ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Generalized Additive Models

Simon N. Wood

School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom;
email: simon.wood@ed.ac.uk

Annu. Rev. Stat. Appl. 2025. 12:497–526

First published as a Review in Advance on October 7, 2024

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-112723-034249

Keywords

cross validation, marginal likelihood, regression, smooth, smoothing parameter

Abstract

Generalized additive models are generalized linear models in which the linear predictor includes a sum of smooth functions of covariates, where the shape of the functions is to be estimated. They have also been generalized beyond the original generalized linear model setting to distributions outside the exponential family and to situations in which multiple parameters of the response distribution may depend on sums of smooth functions of covariates. The widely used computational and inferential framework in which the smooth terms are represented as latent Gaussian processes, splines, or Gaussian random effects is reviewed, paying particular attention to the case in which computational and theoretical tractability is obtained by prior rank reduction of the model terms. An empirical Bayes approach is taken, and its relatively good frequentist performance discussed, along with some more overtly frequentist approaches to model selection. Estimation of the degree of smoothness of component functions via cross validation or marginal likelihood is covered, alongside the computational strategies required in practice, including when data and models are reasonably large. It is briefly shown how the framework extends easily to location-scale modeling, and, with more effort, to techniques such as quantile regression. Also covered are the main classes of smooths of multiple covariates that may be included in models: isotropic splines and tensor product smooth interaction terms.

1. INTRODUCTION

A generalized additive model (GAM; Hastie & Tibshirani 1986, 1990) is a regression model in which the mean or location parameter of the response variable depends on a sum of smooth functions of covariates. Originally GAMs grew out of work on smoothing (e.g., Kimeldorf & Wahba 1970, Silverman 1985, O'Sullivan et al. 1986, Wahba 1990) and referred to generalized linear models (GLMs) (Nelder & Wedderburn 1972) in which the linear predictor depends on a sum of unknown smooth functions of covariates, those smooth functions being the target of inference. The GLM restriction to exponential family distributions no longer applies. Indeed, the restriction to modeling only the location parameter has also been abandoned, and it is common to model multiple parameters of the response distribution using sums of smooth functions of predictors (e.g., Yee & Wild 1996, Rigby & Stasinopoulos 2005, Mayr et al. 2012, Yee 2015, Wood et al. 2016, Stasinopoulos et al. 2017).

The original approach to GAM estimation used the elegant backfitting algorithm (Hastie & Tibshirani 1990), in which smooth components of a model are estimated by iterative nonparametric smoothing of partial residuals with respect to covariates. But well-founded estimation of the degree of smoothness of the components is difficult with backfitting. An alternative (Gu & Wahba 1991, Gu 1992) used direct fitting with full smoothing spline models, allowing smoothness estimation via cross validation (Craven & Wahba 1979) or marginal likelihood maximization (Wahba 1985), but at $O(n^3)$ computational cost for n data.

It later became clear that many of the advantages of the spline approach to GAMs, including smoothness estimation, could be retained at much lower cost by making use of rank reduced splines (e.g., Wood 2000). This article reviews the framework that then emerges, developing it from basic one-dimensional penalized regression. The framework underpins the R package `mgcv` and (in whole or part) several other software packages. It is also very widely used in applied statistical work. Readers are directed to Wood (2017a) for further detail on some topics covered below plus many applications.

2. SMOOTHING IN ONE DIMENSION

The key concepts underpinning modeling with smooth functions are most easily grasped by considering a one-dimensional model for x_i, y_i data,

$$y_i = f(x_i) + \epsilon_i,$$

where f is an unknown function and the ϵ_i are independent zero mean random variables, with variance σ^2. The obvious way to estimate this would be to seek f to minimize the residual sum of squares $\|\mathbf{y} - \mathbf{f}\|^2$, where $\mathbf{y}$ and $\mathbf{f}$ are the vectors of y_i and $f(x_i)$, respectively. But if f can be any function, this is clearly neither identifiable nor desirable: Any interpolant of the data would reduce the objective to zero. Fitting the noise in this way seldom makes statistical sense, even if it can be done uniquely.

Instead we might restrict attention to functions that are smooth according to some appropriate measure. For example, we might measure lack of smoothness (wiggliness) using the cubic spline penalty $\int f''(x)^2 \mathrm{d}x$, the integrated squared second derivative of f. Then we might seek $\hat{f}$ to minimize

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda \int f''(x)^2 \mathrm{d}x, \qquad 1.$$

where the smoothing parameter, λ, controls the trade-off between smoothness and fidelity to the data. We restrict our search to the space of functions for which the spline penalty is well defined, but that is no practical restriction if we are looking for smooth functions.

The result of this optimization is known as a cubic smoothing spline. Conceptually there are a number of ways that we might compute $\hat{f}$. An interesting starting point is brute force numerical discretization, as this makes clear the rather fruitful link between smoothing and the solution of partial differential equations (Heckman & Ramsay 2000, Ramsay 2002, Ramsay et al. 2007, Wood et al. 2008, Lindgren et al. 2011, Ettinger et al. 2016, Sangalli 2021). Without loss of generality assume that the x_i are arranged in ascending order from x_1 to x_n. Then define $f_j = f(x_1 + jh - h)$, where $h = (x_n - x_1)/(m - 1)$, for some relatively large m. That is, the f_j represent $f(x)$ evaluated on a fine grid over the range of the x data. The finite difference approximation to the penalty is therefore

$$\int f''(x)^2 \mathrm{d}x \simeq \sum_{j=2}^{m-1} (f_{j-1} - 2f_j + f_{j+1})^2/h;$$

note that the difference terms for $j = 1$ and m are not needed, as the minimizing f_0 and f_{m+1} would have to zero these. To avoid irrelevant complication, assume that $x_i = x_1 + j(i)h - h$, where $j(i)$ is an integer between 1 and m; i.e., the x_i correspond to evaluation grid points. So Expression 1 is now

$$\sum_{i=1}^{n} \left(y_i - f_{j(i)}\right)^2 + \lambda \sum_{j=2}^{m-1} (f_{j-1} - 2f_j + f_{j+1})^2/h.$$

Differentiating with respect to f_k and equating to zero, we have that if $k = j(i)$ for some i then[1]

$$-2y_i + 2f_{j(i)} + \lambda(12f_k - 8f_{k-1} - 8f_{k+1} + 2f_{k-2} + 2f_{k+2})/h = 0, \qquad 2.$$

and otherwise,

$$\lambda(12f_k - 8f_{k-1} - 8f_{k+1} + 2f_{k-2} + 2f_{k+2})/h = 0. \qquad 3.$$

Equations 2 and 3 define a sparse system of equations $\mathbf{Af} = \tilde{\mathbf{y}}$, where $\mathbf{A}$ is a sparse matrix of coefficients and the elements of $\tilde{\mathbf{y}}$ are either zero or a y_i value, as appropriate. The term in parentheses is also a finite difference estimate of $h^4 f''''(x_k)$, so Equations 2 and 3 can be interpreted as defining a differential equation. Clearly the solution (which turns out to exist) has an undefined fourth derivative at the x_i and zero fourth derivative between the x_i, corresponding to a piecewise constant third derivative, a piecewise linear second derivative, and hence a piecewise cubic $\hat{f}(x)$.

Figure 1 shows the result of solving the sparse system/discretized differential equation for some test data, with an intermediate λ value. Repeating the exercise for $\lambda \to 0$ results in a wiggly $\hat{f}(x)$ that interpolates the data, while with a sufficiently large λ a straight line fit is produced, corresponding to the least squares regression line. Solving the sparse system is efficient, but there is a small approximation error associated with the discretization, and if $\hat{f}(x)$ is to be evaluated between grid points then a further approximation is needed, such as linearly interpolating between the gridded $\hat{f}$ values.

Alternatively, the differential equation implies that $\hat{f}$ is made up of piecewise cubic sections between adjacent x_i points, with the sections joined to be continuous up to the second derivative. It is not difficult to create a basis for such piecewise cubic functions (see Section 2.1). Let $b_i(x)$ denote the basis functions. Now writing $f(x) = \sum_i \beta_i b_i(x)$, we can rewrite Expression 1 exactly as

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

[1] The expressions are slightly different for the first two and last two grid points.

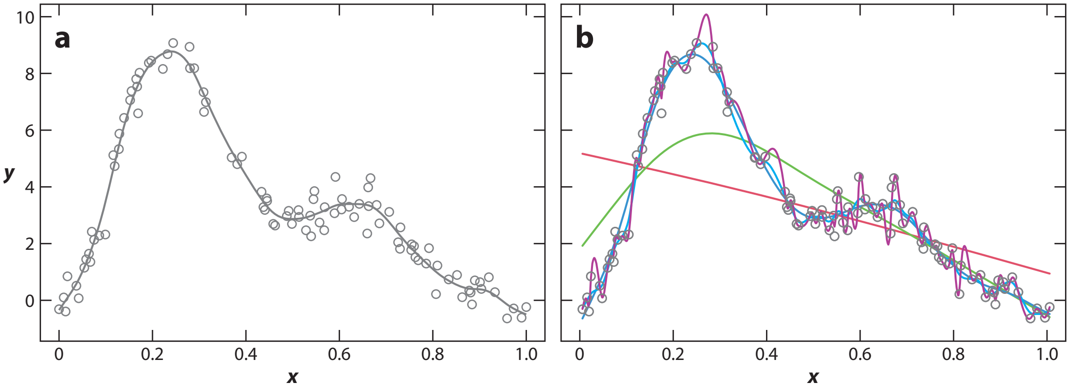

Figure 1

(*a*) A cubic smoothing spline fitted to the data shown, as computed by solving the sparse matrix system (discretized differential equation) given in the text, using a 1,000-point grid and intermediate λ. (*b*) Smoothing splines as λ is varied from very high (*red straight line*) to very low (near interpolation).

where $X_{ij} = b_j(x_i)$ and $S_{ij} = \int b_i''(x) b_j''(x) \mathrm{d}x$. The latter integral can be evaluated analytically. Hence, optimization of Expression 1 reduces to optimization of an n-dimensional system, with formal solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. Using this approach to smooth the data in **Figure 1** gives results visually indistinguishable from the smooths shown, for the same λ values.

The link between smooths (a.k.a. smoothers) and differential equations generalizes to different integrated derivative penalties, and to smoothing with respect to multiple covariates (although some care is required to ensure that the solution of the partial differential equation actually exists in higher dimensions—some appealing penalties fail to guarantee this). Similarly the computational choice between rather direct solution of sparse approximations and lower-dimensional, but less sparse, basis expansion representations also generalizes. Both offer fruitful strategies for GAM computation in different, if overlapping, settings.

2.1. Basis Construction

To understand basis construction, again consider the cubic spline case. We need a basis for piecewise cubic functions that are continuous to second derivative (C^2 functions). The third derivative changes discontinuously where the cubic sections join. That is, we seek known functions $b_j(x)$ such that the spline can be written as $f(x) = \sum_j \beta_j b_j(x)$, where the β_j are to be estimated. If $f(x)$ is piecewise cubic, then so must be the $b_j(x)$. Suppose that the cubic sections join at knots $x_1 < x_2 < \cdots$. An obvious basis would consist of $b_j(x)$ that were C^2 piecewise cubic with $b_j(x_i) = 0$ for all $i \neq j$ and $b_j(x_j) = 1$ (plus the condition of zero second derivatives at the first and last knot). Obviously $f(x)$ is then a spline for which the coefficients have the interpretation $\beta_j = f(x_j)$. Such cardinal basis functions can be constructed algebraically, or by solving the differential equation $b_j''''(x) = 0$ subject to the given conditions on the solution. A 10-dimensional example is shown in **Figure 2*a***.

An appealing property for a basis is compact support, meaning that each basis function is only nonzero over some limited x range. For a C^2 piecewise cubic to achieve this, it must have $b_j = b_j' = b_j'' = 0$ at the first and last knots of its nonzero interval, in order to join correctly to the sections where $b_j(x) = 0$. To obtain a nonzero $b_j(x)$ within the nonzero interval, it suffices to impose the condition that $b_j(x) = 1$ at the interval midpoint. So we have seven constraints defining b_j. A C^2 piecewise cubic over four intervals is uniquely defined by seven constraints: the cubic over the first section is defined by four parameters, plus there is an extra

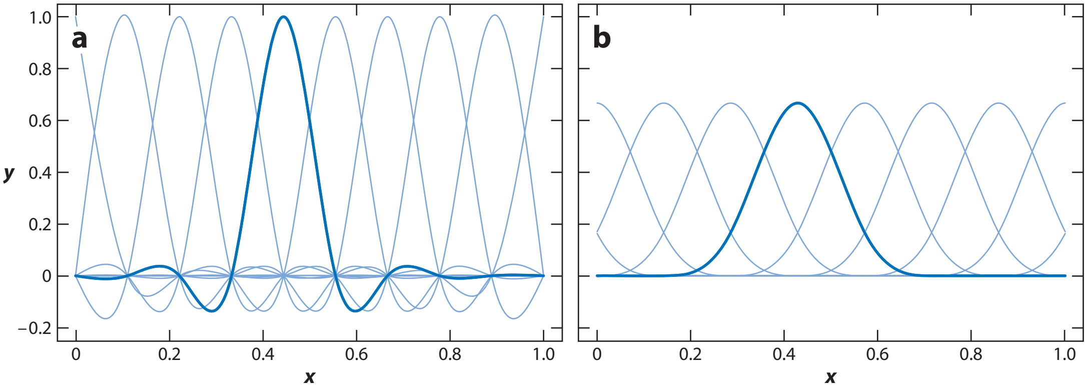

Figure 2

(*a*) Cardinal cubic spline basis of dimension 10. (*b*) B-spline basis of dimension 10, scaled so that at any x value the basis functions sum to one. In both panels the knot locations correspond to the basis function maxima. Bases can be computed just as well for unevenly spaced knots.

parameter for the change in third derivative at each of the three interior knots. So the desired b_j must be constructed by four piecewise cubic sections, and the conditions can be written as $b_j(x_{j-2}) = b'_j(x_{j-2}) = b''_j(x_{j-2}) = b_j(x_{j+2}) = b'_j(x_{j+2}) = b''_j(x_{j+2}) = 0$ and $b_j(x_j) = 1$. Again $b_j(x)$ can be obtained algebraically, or by solving the differential equation $b''''_j(x) = 0$ over the four intervals, subject to the given conditions. The results are known as B-splines[2] (for more detail, see **Figure 2** and de Boor 2001). The evaluation of the penalty matrix elements $S_{ij} = \int b''_i(x)b''_j(x)\mathrm{d}x$ is rather straightforward, as the $b''_j(x)$ are piecewise linear and nonzero over only four intervals, resulting in a banded **S**.

2.2. Rank Reduction

The spline basis function approach allowed exact minimization of Expression 1 while reducing the spline problem to an n-dimensional optimization. But an n-parameter model for n data seems statistically wasteful, especially when the statistician will typically have no interest in allowing sufficient flexibility that n degrees of freedom are used, leading to exact interpolation of noisy data. Could we use fewer than n of these basis functions for piecewise cubic functions? Addressing this requires consideration of the approximation error of splines used as interpolants, as this gives a lower bound for the minimum achievable error (see de Boor 2001).

