the $O(n^3)$ required for generalized smoothing spline models, but the computational saving is actually greater than the simple comparison of leading order terms would suggest, since the computationally costly model selection algorithm is $O(n^3)$ for the generalized smoothing spline case but only $O(nk^2)$ in the thin plate regression spline case.

The obvious expected loss from using thin plate regression splines would be a degradation of MSE performance. To examine this, experiments were performed using two test functions:

$$f_1(x, z) = 1.9[1.45 + \exp(x)\sin\{13(x - 0.6)^2\}]\exp(-z)\sin(7z)$$

and

$$f_2(x, z) = \exp[\{-(x - 0.25)^2 - (z - 0.25)^2\}/0.1] + 0.5\exp[\{-(x - 0.7)^2 - (z - 0.7)^2\}/0.07].$$

**Table 2.** Central processor unit times required to fit full thin plate spline and thin plate regression spline based GAMS†

| $n$ | *Central processor unit times (s) from the following methods:* | |
|---|---|---|
| | *Generalized spline smoothing* | *Thin plate regression spline* |
| 100 | 2.68 | 1.75 |
| 200 | 11.31 | 2.82 |
| 400 | 88.07 | 4.38 |
| 600 | 316.49 | 6.46 |
| 1200 | 2530‡ | 15.65 |

†See Section 3.2 for full details.
‡Estimated time.

Each test function was sampled at a set of 200 randomly chosen points in the unit square, and the function values at these points were perturbed with additive independent normal random deviates ($\sigma = 0.5$ for $f_1$ and $\sigma = 0.05$ for $f_2$). 100 replicate data sets were generated for each model (design points and errors were different for each replicate), and for each replicate the MSE in reconstructing $f_1$ and $f_2$ was assessed for a thin plate regression spline, a knot-based spline and a full thin plate spline. For each replicate the MSE was averaged over all design points. The basis dimensions for the knot basis and the thin plate regression spline were both 49 for $f_1$ and 36 for $f_2$, choices made to ensure that the knot basis operated on a favourable square regular grid, whereas the model-estimated degrees of freedom were below three-quarters of the basis dimension. For all models the smoothing parameter $\lambda$ was selected by GCV. The full thin plate spline model was fitted by using R package `gss`; other models were fitted by using `mgcv`. The results are summarized in Table 3.

The first row gives the number of times that the MSE of each competing method was lower than the MSE of the thin plate regression spline. The thin plate regression spline has superior performance in most cases. The second row shows the largest MSE difference between the thin plate regression spline and a competing method when the thin plate regression spline had the larger MSE. The third row shows the largest difference between a competing method and a thin plate regression spline when the thin plate regression spline had the smaller MSE. The fourth row shows the mean difference in MSE between each competing method and the thin plate regression spline: note that the thin plate regression spline has the lower mean MSE in all cases. The mean MSE for the thin plate regression spline method was 0.050 for $f_1$ and $3.8 \times 10^{-4}$ for $f_2$. Table 3 shows that the thin plate regression spline usually has the better MSE performance and has better MSE performance on average: in the most extreme cases the improvement is of the same order as the MSE for the thin plate regression spline, whereas the occasional improvements of the competing methods over the thin plate regression spline are quite modest in size.

Fig. 3 shows a randomly chosen example comparison of reconstructions of $f_1$ using the three alternative methods. Fig. 4 shows equivalent example comparative reconstructions for $f_2$—for illustration this shows the worst overfit by the full thin plate spline model obtained in three trial runs.

At first sight the improvement of the truncated model relative to the full spline model is counter-intuitive, but it almost certainly reflects the fact that for the full thin plate spline the degree of model complexity is chosen entirely by GCV, whereas the thin plate regression spline

**Table 3.** Results of comparing MSE performance of thin plate regression splines with full thin plate splines and knot-based regression splines†

| | $f_1$ | | $f_2$ | |
|---|---|---|---|---|
| | *Full thin plate spline* | *Knot basis* | *Full thin plate spline* | *Knot basis* |
| Outperformed thin plate regression spline/100 | 5 | 30 | 0 | 9 |
| Best MSE advantage | 0.0033 | 0.0075 | — | $2.2 \times 10^{-5}$ |
| Worst MSE disadvantage | 0.18 | 0.050 | $2.2 \times 10^{-2}$ | $1.9 \times 10^{-4}$ |
| Mean MSE difference | 0.0091 | 0.0084 | $1.7 \times 10^{-4}$ | $5.5 \times 10^{-5}$ |
| Mean MSE | 0.059 | 0.058 | $5.5 \times 10^{-4}$ | $4.3 \times 10^{-4}$ |

†Full details are given in Section 3.2.

