Supplementary material: Smoothing parameter and model selection for general smooth models

Simon N. Wood[0], Natalya Pya[1] and Benjamin Säfken[2]
[0] School of Mathematics, University of Bristol, Bristol BS8 1TW U.K.
[1]Mathematical Sciences, University of Bath, Bath BA2 7AY U.K.
[2]Georg-August-Universität Göttingen, Germany
s.wood@bath.ac.uk

October 10, 2016

A Consistency of regression splines

There is already a detailed literature on the asymptotic properties of penalized regression splines (e.g. Gu and Kim, 2002; Hall and Opsomer, 2005; Kauermann et al., 2009; Claeskens et al., 2009; **?**; Yoshida and Naito, 2014). Rather than reproduce that literature, the purpose of this section and the next is to demonstrate the simple way in which the properties of penalized regression splines are related to the properties of regression splines, which in turn follow from the properties of interpolating splines. We will mostly focus on cubic splines and 'infill asymptotics' in which the domain of the function of interest remains fixed as the sample size increases. We use the expression 'at most $O(n^a)$' as shorthand for '$O(n^b)$ where $b \leq a$', and use $O(\cdot)$ to denote stochastic boundedness when referring to random quantities.

A.1 Cubic interpolating splines

Let $g(x)$ denote a 4 times differentiable function, observed at k points $x_j, g(x_j)$, where the x_j are strictly increasing with j. The cubic spline interpolant, $\hat{g}(x)$, is constructed from piecewise cubic polynomials on each interval $[x_j, x_{j+1}]$ constructed so that $\hat{g}(x_j) = g(x_j)$, the first and second derivatives of $\hat{g}(x)$ are continuous, and two additional end conditions are met. Example end conditions are the 'natural' end conditions $\hat{g}''(x_1) = \hat{g}''(x_k) = 0$ or the 'complete' end conditions $\hat{g}'(x_1) = g'(x_1)$, $\hat{g}'(x_k) = g'(x_k)$. $\hat{g}(x)$ is unique given the end conditions. See figure 1a. A cubic spline interpolant with natural boundary conditions has the interesting property of being the interpolant minimizing $\int g''(x)^2 dx$ (see e.g. Green and Silverman, 1994, theorem 2.3).

Let $h = \max_j(x_{j+1} - x_j)$, the 'knot spacing'. By Taylor's theorem, a piecewise cubic interpolant must have an upper bound on interpolation error $O(h^\alpha)$ where $\alpha \geq 4$. In fact if $g^{(i)}(x)$ denotes the i^{th} derivative of g with respect to x

$$|\hat{g}^{(i)}(x) - g^{(i)}(x)| = O(h^{4-i}), \quad i = 0, \ldots, 3 \tag{1}$$

where x is anywhere in $[x_1, x_k]$ for complete (or deBoor's 'not-a-knot') end conditions, or is sufficiently interior to $[x_1, x_k]$ for natural end conditions. de Boor (2001, chapter 5) provides especially clear derivation of these results, while Hall and Meyer (1976) provides sharp versions.

A.2 Regression splines

The space of interpolating splines with k knots can be spanned by a set of k basis functions. Various convenient bases can readily be computed: for example the B-spline basis functions have compact support, while the j^{th} cardinal basis function takes the value 1 at x_j and 0 at any other knot x_i (see e.g. Lancaster and Šalkauskas, 1986; de Boor, 2001). For the cardinal basis, the spline coefficients are $g(x_j)$, the values of the spline at the knots. Given a set of basis functions and $n > k$ noisy observations of $g(x)$, it is possible to perform spline regression. Agarwal

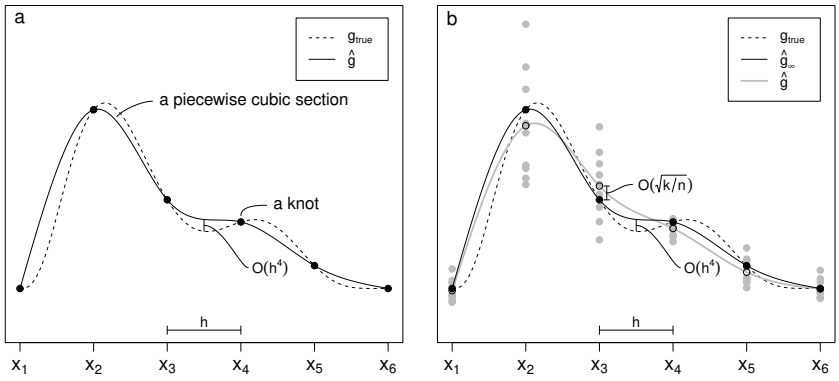

Figure 1: **a.** A cubic interpolating spline (continuous curve), interpolating 6 'knot' points (black dots) with evenly spaced x co-ordinates, from a true function (dashed curve). The spline is made up of piecewise cubic sections between each consecutive pair of knots. The approximation error is $O(h^4)$, where h is the knot spacing on the x axis. **b.** A simple regression spline (grey curve) fitted to n noisy observations (grey dots) of the true function (dashed curve), with n/k data at each of the k knot locations x_j. As $n/k \to \infty$ the regression spline tends to the limiting interpolating spline (black curve), which has $O(h^4) = O(k^{-4})$ approximation error.

and Studden (1980) and Zhou and Wolfe (2000) study this in detail, but a very simple example serves to explain the main results.

Consider the case in which we have n/k noisy observations for each $g(x_j)$, and a model that provides a regular likelihood for $g(x_j)$ such that $|\hat{g}(x_j) - g(x_j)| = O(\sqrt{k/n})$, where $\hat{g}(x_j)$ is the MLE for $g(x_j)$ (which depends only on the n/k observations of $g(x_j)$, as is clear from considering the cardinal basis representation). Suppose also that the x_j are equally spaced. In this setting the cubic regression spline estimate of $g(x)$ is just the cubic spline interpolant of $x_j, \hat{g}(x_j)$, and the large sample limiting $\hat{g}(x)$ is simply the cubic spline interpolant of $x_j, g(x_j)$. By (1) the limiting approximation error is $O(h^4) = O(k^{-4})$. Since the interpolant is linear in the $\hat{g}(x_j)$ the standard deviation of $\hat{g}(x)$ is $O(\sqrt{k/n})$. So if the limiting approximating error is not to eventually dominate the sampling error, we require $O(k^{-4}) \leq O(\sqrt{k/n})$, and for minimum sampling error we would therefore choose $k = O(n^{1/9})$, corresponding to a mean square error rate of $O(n^{-8/9})$ for $g(x)$ and $O(n^{-4/9})$ for g''. See figure 1b.