Consider a cubic spline, $f(x)$, used to interpolate x_i, y_i data, where $y_i = g(x_i)$, and the true function g has at least four continuous bounded derivatives. Without loss of generality consider an interval between $x_i = 0$ and $x_{i+1} = h$. The interval is not the first or the last. Within this interval, the difference between f and the true g can be expressed as

$$f(x) - g(x) = \{f'(0) - g'(0)\}x + \{f''(0) - g''(0)\}x^2/2 + \{f'''(0) - g'''(0)\}x^3/6 + O(x^4),$$

and the $O(x^4)$ term involves only g since f is cubic. Because f interpolates g, $f(h) - g(h) = 0$, so

$$\{f'(0) - g'(0)\}h + \{f''(0) - g''(0)\}h^2/2 + \{f'''(0) - g'''(0)\}h^3/6 + O(h^4) = 0,$$

[2] To see why the unweighted B-splines sum to a constant function, note that the sum of piecewise cubics is piecewise cubic. At each knot the B-splines clearly sum to the same constant and their derivatives obviously sum to zero. The constant value and zero derivative at each knot constrain the piecewise cubic to be constant.

and hence the first three terms must be $O(h^4)$. Since $x = O(h)$, this implies that $|f(x) - g(x)| = O(h^4)$ within the interval. The same applies for any other interval except those at the end, where the end conditions $f''(x_1) = f''(x_n) = 0$ will lower the order in the above argument to $O(h^2)$.

Now suppose that we perform regression for n noisy x_i, y_i data points, using the basis for a cubic spline made up of $k - 1$ cubic segments of equal width, h, where $k < n$. Clearly the error in the regression estimate cannot be less than the spline approximation error $O(h^4) = O(k^{-4})$. We also need to consider the error arising from regression with noisy data, which has standard deviation $O(\sqrt{k/n})$, assuming a nonpathological asymptotic process in which observation density builds up at similar rates across the function domain. Hence, to avoid either the approximation error or statistical error from dominating, $k \propto n^{1/9}$ is required, leading to a lowest achievable error rate of $O(n^{-4/9})$. Although only asymptotic, this result suggests that $k \ll n$ is justifiable.

In fact $k \propto n^{1/9}$ is too slow to be a sensibly informative regime for penalized regression. Careful consideration of the optimization problem shows that $k \propto n^{1/9}$ leads to unpenalized regression when n is large, and $k \propto n^{1/5}$ is required to avoid this (Claeskens et al. 2009). This latter regime is more meaningful, as in practice the statistician is likely to view near-zero smoothing parameter estimates as an indication of an overrestrictive k, and to increase k to avoid this. That said, the finite-sample implication that we can use $k \ll n$ is unchanged.

2.3. Eigen Rank Reduction, Smoothing Bias, and Effective Degrees of Freedom

The simplest approach to rank reduction is to select $k \ll n$ nicely spread out knots from the range of x_i to set up a spline basis. An alternative is to reparameterize so that the evaluated basis functions (the columns of $\mathbf{X}$) are orthogonal and the penalty matrix is diagonal, and then to retain the k least heavily penalized components. Specifically, we might form the QR decomposition $\mathbf{QR} = \mathbf{X}$ and then the orthogonalized[3] eigen decomposition $\mathbf{UDU}^{\mathsf{T}} = \mathbf{R}^{-\mathsf{T}}\mathbf{SR}^{-1}$. Reparameterizing so that $\boldsymbol{\beta}' = \mathbf{U}^{\mathsf{T}}\mathbf{R}\boldsymbol{\beta}$ gives $\mathbf{X}' = \mathbf{QU}$, which has orthogonal columns, while $\mathbf{S}' = \mathbf{D}$ is diagonal. If we retain the columns of $\mathbf{QU}$ corresponding to the k smallest leading diagonal elements of $\mathbf{D}$, then we retain the k smoothest orthogonal components of the model. **Figure 3*a*** shows a truncated eigen basis of rank 12 for the spline in **Figure 1**. **Figure 3*b*** illustrates a fit using this basis.

Since $\hat{\boldsymbol{\beta}}' = (\mathbf{I} + \lambda\mathbf{D})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{y}$, its covariance matrix is $(\mathbf{I} + \lambda\mathbf{D})^{-2}\sigma^2$. It follows that there is also a principal component analysis interpretation of the truncation. The covariance matrix of $\hat{\mathbf{f}}$ is $\mathbf{QU}(\mathbf{I} + \lambda\mathbf{D})^{-2}\mathbf{U}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\sigma^2$, which is an eigen decomposition in which the columns of $\mathbf{QU}$ are eigenvectors and $(\mathbf{I} + \lambda\mathbf{D})^{-2}\sigma^2$ is a diagonal matrix of eigenvalues. Hence retention of the k smallest diagonal elements of $\mathbf{D}$, and corresponding $\mathbf{QU}$ columns, retains the k highest-variance principal components of $\hat{\mathbf{f}}$. The same applies in the Bayesian setting discussed below, since then the posterior covariance matrix of $\mathbf{f}$ turns out to be $\mathbf{QU}(\mathbf{I} + \lambda\mathbf{D})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\sigma^2$.

Now consider either the original or the truncated version of the basis and penalty. Without penalization we would have $\hat{\boldsymbol{\beta}}' = \mathbf{U}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{y}$, so the diagonal elements of $(\mathbf{I} + \lambda\mathbf{D})^{-1}$ are directly interpretable as shrinkage factors mapping the unpenalized coefficients to the penalized versions. This suggests using their sum, $\text{tr}\{(\mathbf{I} + \lambda\mathbf{D})^{-1}\}$, as the effective degrees of freedom of $\hat{f}$. When $\lambda \to \infty$ this is just the number of zero values on the leading diagonal of $\mathbf{D}$, the number of zero eigenvalues of $\mathbf{S}$—2 for the cubic spline. When $\lambda = 0$ it is the dimension of $\boldsymbol{\beta}$, the

[3] While any symmetric matrix can be diagonalized by an orthogonal matrix, the eigenvectors of repeated eigenvalues need not be orthogonal and $\mathbf{R}^{-\mathsf{T}}\mathbf{SR}^{-1}$ has as many zero eigenvalues as the dimension of the space of unpenalized functions. Most symmetric eigen routines will return orthogonalized eigenvectors, but in any case orthogonalization can be achieved by replacing the eigenvectors corresponding to a duplicated eigenvalue by the $\mathbf{Q}$ factor from their QR decomposition.

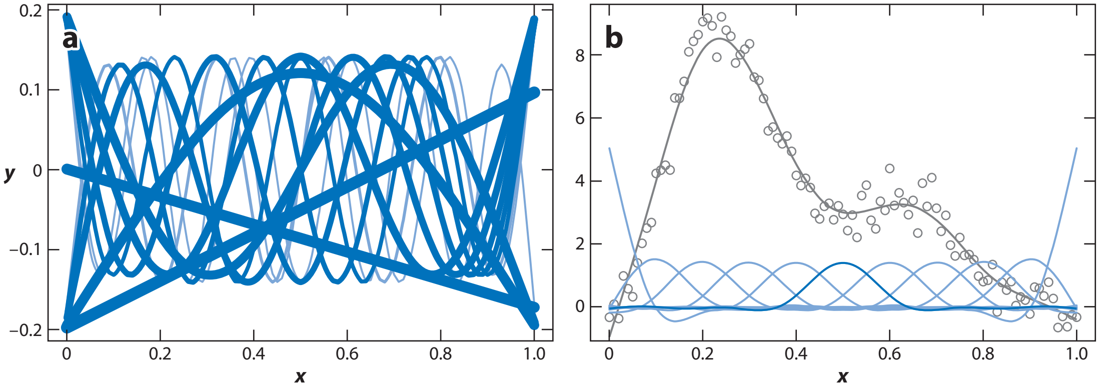

Figure 3

(*a*) An orthogonal eigen basis of rank 12. The evaluated basis functions (columns of **X**) are plotted against x, with thinner lines corresponding to less penalization, and hence higher wiggliness. (*b*) An 8.2 effective degrees of freedom fit to the data from **Figure 1** using the left-hand basis and corresponding diagonal penalty. Also shown are the equivalent kernels corresponding to every 10th datum, multiplied by 20 for plotting. The sum of the equivalent kernels multiplied by the y data gives the fitted curve shown.

unpenalized degrees of freedom. Since the unpenalized covariance matrix of $\hat{\boldsymbol{\beta}}'$ is $\mathbf{I}\sigma^2$, another interpretation of the diagonal elements of $(\mathbf{I} + \lambda\mathbf{D})^{-1}$ is that they are ratios of the Bayesian posterior variances of the coefficients under penalization to their variances without penalization. It is easy to show that in the original parameterization, the effective degrees of freedom is $\text{tr}(\mathbf{F})$, where $\mathbf{F} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}$. By considering the trace of the identity matrix, it is also obvious that $\text{tr}(\mathbf{I} - \mathbf{F}) = \text{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{S}\lambda\}$ degrees of freedom are suppressed by penalization. Furthermore, $\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbf{F}\boldsymbol{\beta}$, so the bias of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}$.

Also note that $\text{tr}(\mathbf{F}) = \text{tr}(\mathbf{A})$, where $\mathbf{A} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}$ is the influence or hat matrix of the spline, such that $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$. The columns of **A** can be viewed as equivalent kernels, if spline smoothing is viewed as a form of kernel smoothing (see **Figure 3*b***). Rewriting **A** in terms of the eigen basis and premultiplying $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$ by $\mathbf{X}'^{\mathsf{T}} = \mathbf{U}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}$ gives

$$\mathbf{X}'^{\mathsf{T}}\hat{\mathbf{f}} = (\mathbf{I} + \lambda\mathbf{D})^{-1}\mathbf{X}'^{\mathsf{T}}\mathbf{y}. \qquad 4.$$

This implies that for the unpenalized columns of $\mathbf{X}'$, which are the last two for the cubic spline example, we have $\mathbf{X}'[,(k-1):k]^{\mathsf{T}}\hat{\mathbf{f}} = \mathbf{X}'[,(k-1):k]^{\mathsf{T}}\mathbf{y}$. In particular, if $\mathbf{1}$ is in the range of the unpenalized columns, as it is for a spline, then $\mathbf{1}^{\mathsf{T}}\hat{\mathbf{f}} = \mathbf{1}^{\mathsf{T}}\mathbf{y}$, despite penalization.

The QR plus eigen scheme is an expensive rank reduction method. A hybrid approach reduces the cost: n_b knots, where $k \ll n_b \ll n$, are randomly selected from the x_i and used to construct a basis and penalty. The orthogonalizing construction given here can then be applied to this partially reduced problem. Alternatively (Section 6.1), one can start from a full basis based on a reproducing kernel or semikernel, for which the penalty matrix and model matrix contain large subblocks that are identical. This block can be replaced by a truncated eigen decomposition computed efficiently by Krylov subspace methods (van der Vorst 2003).

2.4. The Bayesian Viewpoint

Penalization favors smooth function estimates relative to wiggly ones. Presumably we do this because we believe the truth is more likely to be smooth than wiggly, which we might as well formalize in a Bayesian manner by putting a prior on f that penalizes wiggliness (Kimeldorf & Wahba 1970, Silverman 1985). Continuing the cubic spline example, and choosing a later convenient constant of proportionality, the simplest prior might be

$$\pi(f) \propto \exp\left(-\frac{\lambda}{2\sigma^2}\int f''(x)^2 \mathrm{d}x\right).$$

Given a basis expansion for f, this implies a prior $\pi(\boldsymbol{\beta}) \propto \exp\{-\lambda\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}/(2\sigma^2)\}$, recognizable as $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \mathbf{S}^{-}\sigma^2/\lambda)$, where $\mathbf{S}^{-}$ is a pseudoinverse, as the precision matrix $\lambda\mathbf{S}/\sigma^2$ is not full rank.

Following the usual Bayesian approach we have that $\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \pi(\mathbf{y}, \boldsymbol{\beta})$, and

$$\log \pi(\mathbf{y}, \boldsymbol{\beta}) = -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/(2\sigma^2) - \lambda\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}/(2\sigma^2) + c,$$

where c is independent of $\boldsymbol{\beta}$ (see also Silverman 1985). The right-hand side is maximized by $\hat{\boldsymbol{\beta}}$ and, because it is quadratic in $\boldsymbol{\beta}$, it can be exactly replaced by its Taylor expansion about $\hat{\boldsymbol{\beta}}$, so that

$$\log \pi(\mathbf{y}, \boldsymbol{\beta}) = \log \pi(\mathbf{y}, \hat{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + c.$$

On exponentiating we see at once that $\boldsymbol{\beta} \mid \mathbf{y} \sim \mathrm{N}\{\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2\}$. The mathematical structure here is the same as that for modeling using latent Gaussian random fields (e.g., Rue & Held 2005, Rue et al. 2009) or mixed modeling with Gaussian random effects (explored by Ruppert et al. 2003), and there is also a strong link to Gaussian process regression (e.g., Kammann & Wand 2003).

The Bayesian covariance matrix $\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2$ differs from the frequentist covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ for $\hat{\boldsymbol{\beta}}$. As $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$ and $\mathrm{cov}(\mathbf{y}) = \mathbf{I}\sigma^2$, we have $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2$. The difference between the two matrices turns out to be the squared bias matrix expected according to the smoothing prior. This is most easily established by modifying $\mathbf{S}$ to formally have full rank, and thus an inverse, by resetting its zero eigenvalues to some constant ϵ that is too small to make any numerical difference to the fitted model. Then, as the bias is $(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}$, the expected squared bias matrix is

$$\begin{aligned}\mathbb{E}_{\pi}\{(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}}(\mathbf{F} - \mathbf{I})^{\mathsf{T}}\} &= (\mathbf{F} - \mathbf{I})\mathbf{S}^{-1}(\mathbf{F} - \mathbf{I})^{\mathsf{T}}\sigma^2/\lambda \\ &= (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{S}\mathbf{S}^{-1}\mathbf{S}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2\lambda \\ &= (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{S}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2\lambda = \mathbf{V}_{\boldsymbol{\beta}} - \mathbf{V}_{\hat{\boldsymbol{\beta}}}.\end{aligned}$$

Construction of confidence intervals for $f(x)$, based on $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$, is complicated by the smoothing bias. The obvious approach is to undersmooth to reduce the bias, but the price paid is then a more variable estimate. Bayesian credible intervals are wider by virtue of the squared expected bias component in $\mathbf{V}_{\boldsymbol{\beta}}$, without undersmoothing. They also turn out to have reasonably well calibrated frequentist performance, when considered on average across the function, as explained by Nychka (1988). At any fixed x the smoothing bias in $\hat{f}(x)$ may be systematically smaller or larger than its expected value under the prior, since the expectation is taken with respect to all functions plausible under the prior, not just those consistent with the data. However, the bias at a random x is a random variable, so considering an x_i picked at random, we have

$$\hat{f}(x_i) = f(x_i) + \{\mathbb{E}\hat{f}(x_i) - f(x_i)\} + \{\hat{f}(x_i) - \mathbb{E}\hat{f}(x_i)\} = f(x_i) + b_i + e_i,$$

where b_i and e_i are the bias and estimator sampling error at the randomly chosen x_i. Note that $\sum_i b_i = 0$ as a result of Equation 4. Using the Bayesian estimate for the expected squared bias, the variance of $b_i + e_i$ is $A_{ii}\sigma^2$. The estimator sampling error, e_i, is Gaussian and the Bayesian approach would also imply that b_i is Gaussian, but Nychka argues that even without the latter assumption, $(b_i + e_i)A_{ii}^{-1/2}$ will be approximately Gaussian, provided that the variance of $b_iA_{ii}^{-1/2}$ is substantially less than the variance of $e_i A_{ii}^{-1/2}$, as is usually the case. The upshot is that if z_α denotes the α critical point of a standard normal distribution, then intervals

$$\hat{f}(x_i) \pm z_\alpha A_{ii}^{1/2}\sigma,$$

proposed by Wahba (1983), should have close to $100(1-2\alpha)\%$ coverage, on average, over the x_i.