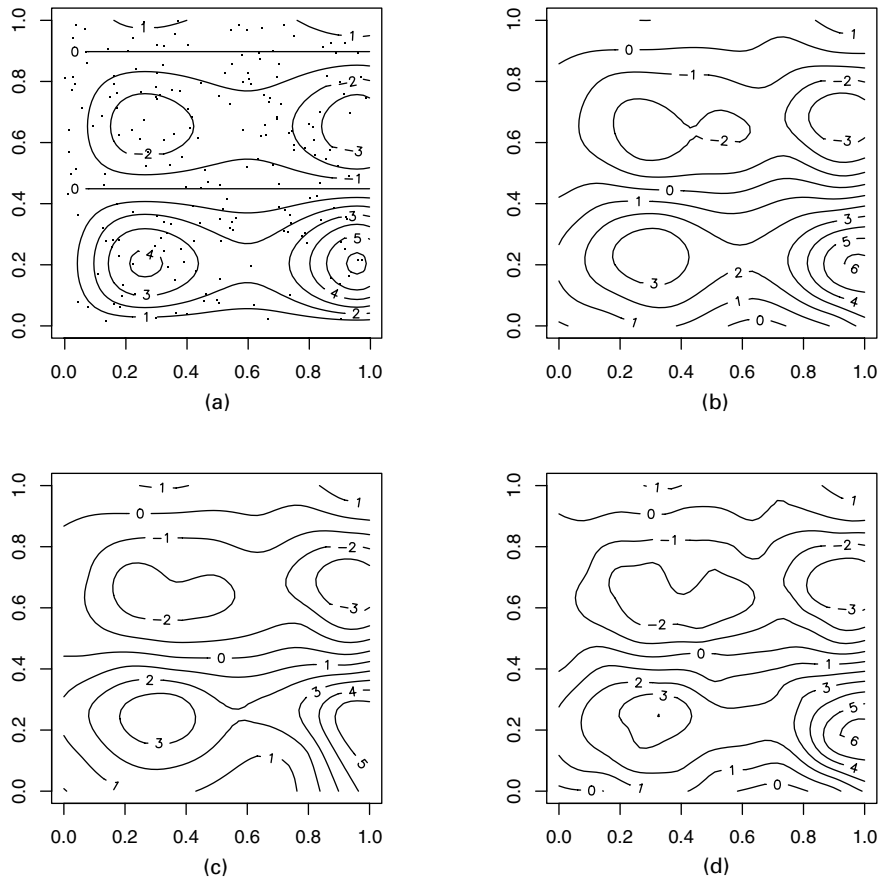


**Fig. 3.** Example reconstructions of the smooth function contoured in (a) from observations of the function taken at the randomly chosen points plotted in (a) and perturbed with zero-mean normal random deviates with standard deviation 0.5: (a) true function; (b) penalized thin plate regression spline reconstruction using a 49-dimensional basis; (c) equivalent reconstruction using a thin plate spline with a 49-knot basis, where the knots are on a regular 7 × 7 grid; (d) full thin plate spline reconstruction (for (b)–(d) GCV was used to select the smoothing parameter)

is already restricted to a space of relatively smooth functions: hence the full thin plate spline is free to overfit the data substantially in a way that the thin plate regresssion spline cannot. For these examples, where the underlying truth is smooth and any possible reduction in bias that might be obtainable by using the full thin plate spline is dwarfed by sampling variability, the thin plate regression spline hence has an advantage.

## 4. Example: generalized additive models

One obvious use of thin plate regression splines is as a way of efficiently incorporating multi-dimensional smoothers into GAMs, in a manner that allows statistically well-founded model selection for such models. GAMs address the problem of modelling response data $y_i$ from an exponential family distribution, in terms of multiple covariates $x_{1i}, x_{2i}, x_{3i}, \ldots,$ by using a model structure of the form

$$g(\mu_i) = \alpha_i + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + \ldots, \tag{11}$$