Agarwal and Studden (1980) shows that the result for $g(x)$ holds when the observations are spread out instead of being concentrated at the knots, while Zhou and Wolfe (2000) confirms the equivalent for derivatives. In summary, cubic regression splines are consistent for $g(x)$ and its first 3 derivatives, provided that the maximum knot spacing decreases with sample size, to control the approximation error. Optimal convergence rates are obtained by allowing h to depend on n so that the order of the approximation error and the sampling variability are equal.

B Penalized regression spline consistency under LAML

Here we show how penalized regression spline estimates retain consistency under LAML estimation of smoothing parameters. To this end it helps to have available a spline basis for which individual coefficients form a meaningful sequence as the basis dimension increases, so we introduce this basis first, before demonstrating consistency and then considering convergence rates.

B.1 An alternative regression basis

An alternative spline basis is helpful in understanding how penalization affects consistency of spline estimation. Without loss of generality, restrict the domain of $g(x)$ to $[0, 1]$ and consider the spline penalty $\int g^{(m)}(x)^2 dx = \int (\nabla^m g)^2 dx$ where ∇^m is the m^{th} order differential operator. Let ∇^{m*} be the adjoint of ∇^m with respect to the inner product $\langle g, h \rangle = \int g(x)h(x)dx$. Then from the definition of an adjoint operator, $\int g^{(m)}(x)^2 dx = \int g\mathcal{K}^m g dx$, where $\mathcal{K}^m = \nabla^{m*}\nabla^m$. Now consider the eigenfunctions of $\mathcal{K}^m$, such that $\mathcal{K}^m \phi_j(x) = \Lambda_j \phi_j(x)$,

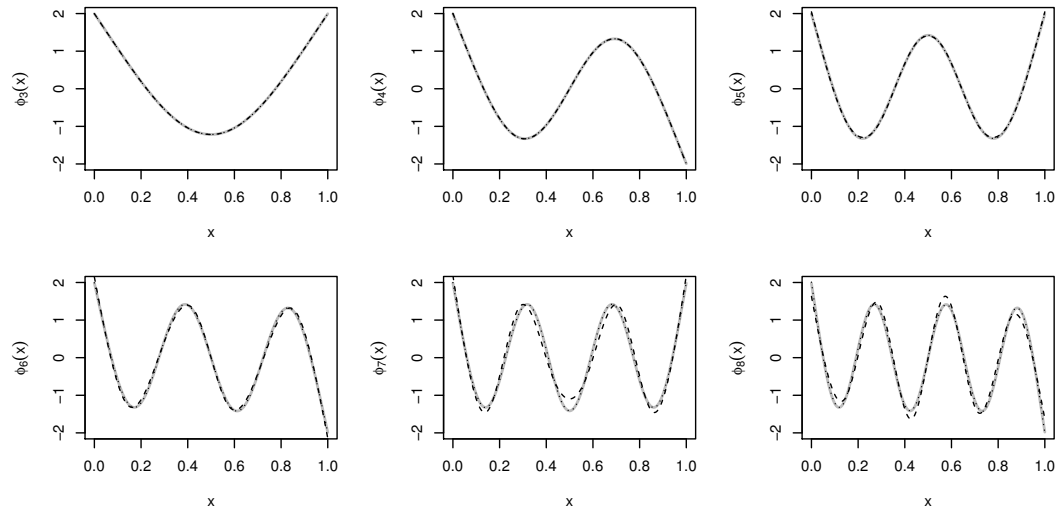

Figure 2: Eigenfunctions of $\mathcal{K}^2$ shown in grey, with Demmler-Reinch spline basis functions overlaid in black. The first two linear functions are not shown. The dashed curves are for a rank 8 cubic spline basis, while the dotted curve, exactly overlaying the grey curves, are for a rank 16 cubic spline basis.

$\Lambda_{j+1} > \Lambda_j \geq 0$. Since $\mathcal{K}^m$ is clearly self adjoint, $\langle \phi_j, \phi_i \rangle = 1$ if $i = j$ and 0 otherwise. Notice that if $\beta_i^* = \langle g, \phi_i \rangle$, then we can write $g(x) = \sum_i \beta_i^* \phi_i(x)$. Finite $\int g^{(m)}(x)^2 dx$ implies that $\beta_i^* \to 0$ as $i \to \infty$. In fact generally we are interested in functions with low $\int g^{(m)}(x)^2 dx$, so it is the low order eigenvalues and their eigenfunctions that are of interest.

To compute discrete approximations to the ϕ_j, first define $\Delta = (n-1)^{-1}$ for some discrete grid size n, and let $\phi_{ji} = \phi_j(i\Delta - \Delta)$ and $g_i = g(i\Delta - \Delta)$. A discrete representation of $\mathcal{K}^2$ is then $\mathbf{K} = \mathbf{D}^\mathsf{T}\mathbf{D}$ where $D_{ij} = 0$ except for $D_{i,i} = D_{i,i+2} = 1/\Delta^2$ and $D_{i,i+1} = -2/\Delta^2$ for $i = 1, \ldots, n-2$ (the approximation for other values of m substitutes m^{th} order differences in the obvious way). The (suitably normalized) eigenvectors of $\mathbf{K}$ then approximate $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots$. Alternatively we can represent $\boldsymbol{\phi}_1 \ldots \boldsymbol{\phi}_k$ and any other $\mathbf{g}$ using a rank k cubic spline basis. Hence we can write $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ has QR decomposition, $\mathbf{X} = \mathbf{QR}$ and $\int g^{(m)}(x)^2 dx = \boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta} = \boldsymbol{\beta}^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{Q}\mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1}\mathbf{Q}^\mathsf{T}\mathbf{Q}\mathbf{R}\boldsymbol{\beta}$. So the approximation of $\mathcal{K}^2$ is $\mathbf{Q}\mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1}\mathbf{Q}^\mathsf{T}$, which has eigenvectors $\mathbf{QU}$ where $\mathbf{U}$ is from the eigen-decomposition $\mathbf{U}\tilde{\mathbf{\Lambda}}\mathbf{U}^\mathsf{T} = \mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1}$.