In practice the coverage properties of these intervals show rather little sensitivity to the exact choice of λ. A contributory factor here is some insensitivity of the interval widths to small changes in λ, around the λ value that approximately minimizes mean square error in $\hat{f}$ (MSE). The MSE is made up of a sampling variance component that decreases with λ and a squared bias component that increases with λ. At a unique minimum MSE, the gradients of the two components must be equal and opposite. Hence, to first order, any small change in λ away from the optimum leads to compensatory changes in the variance and squared bias. In other words, although a small change in λ implies a change in the squared bias to variance ratio, it makes very little difference to the interval width.

The exception to good coverage and insensitivity to λ tends to occur when the true $f(x)$ is close to a function in the penalty null space (Marra & Wood 2012). Then λ may be estimated as effectively infinite with nonnegligible frequency, so that the Bayesian squared bias estimate is zero, the insensitivity argument no longer holds, and Nychka's squared bias less than variance assumption can fail [particularly when $f(x)$ is part of a larger model where it is subject to a sum-to-zero constraint].

2.5. Smoothing Parameter Estimation

Continuing with the Bayesian viewpoint, the log marginal likelihood of λ is

$$\log \pi(\mathbf{y} \mid \lambda) = \log \pi(\mathbf{y} \mid \hat{\boldsymbol{\beta}}, \lambda) + \log \pi(\hat{\boldsymbol{\beta}} \mid \lambda) - \log \pi(\hat{\boldsymbol{\beta}} \mid \mathbf{y}, \lambda), \qquad 5.$$

where the previously implicit dependence on λ is now explicit. So for the smooth model,

$$\log \pi(\mathbf{y} \mid \lambda) = -\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2} - \frac{\lambda \hat{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{S} \hat{\boldsymbol{\beta}}}{2\sigma^2} + \frac{\log|\lambda \mathbf{S}|_+}{2} - \frac{\log|\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \mathbf{S}|}{2} - (n - M)\log\sigma + c, \qquad 6.$$

where $|\mathbf{C}|_+$ denotes the product of the nonzero eigenvalues[4] of $\mathbf{C}$, M is the null space dimension of $\mathbf{S}$, and c a constant. Wahba (1985) suggested $\hat{\lambda} = \operatorname{argmax}_\lambda \log \pi(\mathbf{y} \mid \lambda)$, an empirical Bayes approach. The marginal likelihood is also known as the restricted marginal likelihood (REML).

The Bayesian approach is not the only way to estimate λ. Another is prediction error minimization. The most obvious method is leave-one-out cross validation, which in general seeks λ to minimize

$$V_c(\lambda) = -\sum_{i=1}^{n} \log \pi(y_i \mid \boldsymbol{\beta}^{[-i]}, \lambda), \qquad 7.$$

where $\boldsymbol{\beta}^{[-i]}$ is $\hat{\boldsymbol{\beta}}$ on omission of x_i, y_i from the data fitted. In the Gaussian case, the log likelihood is quadratic in $\boldsymbol{\beta}$, so $\boldsymbol{\beta}^{[-i]}$ can be computed exactly by one step of Newton's method. The best-conditioned approach starts from a Cholesky decomposition of the Hessian for the fit to all the data,

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S} = \mathbf{R}^{\mathsf{T}}\mathbf{R}.$$

[4] The cheap way to compute this is to note that if $\mathbf{C} = \mathbf{K}^{\mathsf{T}}\mathbf{K}$, where $\mathbf{K}$ is of dimension rank$(\mathbf{C}) \times p$ then $|\mathbf{C}|_+ = |\mathbf{K}\mathbf{K}^{\mathsf{T}}|$. Pivoted Cholesky decomposition provides $\mathbf{K}$ efficiently.

Writing $\mathbf{X}_i$ for the ith row of $\mathbf{X}$, the Hessian on omission of the ith datum is

$$\mathbf{R}^{\mathsf{T}}\mathbf{R} - \mathbf{X}_i^{\mathsf{T}}\mathbf{X}_i = \mathbf{R}^{[-i]\mathsf{T}}\mathbf{R}^{[-i]},$$

and the modified Cholesky factor $\mathbf{R}^{[-i]}$, after this rank one update, is computable at $O(p^2)$ cost (Golub & van Loan 2013, section 6.5.4). The corresponding modified gradient is $\mathbf{g}^{[-i]} = \mathbf{g} + 2(y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})\mathbf{X}_i^{\mathsf{T}}$, so

$$\boldsymbol{\beta}^{[-i]} = \hat{\boldsymbol{\beta}} - \mathbf{R}^{[-i]-1}\mathbf{R}^{[-i]-\mathsf{T}}\mathbf{g}^{[-i]}.$$

Hence, V_c is computable at the same $O(np^2)$ cost as evaluation of $\hat{\boldsymbol{\beta}}$. A more familiar presentation of cross validation arrives at $V_c(\lambda) = \sum_i (y_i - \hat{\mu}_i)^2/(1 - A_{ii})^2$. This generalizes less easily beyond single smoothing parameters and exponential families but is the basis for generalized cross validation (GCV; Golub et al. 1979), in which the individual leverage terms, A_{ii}, are replaced by their average, to arrive at $V_g(\lambda) = n\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2/\{n - \mathrm{tr}(\mathbf{A})\}^2$ (the residual variance per residual degree of freedom). Recall that $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{F})$, the latter being computationally preferable. Advantageously, GCV is invariant to orthogonal transformation of the residual vector in the penalized least squares problem defining $\hat{\boldsymbol{\beta}}$, as is $\hat{\boldsymbol{\beta}}$ for a given λ, but not $V_c(\lambda)$; this consideration is less relevant beyond least squares.

3. GENERALIZED ADDITIVE MODELS

GAMs (Hastie & Tibshirani 1986) simply add more smooth terms to the simple one-dimensional smoothing model, while allowing the distribution of y_i to be non-Gaussian and some of the smooth terms to be functions of more than one variable. The most basic form is

$$g(\mu_i) = \eta_i = \mathbf{M}_i\gamma + \sum_j f_j(x_{ij}), \quad y_i \sim \pi(y_i \mid \mu_i, \boldsymbol{\theta}), \tag{8.}$$

where g is a known smooth monotonic link function of the expected value of y_i, as in any GLM; $\mathbf{M}$ is the model matrix for any strictly parametric terms with associated parameter vector $\boldsymbol{\gamma}$; the f_j are smooth functions of covariates x_j (possibly vector); η_i is known as the linear predictor; and π denotes some distribution with location parameter μ_i and possibly some other unknown parameters $\boldsymbol{\theta}$ (usually estimated to optimize the $\boldsymbol{\lambda}$ estimation criterion). Each f_j is represented using its own basis expansion and will have an associated smoothing penalty/prior with its own smoothing parameter, λ_j. In fact, some smooth terms may have multiple smoothing parameters (e.g., Section 6.2).

3.1. Identifiability Constraints

The main immediate nuisance introduced by multiple f_j is the lack of identifiability that it introduces. Clearly the individual f_j are only identifiable to within an additive constant, a problem that can be resolved by imposing the linear sum-to-zero constraint $\sum_i f_j(x_{ij}) = 0$ on each smooth term. Formally various other linear constraints could be used, but, by effectively orthogonalizing against the intercept, narrow confidence intervals for the constrained f_j are obtained, which aids interpretability.

For now let $\mathbf{X}$ be the model matrix for a single f_j. The sum-to-zero constraint can be imposed by subtracting the column mean from each column of $\mathbf{X}$ and dropping the column with the smallest variance,[5] while dropping the equivalent column and row from the penalty matrix $\mathbf{S}$. An alternative

[5] This ensures that any constant column is automatically dropped.

(which readily generalizes to any set of linear constraints) rewrites the constraint as $\mathbf{1}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = 0$ and then reexpresses it in terms of the QR decomposition of the vector $\bar{\mathbf{x}} = \mathbf{X}^\mathsf{T}\mathbf{1}$:

$$\mathcal{Q}\begin{pmatrix} \|\bar{\mathbf{x}}\| \\ \mathbf{0} \end{pmatrix} = \bar{\mathbf{x}} \Rightarrow (\|\bar{\mathbf{x}}\|, \mathbf{0})\begin{pmatrix} \mathbf{q}^\mathsf{T} \\ \mathbf{Q}^\mathsf{T} \end{pmatrix}\boldsymbol{\beta} = 0,$$

where $\mathbf{q}$ is the first column of $\mathcal{Q}$, and $\mathbf{Q}$ denotes the remaining columns. The norm is Euclidean. The reparameterization $\boldsymbol{\beta} = \mathbf{Q}\boldsymbol{\beta}_z$ automatically meets the constraint. This amounts to setting the model matrix for f_j to $\mathbf{XQ}$ and the penalty matrix to $\mathbf{Q}^\mathsf{T}\mathbf{SQ}$. In practice $\mathbf{Q}$ is not formed explicitly, since $\mathcal{Q}$ is defined by a single Householder reflection: It is more efficient to compute with it as such.

When using an estimated model for prediction at new covariate values, it is necessary to produce a model matrix for prediction, or prediction matrix. The original $\mathbf{Q}$ must be used to apply the constraint to this. Or, for column centering constraints, the original column means must be subtracted from the columns and the same column dropped as was dropped in model setup. Recomputing the constraints from the prediction matrix would result in predicting from a different model to that fitted. As the constraints simply impose identifiability, an alternative is to add a quadratic penalty, $\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}$, for each f_j, to the model-fitting problem. This gives the same results as the other methods.

3.2. Beyond Gaussian Likelihood

Having obtained basis expansions and penalties for each smooth term, and imposed identifiability constraints, we can write the $\boldsymbol{\gamma}$ parameters and all the spline basis coefficients in one vector $\boldsymbol{\beta}$, and combine all the smoothing penalty matrices into one block diagonal matrix $\mathbf{S}_\lambda$, which is linear in the elements of the smoothing parameter vector $\boldsymbol{\lambda}$. In what follows, it helps to write $\mathbf{S}_\lambda = \sum_j \lambda_j \mathbf{S}_j$, where the $\mathbf{S}_j$ are zero apart from one diagonal block. Then, we have

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta},$$

where $l(\boldsymbol{\beta}) = \log \pi(\mathbf{y} \mid \boldsymbol{\beta})$. Obviously $\hat{\boldsymbol{\beta}}$ is again the posterior mode under the Bayesian view.

Consider the full posterior distribution of $\boldsymbol{\beta}$. For compactness let $\mathcal{L}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta}$, write $a^i_{\hat{\beta}} = \partial a/\partial\beta_i|_{\hat{\beta}_i}$, and sum over indices occurring on only one side of an equation. By Taylor's theorem,

$$\log \pi(\boldsymbol{\beta}, \mathbf{y}) = \log \pi(\hat{\boldsymbol{\beta}}, \mathbf{y}) + \frac{1}{2}\mathcal{L}^{i\,j}_{\hat{\beta}\hat{\beta}}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) + \frac{1}{6}l^{i\,j\,k}_{\tilde{\beta}\tilde{\beta}\tilde{\beta}}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)(\beta_k - \hat{\beta}_k),$$

where $\tilde{\boldsymbol{\beta}}$ lies on the line from $\boldsymbol{\beta}$ to $\hat{\boldsymbol{\beta}}$. Given $l^{i\,j}_{\hat{\beta}\hat{\beta}}$ and $l^{i\,j\,k}_{\tilde{\beta}\tilde{\beta}\tilde{\beta}}$ both $O(n/p)$,[6] where $p = \dim(\boldsymbol{\beta})$, then for $\beta_i - \hat{\beta}_i = O(\sqrt{p/n})$, the second term in the Taylor expansion is $O(p^2)$ and the third term is $O(p^2\sqrt{p^3/n})$, implying that the second term dominates if $p = o(n^{1/3})$. In that case, since $\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\beta}, \mathbf{y})$, on exponentiating we recognize the asymptotic result

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta),\ \text{where } \mathbf{V}_\beta = (\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1} \qquad 9.$$

and $\hat{\mathcal{I}}_{ij} = -l^{i\,j}_{\hat{\beta}\hat{\beta}}$. Result 9 can be used directly to produce approximate (and component-wise) versions of the confidence intervals discussed in Section 2.4, or for very efficient approximate sampling from the posterior, or as the basis for a proposal distribution when Metropolis–Hastings sampling from $\pi(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\lambda})$ itself.

[6] This is common in regression problems and easy to see when using B-splines.

The result also suggests an approximate marginal likelihood estimate for $\boldsymbol{\lambda}$: Maximize Expression 5 with respect to $\boldsymbol{\lambda}$ with the Gaussian approximation substituted for $\log \pi(\hat{\boldsymbol{\beta}} \mid \mathbf{y}, \boldsymbol{\lambda})$. Neglecting uninteresting constants, the approximate (exact in the Gaussian identity link case) marginal likelihood is

$$V_r(\boldsymbol{\lambda}) = l(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{S}_\lambda \hat{\boldsymbol{\beta}}/2 + \log |\mathbf{S}_\lambda|_+/2 - \log |\hat{\mathcal{I}} + \mathbf{S}_\lambda|/2, \qquad 10.$$

which turns out to be a first-order Laplace approximation to the (log) marginal likelihood.

Cross validation is equally easy to generalize to this setting for any case in which $\hat{\mathcal{I}} = \mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is diagonal. In that case, Equation 7 can be computed to high accuracy starting from $\hat{\mathcal{I}} + \mathbf{S}_\lambda = \mathbf{R}^{\mathsf{T}}\mathbf{R}$ and obtaining $\boldsymbol{\beta}^{[-i]}$ by single Newton steps from $\hat{\boldsymbol{\beta}}$ based on $\mathbf{R}^{[-i]}$ updated as in Section 2.5. A rougher approximation produces a GCV criterion,

$$V_g(\lambda) = nD(\hat{\boldsymbol{\beta}})/\{n - \mathrm{tr}(\mathbf{F})\}^2, \qquad 11.$$

where the deviance, D, is defined as $D = 2\{l_{\mathrm{sat}} - l(\hat{\boldsymbol{\beta}})\}$, with l_{sat} denoting the saturated log likelihood, i.e., the largest value the log likelihood can attain for the given $\mathbf{y}$.