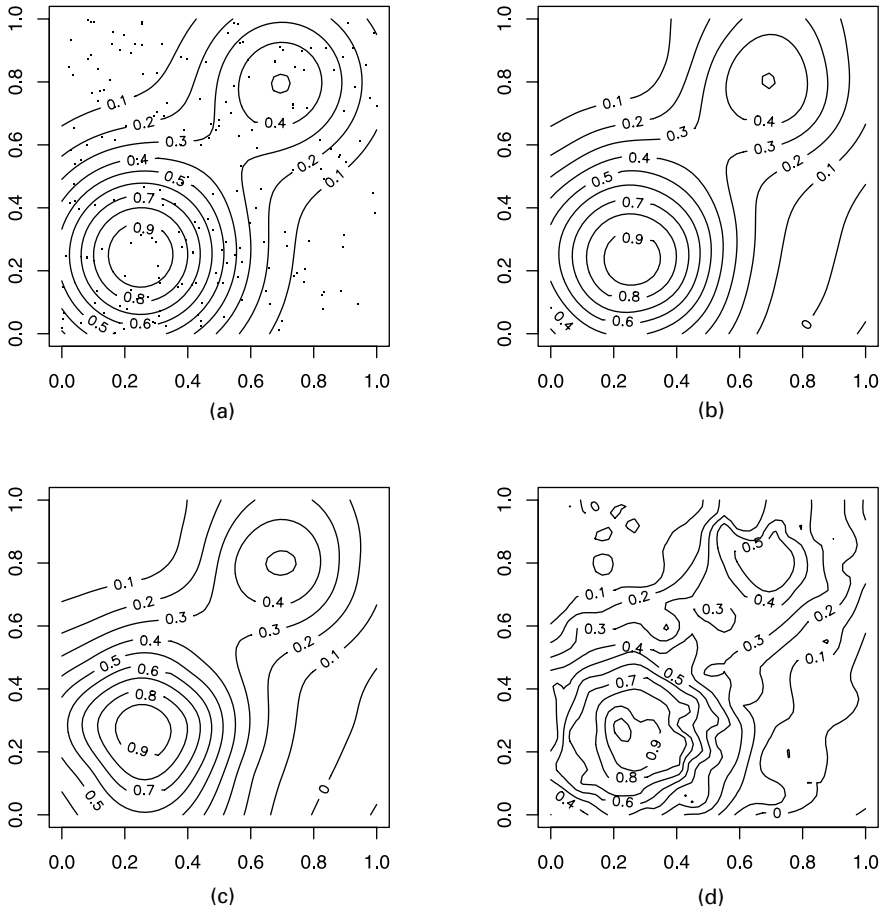


**Fig. 4.** As for Fig. 3, except that the reconstructions are of the function plotted in (a) and both the knot placement basis and the thin plate regression spline basis were of dimension 36 (the noise standard deviation in this case was 0.05): this example shows one of the occasional overfits produced by the full spline model with smoothness selected by GCV

where $\mu_i \equiv E(y_i)$, $g$ is a monotonic link function, the $f$s are smooth functions to be estimated and $\alpha_i$ represents any strictly parametric model components (e.g. a simple constant, or perhaps a linear term in another covariate or whatever), i.e. the key feature of these models is that the mean of the response depends on the covariates through a sum of smooth terms, each of which is a function of only one or a few covariates. Side-conditions are required to ensure identifiability. To date users of these methods have had a choice between the mathematically elegant but computationally very costly generalized smoothing spline approach of Wahba, Gu and co-workers in which the model estimation problem is formulated as a variational problem in an appropriate reproducing kernel Hilbert space (e.g. Wahba (1990), Gu and Wahba (1993), Wahba *et al*. (1995) and Gu (2002)), or of the more *ad hoc*, but much more efficient, methods of Hastie and Tibshirani (1990). The generalized smoothing spline approach has the advantage that it is feasible to select the wiggliness of the components of the GAM by using well-founded criteria such as GCV, generalized maximum likelihood or AIC, and that model inference has a solid theoretical underpinning. To maintain computational efficiency, Hastie and Tibshirani types of GAM rely on more *ad hoc* approaches to model selection and require a slightly less

well-founded approach to inference. In practice the much greater computational efficiency of Hastie and Tibshirani's approach has meant that it is the more widely used.

The thin plate regression splines provide a means by which it should be possible to retain many of the advantages of the generalized smoothing spline approach to GAMs in terms of well-founded model selection and good practical properties deriving from a firm theoretical basis, while also benefiting from the kind of computational efficiency that characterizes Hastie and Tibshirani's approach.

In practice it is straightforward to represent a GAM by producing a thin plate regression spline basis for each model term. Taking the example of model (11) and assuming that the parametric term consists only of a constant, then a design matrix $\mathbf{X}_i$ and wiggliness penalty matrix $\mathbf{S}_i$ would be produced for each component smooth function $f_i$. The design matrix for the whole model is then something like $\mathbf{X} \equiv (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \ldots)$ (where $\mathbf{1}$ is a column of 1s). Writing the parameter vector for the $i$th term as $\boldsymbol{\beta}_i$ then an appropriate side-condition ensuring model identifiability would be that the sum of $f_i$ evaluated over all observed covariate values should be 0, i.e. $\mathbf{1}'\mathbf{X}_i\boldsymbol{\beta}_i = 0$ (where $\mathbf{1}$ is an appropriate column vector of 1s). Writing $\boldsymbol{\beta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots)$ (where $\boldsymbol{\alpha}$ is the vector of parameters of the strictly parametric part of the model $\alpha_i$), we have that the model would be estimated by maximizing the penalized log-likelihood

$$l(\boldsymbol{\beta}) - \tfrac{1}{2}\sum_i \lambda_i \boldsymbol{\beta}_i' \mathbf{S}_i \boldsymbol{\beta}_i, \tag{12}$$