Now if we reparameterize the regression spline basis so that $\boldsymbol{\beta}^* = \Delta^{1/2}\mathbf{U}^\mathsf{T}\mathbf{R}\boldsymbol{\beta}$, we obtain a normalized version of the Demmler-Reinsch basis (Demmler and Reinsch, 1975; Nychka and Cummins, 1996; Wood, 2006, §4.10.4), where $\mathbf{S}$ becomes $\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}\Delta^{-1}$ (the numerical approximation to the first k, Λ_i) and $\mathbf{X}$ becomes $\mathbf{QU}\Delta^{-1/2}$: but the latter is simply the numerical approximation to $\boldsymbol{\phi}_1 \ldots \boldsymbol{\phi}_k$.

Figure 2 shows the first 6 non-linear eigenfunctions of $\mathcal{K}^2$ computed by 'brute-force' discretization in grey, with the normalized cubic Demmler-Reinsch spline basis approximations shown in black for a rank 8 basis (dashed) and a rank 16 basis (dotted). Notice that the rank 8 basis approximation gives visible approximation errors for $\phi_6 \ldots \phi_8$, which have vanished for the rank 16 approximation. (Actually if we use the rank 8 thin plate regression spline basis of Wood (2003) then the approximation is accurate to graphical accuracy.)

In summary each increase in regression spline basis dimension can be viewed as refining the existing normalized Demmler-Reinsch basis functions, while adding a new one. Hence in this parameterization the notion of a sequence of estimates of a coefficient β_j is meaningful even when the basis dimension is increasing.

B.2 Consistency of penalized regression splines

This section explains why the consistency of unpenalized regression splines carries over to penalized regression splines with smoothing parameters estimated by Laplace approximate marginal likelihood. Use of Laplace approximation introduces the extra restriction $k = O(n^\alpha)$, $\alpha \leq 1/3$.

Since consistency and convergence rates of regression splines tell us nothing about what basis size to use at any

finite sample size, it is usual to use a basis dimension expected to be too large, and to impose smoothing penalties to avoid overfit. In the cubic spline basis case the coefficient estimates become

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmax}}\ \ l(\boldsymbol{\beta}) - \frac{\lambda}{2}\int g''(x)^2 dx$$

where λ is a smoothing parameter and the penalty can be written as $\lambda \int g''(x)^2 dx = \lambda \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}$, for known coefficient matrix $\mathbf{S}$. From a Bayesian viewpoint the penalty arises from an improper Gaussian prior $\boldsymbol{\beta} \sim N\{\mathbf{0}, (\lambda\mathbf{S})^{-}\}$.

Consistency of the unpenalized regression spline estimate for g and g'' implies consistency of penalized estimates when the smoothing parameter is estimated by Laplace approximate marginal likelihood (again assuming a regular likelihood and that the true g is 4 time differentiable). To see this, first set the smoothing parameter to

$$\lambda^* = \frac{k-2}{\int g''(x)^2 dx},$$

where the basis size $k = O(n^\alpha)$ for $\alpha \in (0, 1/3)$. Routine calculation shows that this is the value of λ that maximizes the prior density at the true $P(g) = \int g''(x)^2 dx$, although we do not need this fact. Because the regression spline is consistent for g'' it is also consistent for $P(g)$. So in the unpenalized case the evaluated $P(\hat{g})$ would be $O(\int g''(x)^2 dx)$, while in the penalized case it must be at most $O(\int g''(x)^2 dx)$. Hence with the given λ^* the penalty is at most $O(k)$, while the log likelihood is $O(n)$. Intuitively this suggests that the penalty is unlikely to alter the consistency of the unpenalized maximum likelihood estimates.

To see that this intuition is correct, we first reparameterize using the normalized Demmler-Reinsch basis of the previous section. Then the penalized estimate of $\boldsymbol{\beta}$ must satisfy

$$\frac{\partial l}{\partial \boldsymbol{\beta}} - \lambda^* \boldsymbol{\Lambda}\boldsymbol{\beta} = 0. \tag{2}$$

It turns out that if we linearize this equation about the unpenalized $\hat{\boldsymbol{\beta}}$, then in the large sample limit the solution of the linearized version is at the unpenalized $\hat{\boldsymbol{\beta}}$, implying that (2) must have a root at $\hat{\boldsymbol{\beta}}$ in the large sample limit. Specifically, defining $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$, and then solving the linearized version of (2) for $\Delta\boldsymbol{\beta}$ yields

$$\Delta\boldsymbol{\beta} = -\left(\mathbf{H} + \lambda^*\boldsymbol{\Lambda}\right)^{-1}\lambda^*\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}.\ \text{where}\ \ \mathbf{H} = -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathsf{T}}}.$$

Given the reparameterization the elements of $(\mathbf{H} + \lambda^*\boldsymbol{\Lambda})^{-1}$ are at most $O(n^{\delta-1})$ where $0 \le \delta \le \alpha$, while $\lambda^*\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} = O(n^\alpha)$. Hence if all the $|\hat{\beta}_i|$ are bounded below then the $\lambda^*\Lambda_{ii}\hat{\beta}_i$ are at most $O(n^\alpha)$ and the elements of $\Delta\boldsymbol{\beta}$ are at most $O(n^{2\alpha+\delta-1})$ (since each $\Delta\beta_i$ is the sum of $O(n^\alpha)$ terms each of which is the product of an $O(n^{\delta-1})$ and an $O(n^\alpha)$ term). Alternatively, $\hat{\beta}_i = O(n^{(\delta-1)/2})$, in which case $\lambda^*\Lambda_{ii} = O(n^\gamma)$ where $\alpha < \gamma \le \alpha + 1 - \delta$. If $\gamma \le 1 - \delta$ then the elements of $\Delta\boldsymbol{\beta}$ will be at most $O(n^{\alpha+(\delta-1)/2})$. Otherwise the i^{th} row and column of $(\mathbf{H} + \lambda^*\boldsymbol{\Lambda})^{-1}$ are $O(n^{-\gamma})$, but then the elements of $\Delta\boldsymbol{\beta}$ are also $O(n^{\alpha+(\delta-1)/2})$. So $\Delta\boldsymbol{\beta} \to 0$ given the assumption that $\alpha < 1/3$ (of course this is only sufficient here).