The degrees of freedom matrix is $\mathbf{F} = (\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\hat{\mathcal{I}}$ or, better, $\mathbf{F} = (\mathcal{I} + \mathbf{S}_\lambda)^{-1}\mathcal{I}$, where $\mathcal{I} = \mathbb{E}(\hat{\mathcal{I}})$, since $\hat{\mathcal{I}}$ need not be positive definite at $\hat{\boldsymbol{\beta}}$, unlike $\mathcal{I}$. The effective degrees of freedom for an individual smooth term is obtained by summing the F_{ii} corresponding to its coefficients, again following the shrinkage or variance reduction factor approach of Section 2.3. The frequentist sampling covariance matrix becomes $\mathbf{V}_{\hat{\beta}} = \mathbf{F}\mathbf{V}_\beta$.

3.2.1. Scale parameters. When the log likelihood is scaled by a multiplicative scale parameter, ϕ^{-1}, that too must be estimated. Optimization of V_c and V_r with respect to ϕ is one approach, but it cannot be employed with quasi-likelihood (McCullagh & Nelder 1989, section 9). An alternative is to use the usual GLM estimators, i.e., the deviance or Pearson statistic divided by the residual degrees of freedom, $n - \mathrm{tr}(\mathbf{F})$, but these are poorly behaved for count data with a low mean. Fletcher (2012) shows how to fix this. Let $V(\mu)$ be the GLM variance function. Then if $P = \sum_i (y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i)$ and $\hat{\phi}_P = P/\{n - \mathrm{tr}(\mathbf{F})\}$, Fletcher's estimator is $\hat{\phi} = \hat{\phi}_P/(1 + \bar{s})$, where $\bar{s} = n^{-1}\sum_i V'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/V(\hat{\mu}_i)$.

3.3. Model Selection

In addition to point and interval estimation given a model structure, practical use of GAMs usually involves inference about the model structure itself, and as usual this is more difficult. This section covers methods useful when only a few model terms are in question. Higher-dimensional feature selection often requires different approaches (e.g., Chouldechova & Hastie 2015, He & Wand 2024) with boosting appearing particularly promising (e.g., Schmid & Hothorn 2008).

3.3.1. Null space penalties. A good deal of what might usually be considered to be model selection is performed by smoothing parameter estimation, which in effect selects between a large family of models of differing complexity, albeit stopping short of removing a model term entirely. Hence, an obvious option is to add a penalty and smoothing parameter for each smooth term, which penalizes its otherwise unpenalized component toward zero. If all the smoothing parameters for the term then tend to infinity, the term itself tends to the zero function and is selected out of the model. Such null space penalties are easy to construct for a smooth term, f_j, with penalty matrix $\mathbf{S}_j$. Form a symmetric eigen decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}} = \mathbf{S}_j$ and let $\mathbf{U}_0$ denote the eigenvectors corresponding to zero eigenvalues. Then $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{U}_0\mathbf{U}_0^{\mathsf{T}}\boldsymbol{\beta}$ is a penalty for the otherwise unpenalized component of f_j.

3.3.2. Akaike information criterion. It is also helpful to have access to standard tools, such as the Akaike information criterion (AIC) (Akaike 1973) and hypothesis testing for model comparison. Starting with the AIC, let the model likelihood be written in terms of the linear predictor, which we suppose has true value $\boldsymbol{\eta}_t$. Different models result in different estimates $\hat{\boldsymbol{\eta}}$, and the AIC aims to select the one with the highest value of

$$\mathbb{E}_t \log \pi(\mathbf{y} \mid \hat{\boldsymbol{\eta}}) = \int \pi(\mathbf{y} \mid \boldsymbol{\eta}_t) \log \pi(\mathbf{y} \mid \hat{\boldsymbol{\eta}}) \mathrm{d}\mathbf{y},$$

where within the integral, $\hat{\boldsymbol{\eta}}$ is treated as fixed, not as a function of $\mathbf{y}$. So we seek the model that would have the highest expected log likelihood for new data not used in fitting. A Taylor expansion gives

$$\log \pi(\mathbf{y} \mid \hat{\boldsymbol{\eta}}) = \log \pi(\mathbf{y} \mid \boldsymbol{\eta}_t) + \frac{\partial l}{\partial \boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t) + \frac{1}{2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)^{\mathsf{T}} \frac{\partial^2 l}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathsf{T}}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t),$$

where the first derivative of $l = \log \pi(\mathbf{y} \mid \boldsymbol{\eta})$ is evaluated at $\boldsymbol{\eta}_t$ and the Hessian is evaluated somewhere on the line from $\boldsymbol{\eta}_t$ to $\hat{\boldsymbol{\eta}}$. On taking expectations with respect to $\pi(\mathbf{y} \mid \boldsymbol{\eta}_t)$ (keeping $\hat{\boldsymbol{\eta}}$ fixed), the first derivative vanishes and we have

$$\mathbb{E}_t \log \pi(\mathbf{y} \mid \hat{\boldsymbol{\eta}}) = \mathbb{E}_t \log \pi(\mathbf{y} \mid \boldsymbol{\eta}_t) - (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)^{\mathsf{T}} \boldsymbol{\mathcal{I}}_\eta (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)/2 \simeq \mathbb{E}_t \log \pi(\mathbf{y}|\boldsymbol{\eta}_t) - \operatorname{tr}(\mathbf{F})/2,$$

where the information matrix is evaluated between $\boldsymbol{\eta}_t$ and $\hat{\boldsymbol{\eta}}$, and the approximation replaces the quadratic term by its expectation when $\boldsymbol{\mathcal{I}}_\eta$ is evaluated at $\hat{\boldsymbol{\eta}}$. Specifically, using the Bayesian squared bias matrix estimate yields

$$\mathbb{E}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)^{\mathsf{T}} \boldsymbol{\mathcal{I}}_{\hat{\eta}} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)\} = \operatorname{tr}[\boldsymbol{\mathcal{I}}_{\hat{\eta}} \mathbb{E}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_t)^{\mathsf{T}}\}] = \operatorname{tr}\{\boldsymbol{\mathcal{I}}_{\hat{\eta}} \mathbf{X}(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1} \mathbf{X}^{\mathsf{T}}\} = \operatorname{tr}(\mathbf{F}),$$

given that $\boldsymbol{\mathcal{I}} = \mathbf{X}^{\mathsf{T}} \boldsymbol{\mathcal{I}}_{\hat{\eta}} \mathbf{X}$.

So on new data we expect the model log likelihood for $\hat{\boldsymbol{\eta}}$ to be lower than that for the true $\boldsymbol{\eta}_t$. However, on the original data the advantage is reversed. In particular if we now expand around $\hat{\boldsymbol{\eta}}$, we obtain $l(\hat{\boldsymbol{\eta}}) - l(\boldsymbol{\eta}_t) \simeq (\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}})^{\mathsf{T}} \boldsymbol{\mathcal{I}}_\eta (\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}})/2$. Hence ,

$$\mathbb{E}_t l(\hat{\boldsymbol{\eta}}) = \mathbb{E}_t\{l(\hat{\boldsymbol{\eta}}) - l(\boldsymbol{\eta}_t)\} + \mathbb{E}_t l(\boldsymbol{\eta}_t) \simeq \operatorname{tr}(\mathbf{F})/2 + \mathbb{E}_t \log \pi(\mathbf{y}|\boldsymbol{\eta}_t).$$

Estimating $\mathbb{E}_t l(\hat{\boldsymbol{\eta}})$ by $l(\hat{\boldsymbol{\eta}})$ and substituting gives $\mathbb{E}_t \log \pi(\mathbf{y}|\hat{\boldsymbol{\eta}}) \simeq l(\hat{\boldsymbol{\eta}}) - \operatorname{tr}(\mathbf{F})$. Multiplying by -2 (by convention), we get $\mathrm{AIC} = -2l(\hat{\boldsymbol{\eta}}) + 2\operatorname{tr}(\mathbf{F})$ as the criterion minimized by the best fitting model.

Greven & Kneib (2010) show that this AIC tends to overselect models containing simple Gaussian random effects, a problem attributable to neglected smoothing parameter uncertainty in tr($\mathbf{F}$). Wood et al. (2016) and Säfken et al. (2014) suggest fixes.

3.3.3. Hypothesis testing. Another approach to model selection is to test the null hypothesis that $f_j(x_j) = 0$ over the range of x_j.[7] It is tempting to form a Wald statistic directly in terms of the coefficients of $\hat{f}_j$, but this produces poor results. The resulting statistic most up-weights exactly those components of the smooth term that the penalty has most suppressed. Since $\boldsymbol{\lambda}$ selection aims to most suppress those components with the least support from the data, this is a recipe for low power. The literature contains a number of better, if relatively complicated, proposals; in particular, readers are directed to Crainiceanu et al. (2005), Greven et al. (2008), and Wood (2013a,b).

An alternative is to use the test statistic $T = \|\mathbf{W}_j \mathbf{X}_j \hat{\boldsymbol{\beta}}_j\|^2$, where $\mathbf{W}_j$ is a diagonal matrix of the reciprocal standard errors of $\hat{f}_j$, while $\mathbf{X}_j$ and $\boldsymbol{\beta}_j$ are the model matrix and coefficient vector for

[7] When interpreting the null hypothesis, recall that the f_j are subject to sum-to-zero identifiability constraints.

the term. Denoting the covariance matrix for $\hat{\boldsymbol{\beta}}_j$ by $\mathbf{V}_j$, the variances of the $f_j(x_{ij})$ are the leading diagonal elements of $\mathbf{X}_j\mathbf{V}_j\mathbf{X}_j$. These are most efficiently computed as $\{(\mathbf{X}_j\mathbf{V}_j) \odot \mathbf{X}_j\}\mathbf{1}$, where $\odot$ is the Hadarmard (element-wise) product. The reciprocal square roots of these variances give the leading diagonal elements of $\mathbf{W}_j$.

Under the null hypothesis (using the large sample Gaussian approximation), we have $\hat{\boldsymbol{\beta}}_j \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_j)$. Writing $T = \|\mathbf{t}\|^2$, where $\mathbf{t} = \mathbf{W}_j\mathbf{X}_j\hat{\boldsymbol{\beta}}_j$, we can obtain the covariance matrix of $\mathbf{t}$ and its eigen decomposition,

$$\mathbf{V}_t = \mathbf{W}_j\mathbf{X}_j\mathbf{V}_j\mathbf{X}_j^{\mathsf{T}}\mathbf{W}_j = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}.$$

Defining $\mathbf{t}' = \mathbf{U}^{\mathsf{T}}\mathbf{t}$, then $T = \|\mathbf{t}'\|^2$ and hence $\mathbf{V}_{t'} = \boldsymbol{\Lambda}$. That is, under the null, the t'_i are zero mean independent Gaussian variables with variances Λ_{ii}. Hence, $t_i'^2 = \Lambda_{ii}z_i^2$, where the z_i are independent N(0, 1). In other words T has the distribution of a weighted sum of χ_1^2 random variables. Hence, the p-value is $\Pr(\sum_i \Lambda_{ii}z_i^2 > T)$. When $\mathbf{V}_j$ involves an unknown scale parameter, the p-value is better computed as

$$\Pr\left(\sum_i \Lambda_{ii}z_i^2 > T\chi_\kappa^2/\kappa\right) = \Pr\left(\sum_i \Lambda_{ii}z_i^2 - T\chi_\kappa^2/\kappa > 0\right),$$

where χ_κ^2 denotes a chi-squared random variable with κ degrees of freedom—the effective residual degrees of freedom of the model fit. These probabilities involving weighted sums of χ^2 random variables can readily be computed (Davies 1973, 1980) or accurately approximated (Kuonen 1999).

The eigenvalues are not obtained directly from $\mathbf{V}_t$. For computational efficiency a QR decomposition $\mathbf{QR} = \mathbf{X}_j$ is formed first, so the nonzero elements of $\boldsymbol{\Lambda}$ are the eigenvalues of $\mathbf{R}\mathbf{V}_j\mathbf{R}^{\mathsf{T}}$.

What should be used for $\mathbf{V}_j$? There are two distinct penalization scenarios to consider. The first is where penalization shrinks toward the null hypothesis—for example, if f_j is a smooth function subject to the penalty $\int f_j'(x)^2 \mathrm{d}x$, or even a Gaussian random effect term. Under the null hypothesis, shrinkage is toward the truth and simply reduces the sampling variance of the estimator around the null function, as captured by $\mathbf{V}_{\hat{\beta}}$, which therefore supplies the appropriate $\mathbf{V}_j$. The second scenario is when shrinkage is toward some function other than that specified by the null. For example, the usual cubic spline penalty shrinks toward the simple linear regression fit to the data. Unless this regression has exactly zero slope, this shrinkage is away from the null function. Hence, shrinkage now acts to reduce the sampling variability as before, but it becomes necessary to also account for the shrinkage away from the null function using the argument of Section 2.4, and hence employing the appropriate block of the Bayesian covariance matrix, $\mathbf{V}_\beta$ for $\mathbf{V}_j$.[8]

4. COMPUTATIONAL STRATEGIES

This section discusses computation using the reduced rank empirical Bayes approach. Notable alternative computational strategies are backfitting (Hastie & Tibshirani 1990), Markov chain Monte Carlo (MCMC) (e.g., Fahrmeir & Lang 2001), integrated nested Laplace approximation (Rue et al. 2009), or boosting (e.g., Schmid & Hothorn 2008). For the empirical Bayes methods, computation of $\hat{\boldsymbol{\beta}}$ given the log smoothing parameters, $\boldsymbol{\rho} = \log \boldsymbol{\lambda}$, is relatively straightforward, using Newton's method on the penalized log likelihood. But the optimization required to estimate $\boldsymbol{\rho}$ is more involved.

[8] Using the Bayesian $\mathbf{V}_\beta$ under the first scenario incorrectly adds a correction for a nonexistent shrinkage away from the hypothesized null function. This erroneous inflation of the estimator variance inflates p-values.

4.1. Preliminaries

Before discussing optimization strategies, some preliminary matters must be discussed.

4.1.1. Maintaining stability. An $\hat{f}_j$ for which the corresponding penalty, $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$, is zero is a legitimate estimate (a cubic spline being estimated to be simply a straight line, for example) but corresponds to $\hat{\lambda}_j \to \infty$, which can cause numerical difficulties. In particular, for a rank-deficient $\mathbf{S}_j$, in finite precision arithmetic, $\lambda_j \to \infty$ can cause the penalty to so dominate the likelihood that the data have no influence on $\hat{f}_j$ even for components in the penalty null space. The log determinant terms in V_r can also lose all precision if computed naively.