subject to the linear identifiability constraints on $\boldsymbol{\beta}$, where $l$ is the log-likelihood of the model and the penalty terms penalize model components for being wiggly. Expression (12) is solved by penalized iteratively reweighted least squares (see for example Wood (2000)), so that, given the $k$th estimate of the parameter vector, $\boldsymbol{\beta}^{[k]}$, $\boldsymbol{\beta}^{[k+1]}$ is found by solving the weighted penalized least squares problem

$$\text{minimize } \|\mathbf{W}^{[k]}(\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum_i \lambda_i \boldsymbol{\beta}_i' \mathbf{S}_i \boldsymbol{\beta}_i$$

where $\mathbf{z}^{[k]} \equiv \mathbf{X}\boldsymbol{\beta}^{[k]} + \boldsymbol{\Gamma}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$, $\mathbf{W}^{[k]}$ is a diagonal matrix with $W_{ii}^{[k]} \equiv \{g'(\mu_i^{[k]})^2 V_i^{[k]}\}^{-1/2}$, $V_i^{[k]}$ is the variance of $y_i$ according to the estimates $\mu_i^{[k]}$, implied by $\boldsymbol{\beta}^{[k]}$, and $\boldsymbol{\Gamma}^{[k]}$ is a diagonal matrix with $\Gamma_{ii}^{[k]} \equiv g'(\mu_i^{[k]})$. Again, solution is subject to the linear constraints on $\boldsymbol{\beta}$ that ensure identifiability. Estimation of smoothing parameters is performed at each iterate by using a generalization (Wood, 2000) of the multiple smoothing parameter GCV method of Gu and Wahba (1991).

As an example of the use of these methods I modelled some fisheries data that were first analysed using GAMs by Borchers *et al.* (1997). The response data are densities per metre squared of sea surface of mackerel eggs produced per day at each of 634 survey locations, along with covariates at each station. The response data have been gathered from research vessels by hauling sampling nets vertically through the water column. Some preprocessing was done to convert the raw data to egg densities produced per day. The purpose of the surveys is to estimate the total egg production rate, to be able to estimate the total mass of parent fish required to produce this rate. An important part of the estimation process is the modelling of the egg distribution. The method is one of the few feasible ways of assessing the size of fish stock without recourse to commercial fisheries data. The latter tend to suffer from severe biases.

The data to be modelled are shown in Fig. 5(a). Since the modelling of these data is not the primary purpose of this paper, I shall use a relatively unsophisticated error structure and shall not discuss model term selection here; rather I shall assume that the important covariates to consider are seabed depth, distance from the 200 m seabed depth contour, longitude and lati-

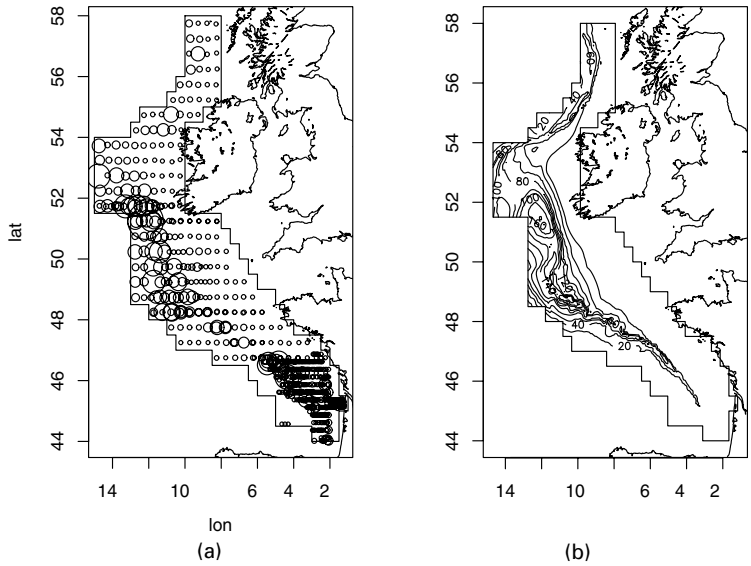


**Fig. 5.** Raw data and fitted model predictions for the mackerel egg GAM example: (a) raw egg density (the areas of the circles are proportional to the densities, with the circles being centred on the sample haul locations); (b) fitted model daily egg production densities

tude. Each of the covariates is available for all 634 egg density estimates. In the original GAM analysis of these data, Borchers *et al.* (1997) modelled egg abundance by using a sum of four univariate smoothers of these four covariates, but there are strong arguments for not modelling the dependence on longitude and latitude in this way, instead using a bivariate smooth function of longitude and latitude (given the isotropic nature of thin plate regression splines there is also an argument for replacing longitude and latitude with co-ordinates on a squarer grid, but to maintain comparability with the analysis of Borchers *et al.* (1997) I have not done that here). Consequently, the model structure used was

$$\sqrt{y_i} = \alpha + f_1(\mathrm{lo}_i, \mathrm{la}_i) + f_2(\mathrm{c.dist}_i) + f_3(\mathrm{b.depth}_i) + \varepsilon_i$$

where the $\varepsilon_i$ are independent and identically distributed normal random variables, $y_i$ is the $i$th observation of egg density produced per day and lo, la, b.depth and c.dist are longitude, latitude, seabed depth and distance to the 200 m contour respectively. $f_2$ and $f_3$ were represented by using rank 10 thin plate regression spline bases, whereas $f_1$ used a rank 50 thin plate regression spline basis. The square-root transform was employed to stabilize variances. Using the `mgcv` package this model was fitted with the command

```
mack.mod<-gam(y^0.5 ~ s(lo,la,k=50)+s(c.dist)+s(b.depth),data=mack)
```