Since the true g is unknown we can not use λ^* in practice. Instead λ is chosen to maximize the Laplace approximate marginal likelihood (LAML),

$$\mathcal{V} = \log f(\mathbf{y}|\hat{\boldsymbol{\beta}}_\lambda) + \log f_\lambda(\hat{\boldsymbol{\beta}}_\lambda) - \frac{1}{2}\log|\mathcal{H}_\lambda| + \frac{k}{2}\log(2\pi) \simeq \log\int f(\mathbf{y}|\boldsymbol{\beta}) f_\lambda(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ denotes the posterior mode/ penalized MLE of $\boldsymbol{\beta}$ for a given λ, and $\mathcal{H}_\lambda$ is the Hessian of the negative log of $f(\mathbf{y}|\hat{\boldsymbol{\beta}})f_\lambda(\hat{\boldsymbol{\beta}})$. Shun and McCullagh (1995) show that in general we require $k = O(n^\alpha)$ for $\alpha \le 1/3$ for the Laplace approximation to be well founded. If $g = \alpha_0 + \alpha_1 x$ for finite real constants α_0 and α_1, then the smoothing penalty is 0 for the true g and consistency follows from the consistency in the un-penalized case, irrespective of λ.

Now suppose that g is not linear. A maximum of $\mathcal{V}$ must satisfy

$$\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}\lambda} = \left(\left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_\lambda} + \left.\frac{\partial \log f_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_\lambda}\right)\frac{\mathrm{d}\hat{\boldsymbol{\beta}}_\lambda}{\mathrm{d}\lambda} + \frac{\partial \log f_\lambda(\hat{\boldsymbol{\beta}}_\lambda)}{\partial\lambda} - \frac{1}{2}\mathrm{tr}\left(\mathcal{H}_\lambda^{-1}\mathbf{S}\right) - \frac{1}{2}\mathrm{tr}\left(\mathcal{H}_\lambda^{-1}\frac{\mathrm{d}\mathbf{H}}{\mathrm{d}\lambda}\right) = 0 \tag{3}$$

The first term in brackets is zero by definition, so the maximizer of $\mathcal{V}$ must satisfy $2\partial \log f_\lambda(\hat{\boldsymbol{\beta}}_\lambda)/\partial\lambda - \text{tr}\left(\mathcal{H}_\lambda^{-1}\mathbf{S}\right) - \text{tr}\left(\mathcal{H}_\lambda^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda\right) = 0$ implying (after some routine manipulation) that the maximiser, $\hat{\lambda}$, must satisfy $\lambda'(\hat{\lambda}) = \hat{\lambda}$, where

$$\lambda'(\lambda) = \frac{k-2}{\hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}\hat{\boldsymbol{\beta}}_\lambda + \text{tr}\left(\mathcal{H}_\lambda^{-1}\mathbf{S}\right) + \text{tr}\left(\mathcal{H}_\lambda^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda\right)}. \tag{4}$$

$\partial\mathcal{V}/\partial\lambda|_\epsilon \leq 0$ for arbitrarily small $\epsilon > 0$ would imply a LAML optimal smoothing parameter $\lambda = 0$, otherwise $\partial\mathcal{V}/\partial\lambda|_\epsilon > 0$ implying that the right hand side of (4) is positive at $\lambda = \epsilon$. Hence if $\lambda' \leq \lambda^*$ when $\lambda = \lambda^*$, then LAML must have a turning point in $(0, \lambda^*)$[1]. In fact $\text{tr}\left(\mathcal{H}_{\lambda^*}^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda^*\right) \to 0$ as $n \to \infty$ (see B.2.1), while consistency of $\hat{\boldsymbol{\beta}}_{\lambda^*}$ implies that the limiting value of $\hat{\boldsymbol{\beta}}_{\lambda^*}^\mathsf{T}\mathbf{S}\hat{\boldsymbol{\beta}}_{\lambda^*}$ is $\int g''(x)^2 dx$. Hence in the large sample limit, since $\text{tr}\left(\mathcal{H}_\lambda^{-1}\mathbf{S}\right) > 0$, we have that $\lambda' < \lambda^*$ as required (the latter is equivalent to $\partial\mathcal{V}/\partial\lambda|_{\lambda^*} < 0$ confirming that there is a *maximum* in $(0, \lambda^*)$). Notice how straightforward this is relative to what is needed for full spline smoothing where $k = O(n)$ and much more work is required.

The result is unsurprising of course. Restricted marginal likelihood is known to smooth less that Generalized Cross Validation (Wahba, 1985), but the latter is a prediction error criterion and smoothing parameters resulting in consistent estimates are likely to have lower prediction error than smoothing parameters that result in inconsistent estimation, at least asymptotically.

B.2.1 tr $\left(\mathcal{H}_\lambda^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda\right)$

In section B.2 we require that $\text{tr}\left(\mathcal{H}_{\lambda^*}^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda^*\right) \to 0$ in the large sample limit. Unfortunately there are too many summations over k elements involved in this computation for the simple order bounding calculations used in section B.2 to yield satisfactory bounds for all $\alpha \in (0, 1/3)$. This can be rectified by another change of basis, to a slightly modified normalized Demmler-Reinsch type basis in which $\mathbf{H} + \lambda\mathbf{S}$ is diagonal. Specifically let $\mathbf{H} = \mathbf{R}^\mathsf{T}\mathbf{R}$, $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T} = \mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1}$, and let the reparameterization be $\boldsymbol{\beta}^* = n^{-1/2}\mathbf{U}^\mathsf{T}\mathbf{R}\boldsymbol{\beta}$. In the remainder of this section we work with this basis.