There are two ways to avoid difficulty. One is to limit the smoothing parameters during optimization, by not increasing any ρ_j for which the degrees of freedom suppressed by the penalty are too close to the rank of the penalty matrix. Too close is when $\text{rank}(\mathbf{S}_j) - \text{tr}\{(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\lambda_j\} < 0.02$, say. When the same smooth term is subject to multiple penalties, then $\text{rank}(\mathbf{S}_j)$ has to be replaced by the penalty's effective rank $\text{tr}(\mathbf{S}_\lambda^-\mathbf{S}_j\lambda_j)$. This is the maximum degrees of freedom that could be associated with the jth penalty given the other penalties and their smoothing parameter values.

The second approach leaves the λ_j unconstrained but uses reparameterization to stabilize the computation with very large λ_j values. The most challenging term is $|\mathbf{S}_\lambda|_+$, and a reparameterization stabilizing this also stabilizes the rest of the computation. $\mathbf{S}_\lambda$ is block diagonal—for example,

$$\mathbf{S}_\lambda = \begin{pmatrix} \lambda_1\boldsymbol{\mathcal{S}}_1 & \mathbf{0} & \mathbf{0} & \cdot \\ \mathbf{0} & \lambda_2\boldsymbol{\mathcal{S}}_2 & \mathbf{0} & \cdot \\ \mathbf{0} & \mathbf{0} & \sum_j \lambda_j\boldsymbol{\mathcal{S}}_j & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Blocks with one smoothing parameter are easier to handle than those with several, which occur for tensor product and adaptive smooths. Any number of both block types can occur in any order, and $\mathbf{0}$ blocks on the diagonal are also possible. For the given example, it is easy to show that

$$\log|\mathbf{S}_\lambda|_+ = \text{rank}(\boldsymbol{\mathcal{S}}_1)\log\lambda_1 + \log|\boldsymbol{\mathcal{S}}_1|_+ + \text{rank}(\boldsymbol{\mathcal{S}}_2)\log\lambda_2 + \log|\boldsymbol{\mathcal{S}}_2|_+ + \log\left|\sum_j \lambda_j\boldsymbol{\mathcal{S}}_j\right|_+ + \cdots.$$

For the single smoothing parameter blocks, a symmetric eigen decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}} = \boldsymbol{\mathcal{S}}_j$ can be formed and then the reparameterization $\boldsymbol{\beta}_j \leftarrow \mathbf{U}^{\mathsf{T}}\boldsymbol{\beta}_j$ applied so that the penalty matrix becomes $\boldsymbol{\Lambda}$, with the zero eigenvalues set to exact zeroes. In that case $\log|\boldsymbol{\mathcal{S}}_j|_+$ is just the sum of the logarithms of the positive eigenvalues; also, $\partial\log|\mathbf{S}_\lambda|_+/\partial\rho_j = \text{rank}(\boldsymbol{\mathcal{S}}_j)$.

For multiple smoothing parameter blocks, difficulties arise when a λ_k becomes so large that the formally zero eigenvalues of rank-deficient $\lambda_k\boldsymbol{\mathcal{S}}_k$ evaluate numerically to values of significant size relative to the smallest positive eigenvalues of $\sum_j\lambda_j\boldsymbol{\mathcal{S}}_j$. Then $\log|\sum_j\lambda_j\boldsymbol{\mathcal{S}}_j|_+$ can lose all precision.

A similarity transform method fixes this. In particular we seek to reparameterize so that the dominant block, $\lambda_k\boldsymbol{\mathcal{S}}_k$, is restricted to a $\text{rank}(\boldsymbol{\mathcal{S}}_k) \times \text{rank}(\boldsymbol{\mathcal{S}}_k)$ block of the transformed $\boldsymbol{\mathcal{S}} = \sum_j\lambda_j\boldsymbol{\mathcal{S}}_j$. Let $r = \text{rank}(\boldsymbol{\mathcal{S}}_k)$, p be the dimension of $\boldsymbol{\mathcal{S}}$, and $m = p - r$. Forming the eigen decomposition $\mathbf{U}\mathbf{D}\mathbf{U}^{\mathsf{T}} = \boldsymbol{\mathcal{S}}_k$, let $\mathbf{D}_r$ denote the diagonal matrix containing the positive eigenvalues, $\mathbf{U}_r$ the corresponding columns of $\mathbf{U}$, and $\mathbf{U}_m$ the remaining columns. Now consider the similarity transform $\boldsymbol{\mathcal{S}}' = \mathbf{U}^{\mathsf{T}}\boldsymbol{\mathcal{S}}\mathbf{U}$, equivalent to the reparameterization $\boldsymbol{\beta}'_j = \mathbf{U}^{\mathsf{T}}\boldsymbol{\beta}_j$. $\boldsymbol{\mathcal{S}}'$ and $\boldsymbol{\mathcal{S}}$ obviously have the same generalized determinant. Defining $\boldsymbol{\mathcal{S}}_{\bar{k}} = \sum_{j\neq k} \lambda_j \boldsymbol{\mathcal{S}}_j$ gives

$$\boldsymbol{\mathcal{S}}' = \begin{pmatrix} \mathbf{D}_r + \mathbf{U}_r^{\mathsf{T}} \boldsymbol{\mathcal{S}}_{\bar{k}} \mathbf{U}_r & \mathbf{U}_r^{\mathsf{T}} \boldsymbol{\mathcal{S}}_{\bar{k}} \mathbf{U}_m \\ \mathbf{U}_m^{\mathsf{T}} \boldsymbol{\mathcal{S}}_{\bar{k}} \mathbf{U}_r & \mathbf{U}_m^{\mathsf{T}} \boldsymbol{\mathcal{S}}_{\bar{k}} \mathbf{U}_m \end{pmatrix}.$$

Computing using the above right-hand side restricts the impact of the dominant block strictly to the upper left $r \times r$ block of $\boldsymbol{\mathcal{S}}'$, but the lower right $m \times m$ block, $\mathbf{U}_m^{\mathsf{T}} \boldsymbol{\mathcal{S}}_{\bar{k}} \mathbf{U}_m = \sum_{j\neq k} \lambda_j \mathbf{U}_m^{\mathsf{T}} \boldsymbol{\mathcal{S}}_j \mathbf{U}_m$, now has the same form as the original $p \times p$ block. To ensure its stable computation, we again identify the dominant term in the summation and apply the same similarity transform approach again. This process can be iterated until no trailing block is left, at which point the final M rows and columns of the final $\boldsymbol{\mathcal{S}}'$ will be zero, M being the dimension of the null space of $\boldsymbol{\mathcal{S}}$. The regular determinant of the leading $(p - M) \times (p - M)$ block of the final $\boldsymbol{\mathcal{S}}'$ gives $|\boldsymbol{\mathcal{S}}|_+$. It suffices to consider computing the determinant via QR decomposition to see that the column separation achieved by the similarity transform will lead to stable computation of the determinant (less obviously, it turns out that Cholesky-based computation is equally stable).

4.1.2. Finding $\hat{\boldsymbol{\beta}}$ given λ. Given smoothing parameters, $\hat{\boldsymbol{\beta}}$ can be found using Newton's method, which amounts to iteratively updating via $\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + (\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}_\lambda)^{-1}(\partial l/\partial \boldsymbol{\beta} - \mathbf{S}_\lambda \hat{\boldsymbol{\beta}})$, until convergence. As usual, to guarantee convergence, the penalized Hessian must be perturbed to be positive definite if it is not, and if necessary the update step should be repeatedly halved until the penalized likelihood is improved. In general the update step can be computed using a Cholesky decomposition of the penalized Hessian, but a more stable approach is also often possible. Any required perturbation to positive-definiteness can often be accomplished by replacing the Hessian with its expected value: a Fisher scoring step. However, as the Hessian itself is subsequently required for implicit differentiation it does not make sense to base the whole optimization on Fisher scoring.

In the context of location parameter regression models for independent data, the Hessian of the negative log likelihood with respect to $\boldsymbol{\beta}$ has the form $\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}$, where $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$. For example, in exponential family likelihoods such that $\text{var}(y_i) = V(\mu_i)\phi$, where $\mu_i = \mathbb{E}(y_i)$ and ϕ is a scale parameter, then

$$w_i = \frac{\alpha_i}{\phi g'(\mu_i)^2 V(\mu_i)}, \quad \text{where} \quad \alpha_i = 1 + (y_i - \mu_i)\left\{\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)}\right\}.$$

In such cases Newton's method is equivalent to repeatedly solving a penalized least squares problem with iteratively updated weights, w_i, and pseudodata, z_i:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_i w_i(z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_\lambda\boldsymbol{\beta}.$$

In the smooth GLM context, $z_i = g'(\mu_i)(y_i - \mu_i)/\alpha_i + \hat{\eta}_i$, where $\eta_i = \mathbf{X}_i\boldsymbol{\beta} = g(\mu_i)$. The iteration does not require starting $\hat{\boldsymbol{\beta}}$ values: y_i is used as the initial $\hat{\mu}_i$, perturbed a little if necessary to ensure finite $\hat{\eta}_i$ (e.g., if $y_i = 0$ and $g = \log$).

If the w_i are nonnegative,[9] then the least squares objective function is

$$\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\|^2 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_\lambda\boldsymbol{\beta} = \|\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2, \text{ where } \tilde{\mathbf{z}} = \begin{pmatrix} \sqrt{\mathbf{W}}\mathbf{z} \\ \mathbf{0} \end{pmatrix}, \ \tilde{\mathbf{X}} = \begin{pmatrix} \sqrt{\mathbf{W}}\mathbf{X} \\ \mathbf{E}_\lambda \end{pmatrix},$$

and $\mathbf{E}_\lambda$ is a square root of $\mathbf{S}_\lambda$, obtained by pivoted Cholesky or symmetric eigen decompositions, or as a by-product of the stabilizing transforms of the previous section. Forming a QR decomposition $\mathbf{QR} = \tilde{\mathbf{X}}$, we then have $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^{\mathsf{T}}\tilde{\mathbf{z}}$. The condition number of this system is the square root of the condition number of the Hessian, a considerable stability enhancement. If any $w_i < 0$ then some extra work is needed to employ the QR approach (Wood 2011).

[9] This is always true when using a canonical link.

4.1.3. Implicit differentiation. Direct optimization of V_r (Equation 10), V_c (Equation 7), or V_g (Equation 11) requires their derivatives with respect to $\boldsymbol{\rho}$, in turn requiring $\mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\boldsymbol{\rho}$, obtained by implicit differentiation. By definition of $\hat{\boldsymbol{\beta}}$,

$$\left.\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{\beta}}\right|_{\hat{\beta}} - \mathbf{S}_\lambda \hat{\boldsymbol{\beta}} = \mathbf{0}.$$

Differentiating with respect to ρ_j gives

$$\left.\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathsf{T}}}\right|_{\hat{\beta}} \frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\rho_j} - \lambda_j \mathbf{S}_j \hat{\boldsymbol{\beta}} - \mathbf{S}_\lambda \frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\rho_j} = \mathbf{0} \quad\Rightarrow\quad \frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\rho_j} = -\lambda_j(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\hat{\boldsymbol{\beta}}.$$

Differentiating again yields second derivatives of $\hat{\boldsymbol{\beta}}$ with respect to the ρ_j.

4.2. Nested Optimization

To estimate the smoothing parameters as the exact optimizers of V_r, V_c, or V_g (generically, V) requires nested optimization. An outer optimization method seeks log smoothing parameters $\hat{\boldsymbol{\rho}}$ to optimize V. For each trial $\boldsymbol{\rho}$ value proposed by the outer optimizer, an inner optimization finds the corresponding $\hat{\boldsymbol{\beta}}$, with implicit differentiation used to compute the derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\rho}$. The approach is less computationally costly than it might appear. As the outer optimization progresses, the previous $\hat{\boldsymbol{\beta}}$ becomes an ever-better starting value for the current $\hat{\boldsymbol{\beta}}$, so that the inner optimization requires very few steps to converge.

The inner optimization is performed by Newton's method (Section 4.1.2). The outer optimization can use a quasi-Newton or full Newton method, and consequently V must be differentiated with respect to $\boldsymbol{\rho}$ to first or second order, respectively. This requires the standard results $\partial \log|\mathbf{A}|/\partial x = \mathrm{tr}(\mathbf{A}^{-1}\partial\mathbf{A}/\partial x)$ and $\partial\mathbf{A}^{-1}/\partial x = -\mathbf{A}^{-1}\partial\mathbf{A}/\partial x\mathbf{A}^{-1}$ and the derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\rho}$. The main challenge is to structure computations so that evaluation of all first derivatives has only the $O(np^2)$ cost of $\hat{\boldsymbol{\beta}}$, while computation of all second derivatives has only $O(mnp^2)$ cost where $m = \dim(\boldsymbol{\rho})$. In principle this is always achievable by reverse mode automatic differentiation, but the trick is to do the same using a tiny fraction of the memory that this would require.

As discussed in Section 4.1.1, some care is required to maintain numerical stability of the computations. To this end the reparameterizations of Section 4.1.1 can be applied for each trial set of smoothing parameters. Even so, V will be indefinite when some ρ_j are large enough that the corresponding $\hat{f}_j$ is fully penalized and hence insensitive to further increases of ρ_j. Such indefiniteness slows optimizer convergence, while the consequent tendency to propose enormous ρ_j updates runs the risk of overflow. When using Newton's method for the outer optimization, we have access to the exact Hessian and gradient, so such indefiniteness is easily diagnosed, and the corresponding large ρ_j can be dropped from updates (while indefiniteness persists). When using quasi-Newton methods, the Hessian is unavailable and its approximation is positive definite by construction, rendering it useless for detecting loss of positive-definiteness. In that case we can instead use proximity of

$$\frac{\mathrm{d}\hat{\boldsymbol{\beta}}^{\mathsf{T}}}{\mathrm{d}\rho_j}\hat{\mathcal{I}}\frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\rho_j}$$

to zero to detect when the fit has lost sensitivity to ρ_j, again dropping such ρ_j from the update process if the update would be positive.

The alternative to reparameterization is to use the ρ_j limiting approach discussed in Section 4.1.1, in which ρ_j are no longer increased once the effective degrees of freedom suppressed by the corresponding penalty term is within some small constant of the maximum possible. This typically maintains sufficient stability and definiteness.

4.3. Performance Iteration/Penalized Quasi-Likelihood

While offering something of a gold standard for numerical stability and reliability, the nested optimization is complex to implement, and it is less amenable to adaptation for large models of large datasets than simpler approximate schemes. The oldest of these (Gu 1992, Breslow & Clayton 1993) exploits the working penalized linear model underlying the working penalized least squares optimization problem, discussed in Section 4.1.2. The idea is that rather than have an outer optimization, $\hat{\boldsymbol{\rho}}$ is estimated using a V appropriate for the working linear model at each step of the iteration for finding $\hat{\boldsymbol{\beta}}$.