Fig. 6 shows the estimated model terms, whereas Fig. 5(b) contours the fitted model egg density daily production estimates (the plots in Fig. 5 are essentially those produced by using `plot.gam` from `mgcv`: `plot(mack.mod)`, although Fig. 5(b) has been modified). The plotted confidence intervals are obtained by using the approximation that the parameter estimators are normally distributed about their true values, with a covariance matrix that can be estimated as

$$\mathbf{Z}(\mathbf{Z}'\mathbf{X}'\mathbf{W}^2\mathbf{X}\mathbf{Z} + \textstyle\sum \lambda_i \mathbf{Z}'\mathbf{S}_i\mathbf{Z})^{-1}\mathbf{Z}'\hat{\sigma}^2,$$

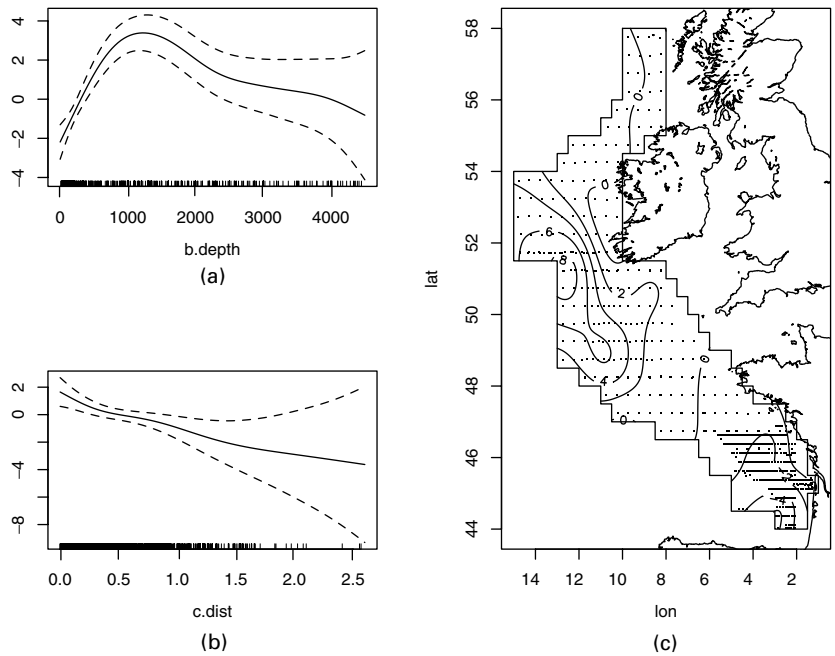


**Fig. 6.** Estimated model terms for the mackerel egg GAM: (a) smooth function of seabed depth and 95% confidence limits; (b) smooth function of distance from the 200 m contour and its 95% confidence limits; contours of the smooth function of longitude and latitude and sample locations

where $\mathbf{Z}$ is a column basis for the null space of the identifiability constraints on the model,

$$\hat{\sigma}^2 = \|\mathbf{W}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\|^2 / \mathrm{tr}(\mathbf{I} - \mathbf{A})$$

and

$$\mathbf{A} \equiv \mathbf{XZ}(\mathbf{Z}'\mathbf{X}'\mathbf{W}^2\mathbf{XZ} + \textstyle\sum \lambda_i \mathbf{Z}'\mathbf{S}_i\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}'\mathbf{W}^2,$$

all quantities being estimated at convergence of the iterative fitting procedure. This is based on a Bayesian argument (see Silverman (1985) and Wood (2000)), and the resulting confidence intervals are quite similar to those proposed by Wahba (1983).

## 5. Discussion

The thin plate regression splines proposed here meet the objectives that were set out in Section 1. For penalized regression modelling they provide optimal low rank approximations to thin plate splines that are both computationally efficient and stable. In pure regression contexts they also provide a way of avoiding the problems of knot placement, while allowing model selection to be carried out by using methods that are dependent on model nesting.