We have

$$\frac{\mathrm{d}H_{ij}}{\mathrm{d}\lambda} = \sum_k \frac{\partial^3 l}{\partial\beta_i\partial\beta_j\partial\beta_k}\frac{\mathrm{d}\hat{\beta}_k}{\mathrm{d}\lambda}$$

where the third derivative terms are $O(n)$ (at most). By implicit differentiation we also have

$$\frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\lambda} = -(\mathbf{I}n + \lambda^*\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$$

in the new parameterization. As in section B.2, for $\hat{\beta}_i$ bounded away from zero, the fact that $\hat{\boldsymbol{\beta}}^\mathsf{T}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} = O(1)$ leads easily to the required result, but again the $\hat{\beta}_i = O(n^{-1/2})$ case makes the bounds slightly less easy to find. In that case $\Lambda_{ii} = O(n^\gamma)$, $0 < \gamma \leq 1$, while $\lambda^*\Lambda_{ii} = O(n^{\gamma+\alpha})$. Then if $\gamma + \alpha \geq 1$ the i^{th} leading diagonal element of $(\mathbf{I}n + \lambda^*\boldsymbol{\Lambda})^{-1}$ is $O(n^{-\gamma-\alpha})$, and $\partial\hat{\beta}_i/\partial\lambda = O(n^{-\alpha-1/2})$. Otherwise $\partial\hat{\beta}_i/\partial\lambda = O(n^{\gamma-3/2})$, which is less than or equal to $O(n^{-\alpha-1/2})$ if $\gamma + \alpha < 1$. In consequence $\partial H_{ij}/\partial\lambda = O(n^{1/2})$, at most. It then follows that $\text{tr}\left(\mathcal{H}_\lambda^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda\right) = -\text{tr}\left((\mathbf{I}n + \lambda^*\boldsymbol{\Lambda})^{-1}\mathrm{d}\mathbf{H}/\mathrm{d}\lambda\right) = O(n^{\alpha-1/2})$.

B.3 Convergence rates

The preceding consistency results reveal nothing about convergence rates. For a cubic spline with evenly spaced knots parameterised using a cardinal spline basis, $\mathbf{S} = O(k^3)$ (see e.g. Wood, 2006, §4.1.2), so $\lambda\mathbf{S}$ has elements of at most $O(k^4)$, while the Hessian of the log likelihood has elements $O(n/k)$. In consequence if $k = O(n^\alpha)$, $\alpha < 1/5$, $\lambda\mathbf{S}$ is completely dominated by the Hessian of the log likelihood in the large sample limit (the elements of the score vector also dominate the elements of the penalty gradient vector), so that the penalty has no effect on any model component. Hence in the limit we have an un-penalized regression spline, and the asymptotic mean square error convergence rate is $O(n^{-8\alpha})$ (bias/approximation error dominated) for $\alpha \leq 1/9$ and $O(n^{\alpha-1})$ (variance dominated) otherwise. Notice that at the $\alpha \to 1/5$ edge of this 'asymptotic regression' regime the convergence rate tends to $O(n^{-4/5})$.

[1] Consider plotting $\lambda'(\lambda)$ against λ for $0 < \lambda < \lambda^*$. The $\lambda'(\lambda)$ curve will start above the line $\lambda' = \lambda$ and finish below it.

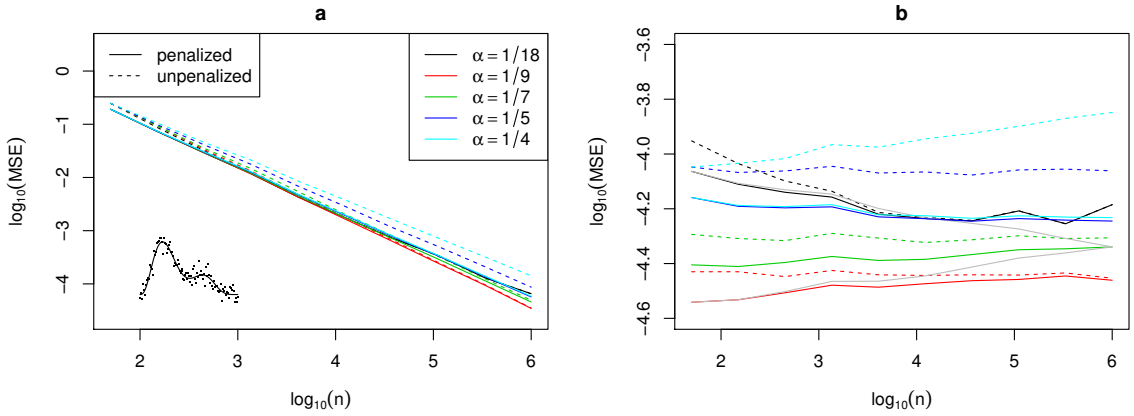

Figure 3: **a.** Example of MSE convergence for simple Gaussian smoothing. The true function is shown at lower left, with 100 noisy samples also shown. The coloured lines show log MSE averaged over 100 replicates against log sample size when the basis size $k \propto n^\alpha$ for various α values (all starting from $k = 12$ at $n = 50$). Dashed lines are for unpenalized regression and solid for penalized. For $\alpha = 1/18$ we eventually see an approximation error dominated rate. For $\alpha < 1/5$ the penalized and unpenalized curves converge, while for $\alpha \geq 1/5$ the penalty always improves the convergence rate. **b.** The same data, but de-trended by subtraction of the log MSE that would have occurred under the theoretical asymptotic convergence rate, if the observed MSE at $n = 10^6$ is correct. The theoretical rate used for $\alpha \geq 1/5$ was $n^{-4/5}$. For reference, the grey curves show curves obtained for $\alpha = 1/7$ if we incorrectly use the theoretical rates for $\alpha = 1/18, 1/9$.

For $\alpha \geq 1/5$ the total dominance of $\lambda \mathbf{S}$ by the Hessian ceases: i.e. as $n \to \infty$ the penalty can suppress overfit, in principle suppressing spurious components of the fit more rapidly than the likelihood alone would do. We do not know how to obtain actual convergence rates in this regime under LAML, although we expect them to lie between $O(n^{-4/5})$ and $O(n^{\alpha-1})$, with simulation evidence suggesting rates close to $O(n^{-4/5})$. Figure 3 shows observed convergence rates for a simple Gaussian smoothing example (a binary example gives a similar plot, but with slower convergence of the penalized case to the unpenalized case for $\alpha < 1/5$).