Specifically, the working linear model implied by iterative penalized least squares is

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where} \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad \text{and} \quad \text{cov}(\boldsymbol{\epsilon}) = \mathbf{W}^{-1}\phi,$$

where we use the Fisher weights for $\mathbf{W}$ to keep them positive. We can apply cross validation to this working linear model, optimizing the appropriate V_c or V_g. But it is not obvious that marginal likelihood is appropriate, since the distribution of $\boldsymbol{\epsilon}$ is generally not Gaussian (or even explicitly known). However, consider the QR decomposition $\mathbf{QR} = \sqrt{\mathbf{W}}\mathbf{X}$. Multiplying both sides of the original model by $\mathbf{Q}^\mathsf{T}\sqrt{\mathbf{W}}$ yields $\mathbf{Q}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{z} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}'$, where $E(\boldsymbol{\epsilon}') = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \mathbf{I}\phi$. Crucially, as $n \to \infty$ and provided p grows only slowly, $\mathbf{Q}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{z}$ tends to multivariate normality by the central limit theorem. In that case, for fixed ϕ, we can write down the log marginal likelihood for the transformed model to use for inference about $\boldsymbol{\lambda}$. It turns out that, to within an additive constant, this log marginal likelihood is identical to that for the original working model with the false assumption that $\boldsymbol{\epsilon}$ is Gaussian. Furthermore, employing a simple Pearson estimator for ϕ produces exactly the estimate obtained by maximizing that same original marginal likelihood of the working model. So there is decent large-sample justification for using the marginal likelihood of the working linear model for smoothing and scale parameter estimation.

Unfortunately this approach is not guaranteed to converge. The iteration can end up in a cyclic state, and on rare occasions it does. The $\hat{\boldsymbol{\lambda}}$ produced by this method do not usually coincide exactly with the nested iteration estimates: The dependence of $\mathbf{W}$ on $\boldsymbol{\rho}$ is entirely neglected at each $\boldsymbol{\rho}$ optimization.

4.4. The Extended Fellner–Schall Method

Another approach to $\boldsymbol{\lambda}$ optimization is based on a method for estimating variance components first proposed by Fellner (1986) and Schall (1991) and generalized by Wood & Fasiolo (2017). It starts from consideration of the derivatives of the marginal likelihood

$$2\frac{\partial V_r}{\partial \lambda_j} = -\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}} + \text{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \text{tr}\{(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} - \text{tr}\{(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\partial\hat{\mathcal{I}}/\partial\lambda_j\} = -a + b - c$$

by definition of terms a, $b = \text{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j)/2 - \text{tr}\{(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$, and c. Wood & Fasiolo (2017) show that $b \geq 0$ when $\hat{\mathcal{I}}$ is positive semidefinite (if it is not, then the expected version, or an estimate of it, can be substituted). Now λ_j should be increased or decreased according to whether $\partial V_r/\partial\lambda_j$ is respectively positive or negative, and it is easy to see that the following update achieves this,

$$\lambda_j^* = \begin{cases} \lambda_j(b-c)/a, & c \leq 0, \\ \lambda_j b/(a+c), & c > 0, \end{cases}$$

and that the update is an ascent direction. Hence, we can estimate the model by alternating optimization for $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{\lambda}$ values, with these updates of $\boldsymbol{\lambda}$ given $\hat{\boldsymbol{\beta}}$ values. Doing so will maximize V_r. Formally we need step length control to guarantee not overshooting, but in practice this rarely seems to be needed, probably because, near convergence, the updates lie between EM (expectation–maximization) and Newton steps.

The c term requires $\mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\lambda_j$ to be evaluated, but an alternative is to set $c = 0$ initially and, when near convergence, to switch to updating one λ_j at a time. Then c can be estimated for that λ_j by finite differencing, the estimate being carried forward to the next update of λ_j. Hence, only the second derivative information required to find $\hat{\boldsymbol{\beta}}$ is needed to also estimate the smoothing parameters.

Going even further we can follow what is done for performance iteration/penalized quasi-likelihood and simply neglect $\partial\hat{\mathcal{I}}/\partial\lambda_j$, setting $c = 0$, so that the update becomes $\lambda_j^* = \lambda_j b/a$—that is,

$$\lambda_j^* = \lambda_j \frac{\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \mathrm{tr}\{(\hat{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}}{\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}}.$$

This can even be used without an explicit exact $\hat{\mathcal{I}}$. Instead $\hat{\boldsymbol{\beta}}$ can be obtained using quasi-Newton optimization, requiring only first derivatives of the log likelihood. At convergence $\hat{\mathcal{I}}$ is then approximated by finite differencing. Such a scheme only has the $O(np^2)$ cost of other methods, but it is appealing in the context of nonstandard nonlinear models formulated in terms of smooth functions, where second derivatives maybe implementationally tedious or computationally prohibitive to obtain (e.g., Wood & Wit 2021). The expression $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j)$ is purely formal. Given the block diagonal structure of $\mathbf{S}_\lambda$, its pseudoinverse is not formed explicitly. For single smoothing parameter terms, $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) = \mathrm{rank}(\mathbf{S}_j)/\lambda_j$, and otherwise computation is blockwise.

4.5. Bigger Data and Models

Very large datasets can be handled rather easily if $p = \dim(\boldsymbol{\beta})$ is not large. All that is needed is some efficient way of accumulating the QR decomposition used in Section 4.3 without explicit formulation of the model matrix, and the remaining computations are then $O(p^3)$ (Wood et al. 2015). It is easy to develop many alternatives. However, the more useful case is when p and n are large. An effective approach uses the performance iteration of Section 4.3 with three enhancements.

1. To the maximum extent possible, base matrix computation on operations dominated by block–block (BLAS level 3) operations, such as Cholesky decomposition, rather than QR or eigen decompositions, with higher matrix–vector (BLAS level 2) loads. This best exploits modern computer hardware, which tends to be memory-bandwidth limited (retrieval of data from memory is typically an order of magnitude slower than a floating point operation).
2. Accelerate the performance iteration by only making single updates of $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ at each step, rather than iterating them to convergence for a working model discarded at the next step.
3. Exploit the natural discretization of many covariates, and finely discretize others, to reduce the operation count associated with products involving the model matrix.

As an example of point 3, consider the matrix vector product $\mathbf{X}^\mathsf{T}\mathbf{y}$, of the type occurring repeatedly in GAM fitting, where $\mathbf{X}$ is the model matrix for a smooth of one variable, x, and $\mathbf{y}$ is a response vector. The product costs $O(np)$ operations to form. Now suppose that x only has $m \ll n$ unique values, either because it is recorded to limited precision (e.g., temperature recorded to the nearest tenth of a degree) or because we discretize; $m = \Theta(n^{1/2})$ should always be acceptable in statistical regression. If $\mathbf{x}$ is the covariate vector, let $\bar{\mathbf{x}}$ be the vector of its unique values and let $k(i)$ be an index such that $x_i = \bar{x}_{k(i)}$. In that case an $m \times p$ matrix $\bar{\mathbf{X}}$ can also be formed, containing the unique rows of $\mathbf{X}$, such that $\mathbf{X}_i = \bar{\mathbf{X}}_{k(i)}$. Forming $\bar{\mathbf{y}}$ where

$$\bar{y}_j = \sum_{\{i:k(i)=j\}} y_i,$$

then $\mathbf{X}^\mathsf{T}\mathbf{y} = \bar{\mathbf{X}}^\mathsf{T}\bar{\mathbf{y}}$, but the right-hand side has $O(mp)$ cost, which is less than the $O(n)$ cost of forming $\bar{\mathbf{y}}$ in the usual case with $m = O(n^{1/2})$ and $p = O(n^{1/5})$, so computing this way is a factor of p more efficient than direct formation of $\mathbf{X}^\mathsf{T}\mathbf{y}$, and we also save substantial storage by forming $\bar{\mathbf{X}}$ and index vector $\mathbf{k}$ instead of $\mathbf{X}$. Weighted cross products $\mathbf{X}_j\mathbf{W}\mathbf{X}_k$ are also needed, where $\mathbf{X}_j$ and $\mathbf{X}_k$ may be model matrices for different smooth terms. In addition it is necessary to deal with the common case (e.g., Section 6.2) in which the model matrix for a term has the form $\mathbf{X} = \mathbf{X}_1 \otimes_r \mathbf{X}_2 \otimes_r \cdots \otimes_r \mathbf{X}_d$, where $\mathbf{X}_j$ denotes a marginal model matrix and $\otimes_r$ the Kronecker product applied row-wise; to maintain accuracy, we then form matrices $\bar{\mathbf{X}}_j$, rather than attempting to form $\bar{\mathbf{X}}$. Wood et al. (2017) and Li & Wood (2020) provide the required algorithms.

5. LOCATION SCALE AND SHAPE, OTHER LOSSES, LINEAR FUNCTIONALS

The methods developed above for regression on a location parameter are rather generally formulated and hence apply equally well if we wish to also model parameters other than the mean. Generically we might wish to work with models of the form (Rigby & Stasinopoulos 2005)

$$y_i \sim \mathcal{D}(\boldsymbol{\theta}^i), \quad \text{where} \quad g_j(\theta^i_j) = \eta^{ij} = \mathbf{M}^{ij}\gamma^j + \sum_k f_{jk}(x_{ik}).$$

Here conditionally independent y_i are observations from some distribution $\mathcal{D}$ with parameter vector $\boldsymbol{\theta}^i$, each element of which is modeled with a GAM structure. The f_{jk} are represented using spline-type basis expansions with smoothing penalties/priors as before, and the same penalized likelihood/empirical Bayes methods are applicable. Cox (1972) regression, with the partial likelihood substituted for the likelihood, is covered by the same basic methods.

A less straightforward generalization is to other loss functions, such as that used for quantile regression. The coherent belief updating framework of Bissiri et al. (2016) provides a way of extending the empirical Bayes approach to this setting. Essentially the loss (or log loss) takes the role of the log likelihood to within a scaling parameter, the learning rate, that determines the relative weighting of loss and prior. Given the learning rate, the estimation methods follow those described above, but the learning rate must also be chosen. Using an optimally smoothed quantile loss, Fasiolo et al. (2021) choose the learning rate to optimize the coverage probability of the implied Bayesian confidence intervals. An alternative is to use cross validation, since the general form of V_c in Equation 7 remains valid if another sufficiently smooth loss function is substituted for the negative log likelihood, and can be computed in the same way.

Another straightforward extension is to allow the linear predictor(s) to depend on not only smooth functions evaluated at single covariate values but also general bounded linear functionals of unknown smooth functions. Prototypical examples are deconvolution and signal regression–type problems, such as when a spectrum is observed as a predictor of a response. The simplest example model is then (Marx & Eilers 1999)

$$y_i = \int f(x)k_i(x)\mathrm{d}x + \epsilon_i,$$

where $k_i(x)$ is the function giving intensity at frequency x, and f is an unknown smooth function. The function $k_i(x)$ is usually observed on a discrete grid, turning the integral into a sum, so given a basis for f, the integral is straightforward to handle: The term involving f results in a model matrix, coefficient vector, and penalty, just like any other smooth term. Another simple example is provided by varying coefficient (Hastie & Tibshirani 1993) or geographic regression models in which terms like $zf(x)$ arise, with both x and z being covariates.

6. SMOOTH MODEL COMPONENTS

A very wide range of model components can be represented using basis expansions and quadratic penalties, and this section briefly touches on the most important for general applied use. The fact that each term has a model matrix and a quadratic penalty/Gaussian prior means that simple Gaussian random effects can be handled using the same methods as smooth terms (with the minor adjustments to p-value computations noted in Section 3.3.3).

There are numerous ways to set up one-dimensional smooth terms. Different orders of differentiation can be used to define the spline penalty, and a popular alternative is to combine an evenly spaced B-spline basis with a discrete difference penalty on the basis coefficients: e.g., $\sum_j(\beta_{j-1} - 2\beta_j + \beta_{j+1})^2$. The latter (P-splines; Eilers & Marx 1996) are popular when using MCMC for inference because of the link between difference penalties and random walks of various orders. It is also easy to mix different orders of basis and penalty this way, although the same is also possible for ordinary spline penalties (e.g., Wood 2017b). P-splines also make it very easy to construct adaptive splines in which the degree of smoothness is allowed to vary with the covariate (i.e., along the spline), controlled by multiple smoothing parameters. Note that a first-order penalty (first-derivative-based or first-difference-based) actually implies a quite rough $f(x)$.

Smooth functions of more than one covariate are more interesting. Here there are two fundamentally different possibilities. The first extends the one-dimensional spline penalty to multiple dimensions in an isotropic manner, with penalization controlled by a single smoothing parameter. Such terms are often felt to be natural for representing the effect of spatial location, but they are inappropriate for modeling general smooth interactions where covariates may not even be measured in the same units, so that the notion of smoothness depends on arbitrary decisions about relative scaling.[10] In this case progress can be made by simple application of the statistical notion of an interaction to the smooth function case.

6.1. Duchon and Thin Plate Splines for Isotropic Smoothing

Consider smoothing with respect to x_1 and x_2. The cubic spline penalty can be generalized to

$$J_{22} = \int \left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 \mathrm{d}\mathbf{x},$$

or, more generally, to d-dimensional smoothing with order m derivative penalization as

$$J_{md} = \int \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\frac{\partial^m f}{x_1^{\nu_1} \dots x_d^{\nu_d}}\right)^2 \mathrm{d}\mathbf{x}.$$

Existence of the corresponding $\hat{f}$ requires $2m > d$, with the result only visually smooth for $2m > d + 1$. These restrictions lead to rather rapid growth in the dimension of the penalty null space with d (56-dimensional by $d = 5$, 3,003-dimensional by $d = 10$), but Duchon (1977) reexpressed J_{md} in the Fourier domain using Plancherel's theorem while adding an extra penalization term $\|\boldsymbol{\tau}\|^{2s}$ to give

$$J_{mds} = \int \|\boldsymbol{\tau}\|^{2s} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\mathcal{F}\frac{\partial^m f}{x_1^{\nu_1} \dots x_d^{\nu_d}}(\boldsymbol{\tau})\right)^2 \mathrm{d}\boldsymbol{\tau},$$

[10] A superficially appealing approach is to scale all covariates to the unit interval. That this is a poor solution is seen by thinking about the case where the data are much more variable with respect to one covariate than to another.

where $\boldsymbol{\tau}$ is frequency and $\mathcal{F}$ denotes Fourier transform. Under the restrictions $-d/2 < s < d/2$ and $m + s > (d+1)/s$, the resulting $\hat{f}$ exists and is continuous to first derivative, having the form

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{msd}(\|\mathbf{x} - \mathbf{x}_i\|) + \sum_j \alpha_j \phi_j(\mathbf{x}),$$

where α_j and δ_i are parameters to be estimated, the $\phi_j(\mathbf{x})$ define a basis for polynomials of order less than m (i.e., a basis of the penalty null space) and

$$\eta_{msd}(t) = \begin{cases} (-1)^{m-s-(d-1)/2}|t|^{2m-2s-d}, & 2m-2s-d \text{ odd}, \\ (-1)^{m-s-d/2}|t|^{2m-2s-d}\log|t|, & 2m-2s-d \text{ even}; \end{cases}$$

we take $\eta_{msd}(0) = 0$. Define $T_{ij} = \phi_j(\mathbf{x}_i)$ and $E_{ij} = \eta_{mds}(\|\mathbf{x}_i - \mathbf{x}_j\|)$. Then for a least squares loss (log likelihood), and smoothing parameter λ, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\delta}}$ are the minimizers of

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{E}\delta\|^2 + \lambda\boldsymbol{\delta}^{\mathsf{T}}\mathbf{E}\boldsymbol{\delta} \text{ subject to } \mathbf{T}^{\mathsf{T}}\boldsymbol{\delta} = \mathbf{0}.$$

In practice the linear constraint can be incorporated by reparameterization using the QR-based method given in Section 3.1, while $\hat{f}$ has the same form when an alternative log likelihood is used rather than the residual sum of squares.