This computational efficiency and stability should be beneficial when non-linear models are employed which contain embedded smooth functions (see Wood (2000), section 5, for example). As demonstrated in Section 4, the method provides a computationally efficient way of incorporating multidimensional smooth terms into GAMs, in a way that facilitates the well-founded approaches to model selection and inference characteristic of the generalized smoothing spline models of Gu, Wahba and co-workers (e.g. Wahba (1990), Wahba *et al.* (1995) and Gu (2002)), but with the considerable practical advantage of greatly enhanced computational efficiency (see

Section 3.2). It would also be straightforward to incorporate thin plate regression spline based smoothers directly into Hastie and Tibshirani's GAM framework, and by using pure (rather than penalized) thin plate regression splines GAMs could also be constructed entirely within a generalized linear model framework.

There are some interesting open questions relating to thin plate regression splines. Firstly, the sense in which a thin plate regression spline is optimal is slightly weak: one could argue that the measures of approximation error are formulated in too large a space. It is not clear whether it is possible to produce more strongly optimal approximations, although the structure of the problem suggests that it is unlikely that stronger results are possible if any computational advantage over full spline models is to be maintained. In response to this a referee suggested concentrating on only one of $\varepsilon_k$ or $e_k$, but working in the space where $\mathbf{T}'\boldsymbol{\delta} = \mathbf{0}$, which implies setting $\boldsymbol{\Gamma}_k$ to the first $k$ right singular vectors of $\mathbf{E}\mathbf{Z}_n$ (ordered by decreasing singular values) or to the first $k$ eigenvectors of $\mathbf{Z}_n'\mathbf{E}\mathbf{Z}_n$ (ordered by decreasing eigenvalues) respectively. This has the appeal that the optimality criterion is now in the 'correct' space, and in limited numerical simulations for the $m = 1, d = 1$ case the results are rather similar to the thin plate regression spline results: the modified $e_k$ basis tends to give slightly more smooth but slightly worse fitting results, whereas the modified $\varepsilon_k$ basis gives slightly better fitting but more wiggly estimates. Computational costs are doubled by use of the modified $\varepsilon_k$ basis (each Lanczos step requiring a vector to be multiplied by $\mathbf{E}\mathbf{Z}_n$ *and its transpose*) but are almost unchanged by using the modified $e_k$ basis. In a pure regression context both alternative bases are even easier to use than the thin plate regression spline basis since reducing the order of the model function is now a simple matter of dropping design matrix columns. However, the minimization of either $\varepsilon_k$ or $e_k$ alone is not a satisfactory way of arriving at an optimal basis in general: minimizing the change in fitted values without regard to a change in the function norm has the potential to lead to very wiggly results, whereas concentrating solely on avoiding big changes in the function norm can result in a poor fit. This is perhaps most readily appreciated by considering a cubic or thin plate spline parameterized directly in terms of the function values $\mathbf{f}$, corresponding to the response data $\mathbf{y}$, so that the fitting problem is to minimize $\|\mathbf{y} - \mathbf{f}\|^2 + \lambda \mathbf{f}'\mathbf{S}\mathbf{f}$, where $\mathbf{S}$ is a positive semidefinite matrix—it is clear that an approach based solely on a suitably modified $\varepsilon_k$ criterion is unlikely to result in a useful truncation here. The likely explanation for the relatively good performance of the bases derived from applying the modified $e_k$ or $\varepsilon_k$ criteria singly to the thin plate spline in its standard parameterization is that the resulting truncated bases in fact almost minimize the neglected criterion in each case, but it is not obvious how to formalize this statement.

Another open question relates to the automatic selection of the basis dimension $k$ in the penalized regression context. If $k$ is not too small then model results should be rather insensitive to its value: practical experience to date suggests that this is so, in which case the pragmatic approach suggested in Section 2.1 is a reasonable approach to adopt. Nevertheless it would be more satisfactory to have some theoretical guidance on this point. A further open issue is the question of when we might want to use a knot-based penalized regression spline, rather than the eigenbasis proposed here. Purely on computational grounds a knot-based scheme should be more efficient because it does not require a truncated eigendecomposition to be obtained. In principle the computational saving could be 'spent' on more knots, in which case the knot-based scheme might approach or exceed the eigenbased scheme in terms of MSE performance: however, for models with multiple smoothing parameters or those requiring an iteratively reweighted least squares method for fitting, I found that the eigenbased models tended to converge faster than the knot-based models for a given basis dimension. If this is a general phenomenon then it clearly deserves further study. In any case the optimal solution in terms of performance for a given amount of computing effort is probably to select a set of knots that is intermediate in size

between the number of data and the desired number of parameters and to obtain a thin plate regression spline basis from this set of knots.

Finally there is the issue of anisotropy. A thin plate regression spline is an isotropic smoother, which is appropriate for spatial co-ordinates, for example, but may not be as suitable if the arguments of the smoother are covariates measured in different units. One could adopt the approach, taken with many smoothing methods, of rescaling covariates to lie in the unit square, cube or hypercube, but this is essentially arbitrary (when used with all methods, not just splines). In principle the problem could be approached in a non-arbitrary way, by treating the relative scaling of axes as extra smoothing parameters in the problem (see for example Wood (2000)), and work on the production of a general method for doing this is on going.