The best mean square error rate possible for a non-parametric estimator of a C^4 function is $O(n^{-8/9})$ (Cox, 1983), which a cubic smoothing spline can achieve under certain assumptions on the the rate of change of λ with n (Stone, 1982; Speckman, 1985). Hall and Opsomer (2005) obtain the same rate for penalized cubic regression splines as considered here. However obtaining rates under smoothing parameter selection (by REML,GCV or whatever) is more difficult. Kauermann et al. (2009) consider inference under LAML selection of smoothing parameters, but assume $k = O(n^{1/9})$ (in the cubic case). As we have seen, under LAML smoothness selection, this leads to penalized regression simply tending to unpenalized regression in the limit. Claeskens et al. (2009) recognise the existence of 2 asymptotic regimes, corresponding to penalizing in the limit and not, but do not treat the estimation of smoothing parameters.

It could be argued that in practice a statistician would tend to view a model fit with very low penalization as an indication of possible underfit, and to increase the basis dimension in response, which implies that under LAML the $\alpha \geq 1/5$ regime (penalizing in the large sample limit) is more informative in practice. The counter argument is that it is odd to choose the regime that gives the lower asymptotic convergence rates. A third point of view simply makes the modelling assumption that the truth is in the space spanned by a finite set of spline basis functions, in which case unpenalized consistency follows from standard maximum likelihood theory, and the effect of penalization with LAML smoothing parameter selection is readily demonstrated to vanish in the large sample limit. In any case the use of penalized regression splines seems to be reasonably well justified.

B.4 Large sample posterior under penalization

Consider a regular log likelihood with second and third derivatives $O(n/k)$, so that we are interested in values of the model parameters such that $|\beta_i - \hat{\beta}_i| = O(\sqrt{k/n})$. By Taylor's theorem (e.g. Gill et al., 1981, §2.3.4) we have

$$\begin{aligned} \log f(\boldsymbol{\beta}|\mathbf{y}) &\propto \log f(\mathbf{y}|\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^{\lambda}\boldsymbol{\beta} \\ &= \log f(\mathbf{y}|\hat{\boldsymbol{\beta}}) - \frac{1}{2}\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}^{\lambda}\hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^{\lambda})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + R \end{aligned} \tag{5}$$

where

$$R = \frac{1}{6}\sum_{ijk} \left.\frac{\partial^3 \log f(\mathbf{y}|\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j\partial\beta_k}\right|_{\boldsymbol{\beta}^*} (\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)(\beta_k - \hat{\beta}_k)$$

and $\boldsymbol{\beta}^* = t\boldsymbol{\beta} + (1-t)\hat{\boldsymbol{\beta}}$ for some $t \in (0,1)$. (5) can be re-written as

$$\log f(\boldsymbol{\beta}|\mathbf{y}) \propto \log f(\mathbf{y}|\hat{\boldsymbol{\beta}}) - \frac{1}{2}\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}^{\lambda}\hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^{\lambda} + \mathbf{R})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

where

$$R_{ij} = \frac{1}{3}\sum_{k} \left.\frac{\partial^3 \log f(\mathbf{y}|\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j\partial\beta_k}\right|_{\boldsymbol{\beta}^*} (\hat{\beta}_k - \beta_k).$$

In the region of interest for $\boldsymbol{\beta}$, R_{ij} are at most $O(\sqrt{kn})$, whereas the elements of $\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^{\lambda}$ are at least $O(n/k)$. Hence if $k = O(n^{\alpha})$, $\alpha < 1/3$ then $\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^{\lambda}$ dominates $\mathbf{R}$ in the $n \to \infty$ limit, and $\log f(\boldsymbol{\beta}|\mathbf{y})$ tends to the p.d.f. of $N(\hat{\boldsymbol{\beta}}, (\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^{\lambda})^{-1})$. Again this is much simpler than would be required for full spline smoothing where $k = O(n)$.

C LAML derivation and log determinants

Consider a model with log likelihood $l = \log f(\mathbf{y}|\boldsymbol{\beta})$ and improper prior $f(\boldsymbol{\beta}) = |\mathbf{S}^{\lambda}|_{+}^{1/2} \exp\{-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^{\lambda}\boldsymbol{\beta}/2\}/\sqrt{2\pi}^{p-M_p}$ where $p = \dim(\boldsymbol{\beta})$. By Taylor expansion of $\log\{f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})\}$ about $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned} \int f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})d\boldsymbol{\beta} &\simeq \int \exp\left\{l(\hat{\boldsymbol{\beta}}) - (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2 - \hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}^{\lambda}\hat{\boldsymbol{\beta}}/2 + \log|\mathbf{S}^{\lambda}|_{+}^{1/2} - \log(2\pi)(p-M_p)/2\right\} d\boldsymbol{\beta} \\ &= \exp\{\mathcal{L}(\hat{\boldsymbol{\beta}})\}|\mathbf{S}^{\lambda}|_{+}^{1/2}\sqrt{2\pi}^{M_p-p}\int \exp\{-(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2\}d\boldsymbol{\beta} \\ &= \exp\{\mathcal{L}(\hat{\boldsymbol{\beta}})\}\sqrt{2\pi}^{M_p}|\mathbf{S}^{\lambda}|_{+}^{1/2}/|\boldsymbol{\mathcal{H}}|^{1/2} \end{aligned}$$

where $\boldsymbol{\mathcal{H}}$ is the negative Hessian of the penalized log likelihood, $\mathcal{L}$.