Rank reduction can be achieved as in the one-dimensional case by selecting a nicely spread out set of $k < n$ knots from the $\mathbf{x}_i$, and then using the corresponding basis and penalty for the whole dataset. Section 2.3 provides an alternative, but the fact that $\mathbf{E}$ defines both the penalty and most of the model matrix offers a more elegant solution. Matrix $\mathbf{E}$ can be replaced by its rank k eigen approximation, which can be obtained efficiently by Lanczos iteration (Wood 2003). The hybrid approach suggested in Section 2.3 can equally be applied here. Obviously, given a basis and quadratic penalty, such smooth terms can be incorporated into a GAM like any one-dimensional smooth term.

6.2. Smooth Interactions: Tensor Product Splines

A statistical interaction occurs when the effect of one covariate is modified by others. In parametric terms, this means that the coefficients describing the effect of one covariate themselves change with another covariate. Generalization to smooth effects is straightforward. For a smooth interaction we require the smooth effect of one covariate to vary smoothly with another covariate. Given bases for the smooth effects of each covariate separately, we allow each coefficient of one covariate's smooth to itself vary smoothly with the other covariate, as shown in **Figure 4**.

To illustrate the construction, consider a smooth interaction of variables x and z. Start by constructing bases and penalties for one-dimensional functions: $f^x(x) = \sum_{j=1}^{k_x} \beta_j^x b_j^x(x)$ and $f^z(z) = \sum_{j=1}^{k_z} \beta_j^z b_j^z(z)$ with penalties $\boldsymbol{\beta}^{x\mathsf{T}}\mathbf{S}^x\boldsymbol{\beta}^x$ and $\boldsymbol{\beta}^{z\mathsf{T}}\mathbf{S}^z\boldsymbol{\beta}^z$ (here, superscripts are labels, not powers). To allow f^z to vary smoothly with x, we can set $\beta_j^z(x) = \sum_{i=1}^{k_x} \beta_{ij}^x b_i^x(x)$, as illustrated in **Figure 5**. On substitution into the expression for f_z, we get

$$f(x, z) = \sum_{i,j} \beta_{ij} b_i^x(x) b_j^x(x),$$

so the basis functions for the interaction are all possible products of the marginal basis functions.

It is important to ensure that the penalty is invariant to the relative scaling of the covariates. This will be the case if we penalize separately in the x and z directions, with a separate smoothing parameter for each direction. A particularly simple approach sums up the marginal x penalty along the curves traced out by the smoothly varying coefficients of f_z (thick curves parallel to the x axis in

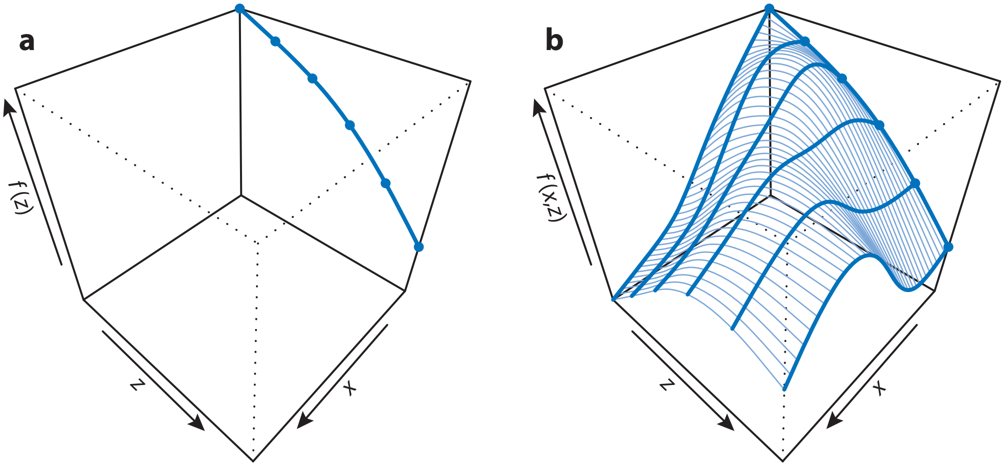

Figure 4

Smooth interaction concept. (*a*) A smooth function of z parameterized in terms of function values at evenly spaced knots (*filled circles*). (*b*) The smooth of z is made to vary smoothly with x by making its coefficients vary smoothly with x (*thick curves*). The thin curves illustrate how this defines a smooth function of x and z.

Figure 5*b*). This gives the penalty on wiggliness in the x direction. A similar construction produces a penalty for wiggliness in the z direction. Mathematically these penalties are $\lambda_x \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}^x \otimes \mathbf{I}_{k_z} \boldsymbol{\beta}$ and $\lambda_z \boldsymbol{\beta}^{\mathsf{T}} \mathbf{I}_{k_x} \otimes \mathbf{S}^z \boldsymbol{\beta}$, under appropriate arrangement of the β_{ij} into a single vector $\boldsymbol{\beta}$. Generalization to more than two covariates is easy. Isotropic smooths of more than one variable can equally well be used as marginals: This provides a useful method for spatiotemporal smoothing using a marginal thin plate spline of spatial location and a marginal cubic spline of time, for example.

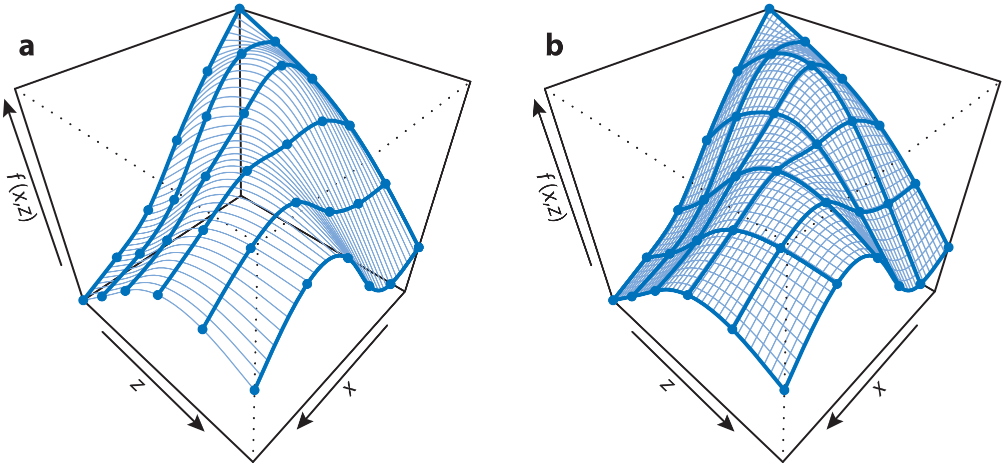

Figure 5

Smooth interaction implementation. (*a*) Each coefficient of the smooth of z is represented as a smooth function of x using the basis for smoothing with respect to x. (*b*) The construction is symmetric whether we start by considering the smooth of x or the smooth of z. Separate penalties in the x and z directions can be constructed by summing the marginal penalties applied along the thick curves in the x and z directions, respectively.

It is often appealing to decompose effects into main effects plus (pure) interactions, e.g., to write $f(x, z) = f_x(x) + f_z(z) + f_{xz}(x, z)$. Construction of an appropriate f_{xz}, excluding the main effects, is easy and follows directly from the usual way that pure interaction terms are constructed in regression models. All that is required is to absorb the sum-to-zero constraint into the marginal bases for f_z and f_x before applying the tensor product smooth construction. Elimination of the constant functions from the margins ensures that main effects are not reproduced when the tensor product basis is constructed.

6.3. Spatial Smoothers

Several special smoothers are available for spatial data. When data cover a significant portion of the globe, thin plate splines can be adapted to splines on the sphere (Wendelberger 1981). Health and administrative data are often only available aggregated by reporting district. It is then common to have a coefficient for each district, but to apply a penalty (define a prior) imposing smooth variation of the coefficients between districts. These Markov random fields (e.g., Rue & Held 2005) can be treated in the same way as any other smooth term in the model, including the Section 2.3 rank reduction. Similarly, many Gaussian process models can be expressed in basis-penalty form exactly like a spline smoother (Kammann & Wand 2003), with rank reduction working in the same way as for Duchon/thin plate splines. Sometimes data are collected within a complicated geographic (or other) region where it is important not to smooth across the boundary, artificially linking regions that may be close in Euclidean distance but far apart in terms of within-domain separation. Leveraging the link between smoothing and partial differential equations, a penalty and smoother can then be constructed that exist only within the domain of interest and avoid such artificial linkages (e.g., Wood et al. 2008).

7. PRACTICAL EXAMPLE

One application of GAMs is in air pollution modeling. **Figure 6*a*** shows daily deaths in Chicago starting in 1995, from Peng & Welty (2004) and available as data frame `chicago` in R package `gamair`. Along with the daily death count, ozone, temperature, and PM$_{10}$ (particles with diameters of 10 μm and smaller) measurements are available for most days. The short period of very high deaths occurs in a heat wave, but such high temperatures also occur at other times without such high deaths. An obvious model to try is

$$\texttt{death}_i \sim \text{Poi}(\mu_i), \text{ where } \log\mu_i = \alpha + f_1(\texttt{day}_i) + f_2(\texttt{o3}_i) + f_3(\texttt{PM10}_i) + f_4(\texttt{tmp}_i).$$

The Poisson rate parameter is modulated by a background time effect, representing all the effects modifying death rate other than the measured covariates, and modulation effects for each covariate. The effects act multiplicatively on the death rate. In the package `mgcv`, supplied with R, such a model (with basis dimension 200 for f_1) can be estimated using

```
m0 <- gam(death~s(time,k=200)+s(o3)+s(pm10)+s(tmp),family=poisson)
```

but standard plots of deviance residuals against day or fitted values ($\hat{\mu}_i$) indicate that this model is untenable. The high deaths remain as enormous outliers.

Two biologically motivated modifications seem sensible. Firstly, the stress effects of high ozone may interact with temperature. Secondly, pollution and temperature effects are unlikely to act instantaneously, but rather to accumulate over several days. A model capturing these modifications is

$$\log(\mu_i) = \alpha + f_1(\texttt{day}_i) + \sum_{\texttt{lag}=0}^{3} f_2(\texttt{PM10}_{i-\texttt{lag}}, \texttt{lag}) + \sum_{\texttt{lag}=0}^{3} f_3(\texttt{o3}_{i-\texttt{lag}}, \texttt{tmp}_{i-\texttt{lag}}, \texttt{lag}),$$

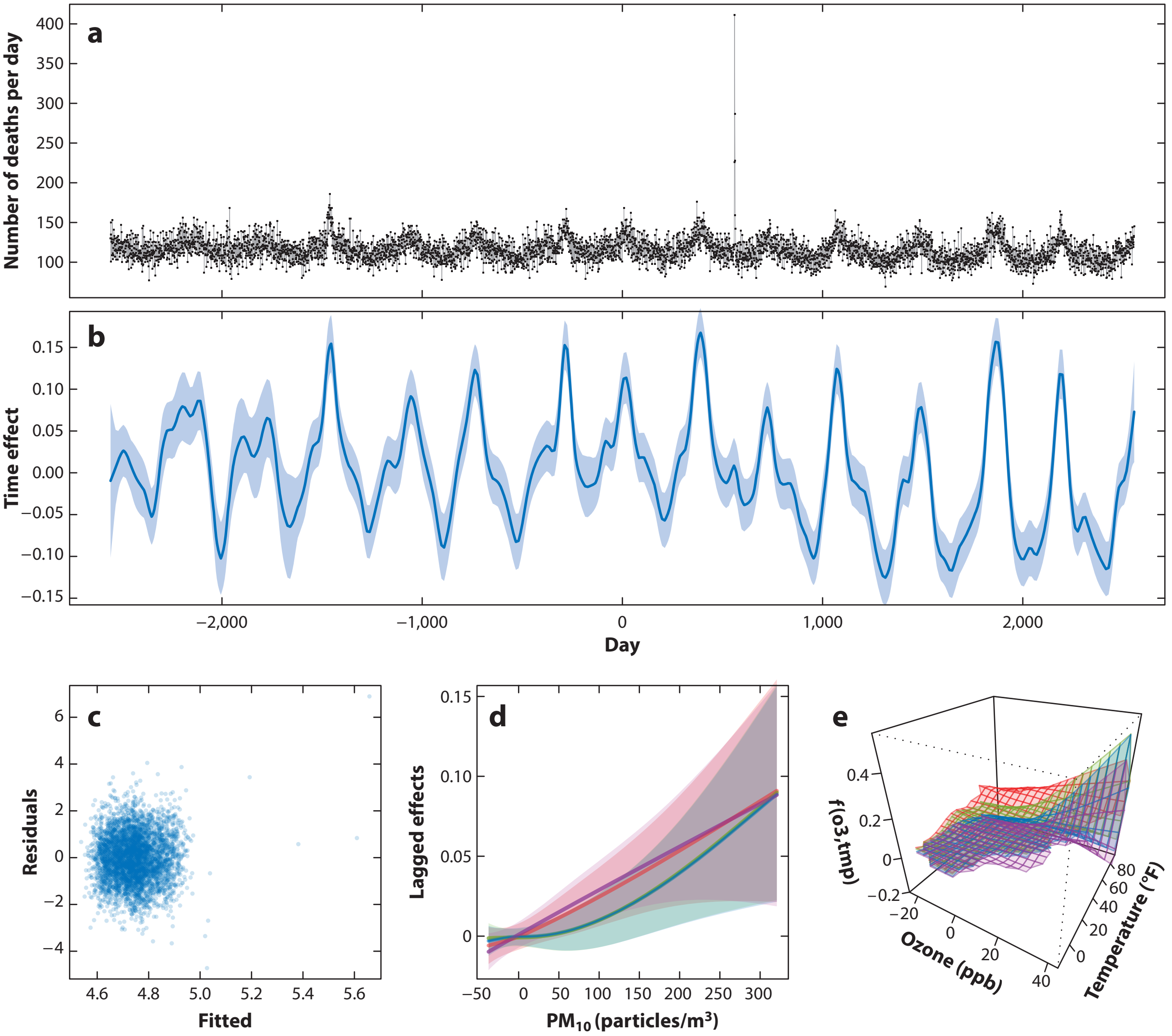

Figure 6

(*a*) Daily deaths in Chicago. (*b*) Estimated background death variation. (*c*) Deviance residuals versus fitted values. (*d*) PM_{10} effects at lag 0 (*red*), 1 (*green*), 2 (*blue*), and 3 (*pink*). (*e*) Smooth temperature–ozone interaction effects at lags 0–3 color coded as in panel *d*. Ozone and PM_{10} measurements are centered.

where f_2 and f_3 are represented using tensor product smooth interaction terms. So the effects of lagged covariates are additive on the log scale, while changing smoothly with lag. This model can also be estimated using `gam`. Tensor product splines are specified using `te()` terms in the model specification formula. The required summations are obtained by supplying matrices to the relevant smooth, which is then evaluated for each row and column of the supplied matrix or matrices; the results for each row are then summed across the columns (summation weights can be supplied but are not needed here). First set up an $n \times 4$ matrix, `lag`, each row of which is simply (0, 1, 2, 3), and $n \times 4$ matrices `PM10`, `o3`, and `tmp`, each of which contains unlagged covariate observations in its first column, and increasingly lagged versions in subsequent columns. Then the model can be estimated in R:

```
m=gam(death~s(time,k=200)+te(pm10,lag,k=c(10,4),bs=c("cr","tp"),m=1)
 +te(o3,tmp,lag,k=c(8,8,4)),family=poisson,method="REML")
```

The `k` argument to `te` specifies the marginal basis dimensions, while the `bs` argument allows the marginal basis types to be specified. Argument `m` allows the order of penalization to be specified for bases where there is a choice. It is used here in order to force identifiability, as both tensor product terms contain a main effect of `lag`—the unpenalized components of these main effects can not be identified, but with $m = 1$ there is no unpenalized component. **Figure *6b–e*** shows the effect estimates for the model, with deviance residuals versus fitted values in panel *c*. The highest daily death count is still an outlier, but far less so than with the simple model. Using negative binomial or Tweedie distributions leads to slightly higher AIC values [using `AIC(m)`] and no improvement in residuals. The tests from Section 3.3.3 are performed as part of `summary(m)` and all terms are significant, with the PM_{10} effect the least so. However, replacing it by either $\sum_{\texttt{lag}=0}^{3} f_2(\texttt{PM10}_{i-\texttt{lag}})$, or by a simple instantaneous effect, leads to substantially higher AIC values. The same goes for the ozone–temperature interaction. The `mgcv` package offers a default `plot` method, but **Figure 6** was created specifically for this example, using the `predict` method function to produce the plot data.