Thin plate regression splines are available as part of the author's R package `mgcv`, available from `www.cran.r-project.org`. The package includes full source code in C.

## Acknowledgements

I am particularly grateful to a referee for numerous helpful and thought-provoking suggestions which improved the paper, and to another referee for pointing me in the direction of some relevant papers in the machine learning literature.

## Appendix A: Implementation by using standard software

Here are the steps required to construct a rank $k$ basis for smoothing.

(a) Form the $n \times n$ matrix $\mathbf{E}$ and the $n \times M$ matrix $\mathbf{T}$ defined in Section 2.
(b) Obtain the truncated spectral decomposition $\mathbf{E}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{U}_k'$, by the use of any standard eigen-routines to find the full spectral decomposition of $\mathbf{E}$.
(c) Using standard routines, form the $QR$-decomposition $\mathbf{QR} = \mathbf{U}_k' \mathbf{T}$ where the last $n - M$ rows of $\mathbf{R}$ are 0 and $\mathbf{Q}$ is orthonormal. Then the final $n - M$ columns of $\mathbf{Q}$ give $\mathbf{Z}_k$, the basis for the null space of the equality constraints. If efficiency matters then $\mathbf{Z}_k$ can be stored as $M$ Householder rotations (see for example Watkins (1991)).
(d) Writing the parameter $k$-vector of the thin plate regression spline as $\boldsymbol{\beta} = (\tilde{\boldsymbol{\delta}}', \boldsymbol{\alpha}')'$, then the $n \times k$ design matrix for the thin plate regression spline is $\mathbf{X} = (\mathbf{U}_k \mathbf{D}_k \mathbf{Z}_k, \mathbf{T})$. Similarly the penalty matrix for using this thin plate regression spline in penalized regression would be

$$\mathbf{S} = \begin{pmatrix} \mathbf{Z}_k' \mathbf{D}_k \mathbf{Z}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where the padding with zero matrices is for notational convenience.
(e) To fit a pure thin plate regression spline to response data $\mathbf{y}$, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is minimized with respect to $\boldsymbol{\beta}$, whereas the incorporation of the thin plate regression spline into any generalized linear model is simply a matter of incorporating the thin plate regression spline design matrix into the generalized linear model design matrix.
(f) To fit a penalized thin plate regression spline requires minimization of

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}' \mathbf{S} \boldsymbol{\beta}$$

with respect to $\boldsymbol{\beta}$, given a value for the smoothing parameter $\lambda$.

The evaluation of any bounded linear functional of $g$ is self-evidently linear in $(\tilde{\boldsymbol{\delta}}', \boldsymbol{\alpha}')'$, with the coefficients easily obtained given $\mathbf{Z}_k$, $\mathbf{D}_k$ and $\mathbf{U}_k$. For example $g(\mathbf{x})$ can be written $g(\mathbf{x}) = \mathbf{a}' \tilde{\boldsymbol{\delta}} + \mathbf{b}' \boldsymbol{\alpha}$ where $a_i$ and $b_i$ are known coefficients, depending only on $\mathbf{x}$:

$$a_i = \sum_{j=1}^{n} \eta_{md}(\|\mathbf{x} - \mathbf{x}_j\|) (\mathbf{U}_k \mathbf{Z}_k)_{ji},$$

$$b_i = \phi_i(\mathbf{x}).$$

## Appendix B: Lanczos iteration

(In this appendix only $\mathbf{E}$ refers to the same quantity that it refers to in the main body of the paper.) The Lanczos algorithm is iterative, and at the $i$th iteration produces an $(i \times i)$ symmetric tridiagonal matrix ($\mathbf{T}_i$, say), the eigenvalues of which approximate the $i$ largest magnitude eigenvalues of the original matrix: these eigenvalues converge as the iteration proceeds, with those of largest magnitude converging first. The eigenvalues and vectors of $\mathbf{T}_i$ can be obtained in order $i^2$ operations (using the usual $QR$-algorithm to find the eigenvalues and then inverse iteration to find the eigenvectors); however, the inverse iteration appears to be insufficiently stable in some cases, so it is probably preferable simply to accumulate the eigenvectors as part of the $QR$-algorithm at a cost of order $i^3$. The eigenvectors of the original matrix are easily obtained from the eigenvectors of $\mathbf{T}_i$. A complete version of the algorithm, which is suitable for finding the truncated decomposition of $\mathbf{E}$, is as follows.