C.1 The problem with log determinants

Unstable determinant computation is the central constraint on the development of practical fitting methods, and it is necessary to understand the issues in order to understand the structure of the numerical fitting methods. A very simple example provides adequate illustration of the key problem. Consider the real 5×5 matrix $\mathbf{C}$ with QR decomposition $\mathbf{C} = \mathbf{QR}$ so that $|\mathbf{C}| = |\mathbf{R}| = \prod_i R_{ii}$. Suppose that $\mathbf{C} = \mathbf{A} + \mathbf{B}$ where $\mathbf{A}$ is rank 2 with non-zero elements of size $O(a)$, $\mathbf{B}$ is rank 3 with non-zero elements of size $O(b)$ and $a \gg b$. Let the schematic non-zero structure of $\mathbf{C} = \mathbf{A} + \mathbf{B}$ be

$$\begin{pmatrix} \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & & \\ & & & & \\ & & & & \end{pmatrix} + \begin{pmatrix} & & & & \\ & & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \end{pmatrix}$$

where $\bullet$ shows the $O(a)$ elements and $\cdot$ those of $O(b)$. Now QR decomposition (see Golub and Van Loan, 2013) operates by applying successive householder reflections to $\mathbf{C}$, each in turn zeroing the subdiagonal elements of

successive columns of $\mathbf{C}$. Let the product of the first 2 reflections be $\mathbf{Q}_2^\mathsf{T}$ and consider the state of the QR decomposition after 2 steps. Schematically $\mathbf{Q}_2^\mathsf{T}\mathbf{C} = \mathbf{Q}_2^\mathsf{T}\mathbf{A} + \mathbf{Q}_2^\mathsf{T}\mathbf{B}$ is

$$\begin{pmatrix} \bullet & \bullet & \bullet & \cdot & \cdot \\ & \bullet & \bullet & \cdot & \cdot \\ & & d_1 & \cdot & \cdot \\ & & d_2 & \cdot & \cdot \\ & & d_3 & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & & \\ & & d_1' & & \\ & & d_2' & & \\ & & d_3' & & \end{pmatrix} + \begin{pmatrix} & & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \\ & & d_1'' & \cdot & \cdot \\ & & d_2'' & \cdot & \cdot \\ & & d_3'' & \cdot & \cdot \end{pmatrix}$$

Because $\mathbf{A}$ is rank 2, $\mathbf{d}_\mathbf{j}'$ should be 0, and d_j should be d_j'' but computationally $d_j' = O(\epsilon a)$ where ϵ is the machine precision. Hence if b approaches $O(\epsilon a)$, we suffer catastrophic loss of precision in $\mathbf{d}$, which will be inherited by R_{33} and the computed value of $|\mathbf{C}|$. Matrices such as $\sum_j \lambda_j^\mathsf{T}\mathbf{S}^j$ can suffer from exactly this problem, since some λ_j can legitimately tend to infinity while others remain finite, and the $\mathbf{S}^j$ are usually of lower rank than the dimension of their non-zero sub-block: hence both log determinant terms in the LAML score are potentially unstable.

One solution is based on similarity transform. In the case of our simple example, consider the similarity transform $\mathbf{UCU}^\mathsf{T} = \mathbf{UAU}^\mathsf{T} + \mathbf{UBU}^\mathsf{T}$ constructed to produce the following schematic

$$\begin{pmatrix} \bullet & \bullet & \cdot & \cdot & \cdot \\ \bullet & \bullet & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & & & \\ & & & & \\ & & & & \\ & & & & \end{pmatrix} + \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

$\mathbf{UCU}^\mathsf{T}$ can then be computed by adding $\mathbf{UBU}^\mathsf{T}$ to $\mathbf{UAU}^\mathsf{T}$ with the theoretically zero elements set to exact zeroes. $|\mathbf{UCU}^\mathsf{T}| = |\mathbf{C}|$, but computation based on the similarity transformed version no longer suffers from the precision loss problem, no-matter how disparate a and b are in magnitude. Wood (2011) discusses the issues in more detail and provides a practical generalized version of the similarity transform approach, allowing for multiple rank deficient components where the dominant blocks may be anywhere on the diagonal.

D Smoothing parameter uncertainty

$\partial\mathbf{R}/\partial\rho_k$: Computation of the $\mathbf{V}''$ term requires $\partial\mathbf{R}'/\partial\rho$ where $\mathbf{R}'^\mathsf{T}\mathbf{R}' = \mathbf{V}_\beta$. Generally we have access to $\partial\mathbf{A}/\partial\rho$ where $\mathbf{A} = \mathbf{V}_\beta^{-1}$. Given Cholesky factorization $\mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{A}$ then $\mathbf{R}' = \mathbf{R}^{-\mathsf{T}}$, and $\partial\mathbf{R}'^\mathsf{T}/\partial\rho = -\mathbf{R}^{-1}\partial\mathbf{R}/\partial\rho\mathbf{R}^{-1}$. Applying the chain rule to the Cholesky factorization yields

$$\frac{\partial R_{ii}}{\partial\rho} = \frac{1}{2}R_{ii}^{-1}B_{ii}, \quad \frac{\partial R_{ij}}{\partial\rho} = R_{ii}^{-1}\left(B_{ij} - R_{ij}\frac{\partial R_{ii}}{\partial\rho}\right), B_{ij} = \frac{\partial A_{ij}}{\partial\rho} - \sum_{k=1}^{i-1}\frac{\partial R_{ki}}{\partial\rho}R_{kj} + R_{ki}\frac{\partial R_{kj}}{\partial\rho},$$

and $\sum_{k=1}^{0} x_i$ is taken to be 0. The equations are used starting from the top left of the matrices and working across the columns of each row before moving on to the next row, at approximately double the floating point cost of the original Cholesky factorization, but with no square roots.