The striking result is the very strong positive effect on daily deaths of a combination of high ozone and high temperature sustained over several days. Also notable is how variable the clear seasonal variation in death rates is: Neither phase nor amplitude are stable, emphasizing the need for flexibility. For the day of peak deaths, the model still predicts only 284 deaths, rather than the 411 observed, but the overall fit is still much better than that of simpler models.

8. CONCLUSIONS

This article has attempted to convey the fullest possible understanding, in the space available, of the empirical Bayes approach to GAMs based on reduced rank splines. The approach gives a reasonably complete and coherent set of applied modeling tools, comparable to that available for parametric GLMs, which has led to wide uptake of the models in areas as diverse as energy forecasting, agricultural production management, environmental modeling, medicine, and finance. But the approach has limitations when high-rank spatial models are required, when tail probabilities poorly captured by Gaussian posterior approximations are needed, or when large numbers of covariates must be screened, or larger model-data combinations than those manageable via Section 4.5's methods are needed.

Some of these limitations are addressable using approaches to GAMs that space has precluded discussing here. Especially notable is integrated nested Laplace approximation (INLA; Rue et al. 2009), which efficiently deals with cases in which low-rank approximation is inappropriate and/or one may be interested in tail probabilities poorly captured by simple Gaussian approximation of the posterior. It does this by extensive development of the sort of sparse direct approach to smoothing discussed in Section 2, while also making effective use of the close relationship between differential equations and smoothers (Lindgren et al. 2011). Equally effective when the Gaussian approximations are inadequate are MCMC approaches taking a direct stochastic simulation approach to the models (e.g., Fahrmeir & Lang 2001, Umlauf et al. 2015). The original backfitting method shares similarities with the boosting approach that builds smooth estimates as the sum of estimates obtained from weak learners iteratively fitted to continuously updated residuals (e.g., Schmid & Hothorn 2008, Mayr et al. 2012), but smoothness selection under boosting amounts only to deciding when to terminate the iteration, with relative degrees of smoothness of model components emerging automatically. Large numbers of covariates are unproblematic for boosting: Weak learners for irrelevant smooth terms are often never selected. Wood (2020) provides an introduction to these approaches and further references.

Several aspects of generalized additive modeling remain challenging. Confidence intervals with good pointwise performance would be a substantial improvement on the across-the-function intervals discussed above. Also desirable are p-value approximations with better theoretical properties, that cope well with highly correlated effects and highly uncertain smoothing parameters. The issue of spatial confounding (e.g., Hodges & Reich 2010), in which spatial smooth effects undermine inference about covariate effects, has recently seen progress (e.g., Dupont et al. 2022, Marques et al. 2023), but methods for general smooth covariate effects are still needed. While the first-order Laplace approximation to the marginal likelihood in fact offers higher-order approximation for the smoothing parameters of the penalty, the same is not true for parameters of the likelihood, such as the Gaussian or Gamma scale parameters or the variance-controlling parameters of the Tweedie or negative binomial distributions, and here improvements are also desirable. A related issue for smooth additive location-scale models is the bias inherent in using the likelihood directly for inference about scale parameters, an issue raised by Nelder in the discussion of Rigby & Stasinopoulos (2005) but still outstanding. For effect screening, boosting (e.g., Schmid & Hothorn 2008) is attractive, but becomes computationally expensive when high-dimensional spatial effects are needed, and smooth interaction terms are also nontrivial to handle. Finally, it remains the case that a far wider variety of theoretically justifiable smooth regression models can be written down, or imagined, than can readily be computed with. Computation remains a major limiting factor. For example, the methods of Section 4.5 remain to be generalized to the models of Section 5, and perhaps most interesting is the open question of how far the method computational costs might be reducible from $O(np^2)$ toward $O(np)$. Very large gains in the applicability of smooth regression would result from such developments.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

My work on GAMs has benefited greatly from several people. David Borchers first suggested implementing the methods in S/R. Kurt Hornick and the R team then encouraged the development of the `mgcv` package, with many users contributing bug reports and suggestions. Mark Bravington introduced me to implicit differentiation and provided the `debug` package, without which `mgcv` would be much worse. Phil Reiss convinced me of the merits of REML. This article has been much improved by detailed comments from reviewers.

LITERATURE CITED

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, ed. B Petran, F Csaaki, pp. 267–81. Budapest: Akadeemiai Kiadi

Bissiri PG, Holmes C, Walker SG. 2016. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B* 78(5):1103–30

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88:9–25

Chouldechova A, Hastie T. 2015. Generalized additive model selection. arXiv:1506.03850 [stat.ML]

Claeskens G, Krivobokova T, Opsomer JD. 2009. Asymptotic properties of penalized spline estimators. *Biometrika* 96(3):529–44

Cox D. 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc. Ser. B* 34(2):187–220

Crainiceanu C, Ruppert D, Claeskens G, Wand MP. 2005. Exact likelihood ratio tests for penalised splines. *Biometrika* 92(1):91–103

Craven P, Wahba G. 1979. Smoothing noisy data with spline functions. *Numer. Math.* 31(5):377–403

Davies RB. 1973. Numerical inversion of a characteristic function. *Biometrika* 60(2):415–17

Davies RB. 1980. Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. *J. R. Stat. Soc. Ser. C* 29(3):323–33

de Boor C. 2001. *A Practical Guide to Splines*. New York: Springer. Rev. ed.

Duchon J. 1977. Splines minimizing rotation-invariant semi-norms in Solobev spaces. In *Construction Theory of Functions of Several Variables*, ed. W Schemp, K Zeller, pp. 85–100. Berlin: Springer

Dupont E, Wood SN, Augustin NH. 2022. Spatial+: a novel approach to spatial confounding. *Biometrics* 78(4):1279–90

Eilers PHC, Marx BD. 1996. Flexible smoothing with *B*-splines and penalties. *Stat. Sci.* 11(2):89–121

Ettinger B, Perotto S, Sangalli LM. 2016. Spatial regression models over two-dimensional manifolds. *Biometrika* 103(1):71–88

Fahrmeir L, Lang S. 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. R. Stat. Soc. Ser. C* 50(2):201–20

Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y. 2021. Fast calibrated additive quantile regression. *J. Am. Stat. Assoc.* 116(535):1402–12

Fellner WH. 1986. Robust estimation of variance components. *Technometrics* 28(1):51–60

Fletcher D. 2012. Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika* 99(1):230–37

Golub GH, Heath M, Wahba G. 1979. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–23

Golub GH, van Loan CF. 2013. *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press. 4th ed.

Greven S, Crainiceanu CM, Küchenhoff H, Peters A. 2008. Restricted likelihood ratio testing for zero variance components in linear mixed models. *J. Comput. Graph. Stat.* 17(4):870–91

Greven S, Kneib T. 2010. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* 97(4):773–89

Gu C. 1992. Cross-validating non-Gaussian data. *J. Comput. Graph. Stat.* 1(2):169–79

Gu C, Wahba G. 1991. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comput.* 12(2):383–98

Hastie T, Tibshirani R. 1986. Generalized additive models (with discussion). *Stat. Sci.* 1(3):297–318

Hastie T, Tibshirani R. 1990. *Generalized Additive Models*. Boca Raton, FL: Chapman & Hall

Hastie T, Tibshirani R. 1993. Varying-coefficient models (with discussion). *J. R. Stat. Soc. Ser. B* 55(4):757–96

He VX, Wand MP. 2024. Bayesian generalized additive model selection including a fast variational option. *AStA Adv. Stat. Anal.* 108:639–68

Heckman NE, Ramsay JO. 2000. Penalized regression with model-based penalties. *Can. J. Stat.* 28(2):241–58

Hodges JS, Reich BJ. 2010. Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.* 64(4):325–34

Kammann EE, Wand MP. 2003. Geoadditive models. *J. R. Stat. Soc. Ser. C* 52(1):1–18

Kimeldorf GS, Wahba G. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* 41(2):495–502

Kuonen D. 1999. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 86(4):929–35

Li Z, Wood SN. 2020. Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Stat. Comput.* 30(1):19–25

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. R. Stat. Soc. Ser. B* 73(4):423–98

Marques I, Kneib T, Klein N. 2023. Mitigating spatial confounding by explicitly correlating Gaussian random fields. *Environmetrics* 34(4):e2801

Marra G, Wood SN. 2012. Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Stat.* 39(1):53–74

Marx BD, Eilers PH. 1999. Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics* 41(1):1–13

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. 2012. Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *J. R. Stat. Soc. Ser. C* 61(3):403–27

McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. London: Chapman & Hall. 2nd ed.

Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *J. R. Stat. Soc. Ser. A* 135(3):370–84

Nychka D. 1988. Bayesian confidence intervals for smoothing splines. *J. Am. Stat. Assoc.* 83(404):1134–43

O'Sullivan FB, Yandall B, Raynor W. 1986. Automatic smoothing of regression functions in generalized linear models. *J. Am. Stat. Assoc.* 81(393):96–103

Peng RD, Welty LJ. 2004. The NMMAPSdata package. *R News* 4(2):10–14

Ramsay JO, Hooker G, Campbell D, Cao J. 2007. Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. R. Stat. Soc. Ser. B* 69(5):741–96

Ramsay T. 2002. Spline smoothing over difficult regions. *J. R. Stat. Soc. Ser. B* 64(2):307–19

Rigby R, Stasinopoulos DM. 2005. Generalized additive models for location, scale and shape (with discussion). *J. R. Stat. Soc. Ser. C* 54(3):507–54

Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman and Hall/CRC

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. Ser. B* 71(2):319–92

Ruppert D, Wand MP, Carroll RJ. 2003. *Semiparametric Regression*. Cambridge, UK: Cambridge Univ. Press

Säfken B, Kneib T, van Waveren CS, Greven S. 2014. A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electron. J. Stat.* 8:201–25

Sangalli LM. 2021. Spatial regression with partial differential equation regularisation. *Int. Stat. Rev.* 89(3):505–31

Schall R. 1991. Estimation in generalized linear models with random effects. *Biometrika* 78(4):719–27

Schmid M, Hothorn T. 2008. Boosting additive models using component-wise P-splines. *Comput. Stat. Data Anal.* 53(2):298–311

Silverman BW. 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Stat. Soc. Ser. B* 47(1):1–52

Stasinopoulos MD, Rigby RA, Heller GZ, Voudouris V, De Bastiani F. 2017. *Flexible Regression and Smoothing: Using GAMLSS in R.* Boca Raton, FL: Chapman and Hall/CRC

Umlauf N, Adler D, Kneib T, Lang S, Zeileis A. 2015. Structured additive regression models: an R interface to BayesX. *J. Stat. Softw.* 63(21):1–46

van der Vorst HA. 2003. *Iterative Krylov Methods for Large Linear Systems*. Cambridge, UK: Cambridge Univ. Press

Wahba G. 1983. Bayesian confidence intervals for the cross validated smoothing spline. *J. R. Stat. Soc. Ser. B* 45(1):133–50

Wahba G. 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Stat.* 13(4):1378–402

Wahba G. 1990. *Spline Models for Observational Data*. Philadelphia: SIAM

Wendelberger J. 1981. *Smoothing noisy data with multidimensional splines and generalized cross validation*. PhD Thesis, Dep. Stat., Univ. Wis., Madison, WI

Wood SN. 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. Ser. B* 62(2):413–28

Wood SN. 2003. Thin plate regression splines. *J. R. Stat. Soc. Ser. B* 65:95–114

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B* 73(1):3–36

Wood SN. 2013a. A simple test for random effects in regression models. *Biometrika* 100(4):1005–10

Wood SN. 2013b. On *p*-values for smooth components of an extended generalized additive model. *Biometrika* 100(1):221–28

Wood SN. 2017a. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.

Wood SN. 2017b. P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Stat. Comput.* 27(4):985–89

Wood SN. 2020. Inference and computation with generalized additive models and their extensions. *Test* 29(2):307–39

Wood SN, Bravington MV, Hedley SL. 2008. Soap film smoothing. *J. R. Stat. Soc. Ser. B* 70(5):931–55

Wood SN, Fasiolo M. 2017. A generalized Fellner–Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* 73(4):1071–81

Wood SN, Goude Y, Shaw S. 2015. Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C* 64(1):139–55

Wood SN, Li Z, Shaddick G, Augustin NH. 2017. Generalized additive models for gigadata: modelling the UK black smoke network daily data. *J. Am. Stat. Assoc.* 112(519):1199–210

Wood SN, Pya N, Säfken B. 2016. Smoothing parameter and model selection for general smooth models (with discussion). *J. Am. Stat. Assoc.* 111:1548–75

Wood SN, Wit EC. 2021. Was $R < 1$ before the English lockdowns? On modelling mechanistic detail, causality and inference about Covid-19. *PLOS ONE* 16(9):e0257455

Yee TW. 2015. *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York: Springer

Yee TW, Wild C. 1996. Vector generalized additive models. *J. R. Stat. Soc. Ser. B* 58(3):481–93