(a) Let $\mathbf{b}$ be an arbitrary non-zero $n$-vector: it may be best to initialize this from a simple random-number generator, to reduce the risk of starting out orthogonally to some eigenvector (exact repeatability can be ensured by starting from the same random-number generator seed).
(b) Set $\mathbf{q}_1 \leftarrow \mathbf{b}/\|\mathbf{b}\|$.
(c) Repeat steps (d)–(l) for $j = 1, 2, \ldots$ until enough eigenvectors have converged.
(d) Form $\mathbf{z} \leftarrow \mathbf{E}\mathbf{q}_j$.
(e) Calculate $\alpha_j \leftarrow \mathbf{q}_j'\mathbf{z}$.
(f) Reorthogonalize $\mathbf{z}$ to ensure numerical stability, by performing the following step *twice*:

$$\mathbf{z} \leftarrow \mathbf{z} - \sum_{i=1}^{j-1} (\mathbf{z}'\mathbf{q}_i)\mathbf{q}_i.$$

(g) Set $\beta_j \leftarrow \|\mathbf{z}\|$.
(h) Set $\mathbf{q}_{j+1} \leftarrow \mathbf{z}/\beta_j$.
(i) Let $\mathbf{T}_j$ be the $(j \times j)$ tridiagonal matrix with $\alpha_1, \ldots, \alpha_j$ on the leading diagonal and $\beta_1, \ldots, \beta_{j-1}$ on the leading subdiagonals and superdiagonals.
(j) If iteration has proceeded sufficiently far to make it worthwhile, find the eigendecomposition (spectral decomposition) $\mathbf{T}_j = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$, where columns of $\mathbf{V}$ are eigenvectors of $\mathbf{T}_j$ and $\mathbf{\Lambda}$ is diagonal with eigenvalues on the leading diagonal.
(k) Compute 'error bounds' for each $\Lambda_{i,i}$: $|\beta_j V_{i,j}|$.
(l) Use the error bounds to test for convergence of the $k$ largest magnitude eigenvalues. Terminate the loop if all have converged.
(m) The $i$th eigenvalue of $\mathbf{E}$ is $\Lambda_{i,i}$. The $i$th eigenvector of $\mathbf{E}$ is $\mathbf{Q}\mathbf{v}_i$, where $\mathbf{Q}$ is the matrix whose columns are the $\mathbf{q}_j$ (for all $j$ calculated) and $\mathbf{v}_i$ is the $i$th column of $\mathbf{V}$ (again calculated at the final iteration). Hence $\mathbf{D}_k$ and $\mathbf{U}_k$ can easily be formed.

This algorithm is stabilized by orthogonalization against all previous vectors $\mathbf{q}_j$: several selective orthogonalization schemes have been proposed to reduce the computational burden of this step, but I experienced convergence problems when trying to use these schemes with $\mathbf{E}$, especially in the one-dimensional case ($d = 1$): in any case the computational cost of the method is dominated by the $O(n^2)$ step $\mathbf{z} \leftarrow \mathbf{E}\mathbf{q}_j$, so the efficiency benefits of using a selective method are unlikely to be very great, in the current case (if $\mathbf{E}$ were sparse then selective methods would offer a benefit).

Finally note that $\mathbf{E}$ need never be formed and stored as a whole: it is only necessary that its product with a vector can be formed. It is this feature that suggests that thin plate regression splines could be used on very large data sets without causing storage problems. For a fuller treatment of the Lanczos method see Demmel (1997), from which the algorithm given here has been modified.

## References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

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.

Demmel, J. (1997) *Applied Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

Duchon, J. (1977) Splines minimizing rotation-invariant semi-norms in Solobev spaces. In *Construction Theory of Functions of Several Variables*. Berlin: Springer.
Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89–121.
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. (2002) *Smoothing Spline ANOVA Models*. New York: Springer.
Gu, C. and Wahba, G. (1991) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, **12**, 383–398.
———(1993) Semiparametric analysis of variance with tensor product thin plate splines. *J. R. Statist. Soc.* B, **55**, 353–368.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hutchinson, M. F. and de Hoog, F. R. (1985) Smoothing noisy data with spline functions. *Numer. Math.*, **47**, 99–106.
Parker, R. L. and Rice, J. A. (1985) Discussion on 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting' (by B. W. Silverman). *J. R. Statist. Soc.* B, **47**, 40–42.
Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Statist. Soc.* B, **47**, 1–52.
Smola, A. J. and Schölkopf, B. (2000) Sparse greedy matrix approximation for machine learning. In *Proc. 17th Int. Conf. Machine Learning*. San Francisco: Morgan Kaufmann.
Wahba, G. (1980) Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In *Approximation Theory III* (ed. W. Cheney), pp. 905–912. New York: Academic Press.
———(1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
———(1990) Spline models for observational data. *CBMS–NSF Regl 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**, 1865–1895.
Watkins, D. S. (1991) *Fundamentals of Matrix Computation*. New York: Wiley.
Williams, C. K. I. and Seeger, M. (2001) Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, vol. 13 (eds T. K. Leen and T. G. Diettrich). Cambridge: MIT Press.
Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc.* B, **62**, 413–428.
Zhou, S. and Shen, X. (2001) Spatially adaptive regression splines and accurate knot selection schemes. *J. Am. Statist. Ass.*, **96**, 247–259.