Ratio of the first order correction terms

In the notation of section 4, we now show that for any smooth g_j, $\partial\hat{\boldsymbol{\beta}}/\partial\rho_j$ tends to dominate $\partial\mathbf{R}_\rho^\mathsf{T}\boldsymbol{z}/\partial\rho_j$ for those components of the smooth that are detectably non zero. First rewrite $\mathbf{S}^\rho = \mathbf{S}_{-j} + \lambda_j\mathbf{S}_j$, by definition of $\mathbf{S}_{-j}$, and then form the spectral decomposition $\boldsymbol{\mathcal{I}} + \mathbf{S}_{-j} = \mathbf{VDV}^\mathsf{T}$. Form a second spectral decomposition $\mathbf{D}^{-1/2}\mathbf{V}^\mathsf{T}\mathbf{S}_j\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}$, so that $\boldsymbol{\mathcal{I}}+\mathbf{S}^\rho = \mathbf{VD}^{1/2}\mathbf{U}(\mathbf{I}+\lambda_j\boldsymbol{\Lambda})\mathbf{U}^\mathsf{T}\mathbf{D}^{1/2}\mathbf{V}^\mathsf{T}$. Now linearly re-parameterize so that $\mathbf{S}_j$ becomes $\boldsymbol{\Lambda}$ and $\boldsymbol{\mathcal{I}} + \mathbf{S}^\rho$ becomes $\mathbf{I} + \lambda_j\boldsymbol{\Lambda}$, while $\mathbf{R}^\mathsf{T} = (\mathbf{I} + \lambda_j\boldsymbol{\Lambda})^{-1/2}$. By the implicit function theorem, in the new parameterization $\mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\rho_j = \lambda_j(\mathbf{I} + \lambda_j\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$. Notice that $\boldsymbol{\Lambda}$ has only rank$(\mathbf{S}_j)$ non-zero entries, corresponding to the parameters in the new parameterization representing the penalized component of g_j. Furthermore $\mathrm{d}\mathbf{R}^\mathsf{T}\mathbf{z}/\mathrm{d}\rho_j \simeq \lambda_j(\mathbf{I} + \lambda_j\boldsymbol{\Lambda})^{-3/2}\boldsymbol{\Lambda}\mathbf{z}$, where we have neglected the indirect dependence on smoothing parameters via the curvature of $\boldsymbol{\mathcal{I}}$ changing as $\boldsymbol{\beta}$ changes with $\boldsymbol{\rho}$. Hence for any penalized parameter β_i of g_j

$$\frac{\mathrm{d}\hat{\beta}_i/\mathrm{d}\rho_j}{\mathrm{d}(\mathbf{R}^\mathsf{T}\mathbf{z})_i/\mathrm{d}\rho_j} \simeq \frac{\hat{\beta}_i}{(1 + \lambda_j\Lambda_{ii})^{-1/2}z_i},$$

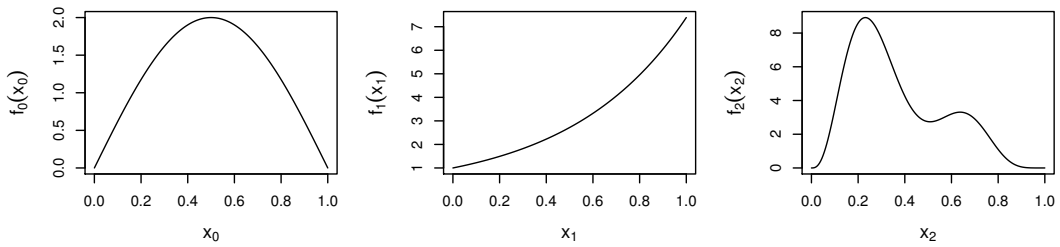

Figure 4: Shapes of the functions used for the simulation study (from Gu and Wahba, 1991). $f_3(x_3) = 0$.

but $(1+\lambda_j\Lambda_{ii})^{-1/2}$ is the (posterior) standard deviation of β_i. So the more clearly non-zero is β_i, the more $\mathrm{d}\hat{\beta}_i/\mathrm{d}\rho_j$ dominates $\mathrm{d}(\mathbf{R}^\mathsf{T}\mathbf{z})_i/\mathrm{d}\rho_j$. The dominance increases with sample size (provided that the data are informative), for all components except those heavily penalized towards zero.

Proof of lemma 1 Form eigen-decompositions $\hat{\boldsymbol{\mathcal{I}}} = \mathbf{V}\mathbf{D}\mathbf{V}^\mathsf{T}$ and $\mathbf{D}^{-1/2}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}$, and linearly re-parameterize $\boldsymbol{\beta}' = \mathbf{U}^\mathsf{T}\mathbf{D}^{-1/2}\mathbf{V}^\mathsf{T}\boldsymbol{\beta}$, so that in the new parameterization $\hat{\boldsymbol{\mathcal{I}}}$ becomes an identity matrix, while the prior becomes $\boldsymbol{\beta}' \sim N(\mathbf{0}, \boldsymbol{\Lambda}^-)$, $\mathbf{V}_{\hat{\beta}'} = (\mathbf{I}+\boldsymbol{\Lambda})^{-2}$ and $\mathbf{V}_{\beta'} = (\mathbf{I}+\boldsymbol{\Lambda})^{-1}$.

$$\begin{aligned}
\mathbf{V}_{\hat{\beta}'} + \mathbb{E}_\pi(\tilde{\boldsymbol{\Delta}}_{\beta'}\tilde{\boldsymbol{\Delta}}_{\beta'}^\mathsf{T}) &= (\mathbf{I}+\boldsymbol{\Lambda})^{-2} + \mathbb{E}_\pi[\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\}\boldsymbol{\beta}'\boldsymbol{\beta}'^\mathsf{T}\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\}] \\
&= (\mathbf{I}+\boldsymbol{\Lambda})^{-2} + \{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\}\boldsymbol{\Lambda}^-\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\} \\
&= (\mathbf{I}+\boldsymbol{\Lambda})^{-1}[(\mathbf{I}+\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Lambda}^-\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\} - (\mathbf{I}+\boldsymbol{\Lambda})\boldsymbol{\Lambda}^-\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\}] \\
&= (\mathbf{I}+\boldsymbol{\Lambda})^{-1}[(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^-\{(\mathbf{I}+\boldsymbol{\Lambda})^{-1} - \mathbf{I}\}] = (\mathbf{I}+\boldsymbol{\Lambda})^{-1}\mathbf{I} = \mathbf{V}_{\beta'}.
\end{aligned}$$

E Further simulation details

Figure 4 shows the functions used in the simulation study in the main paper. In the uncorrelated covariate case x_{0i}, x_{1i}, x_{2i} and x_{3i} were all i.i.d. $U(0,1)$. Correlated covariates were marginally uniform, but were generated as $x_{ji} = \Phi^{-1}(z_{ji})$ where Φ is the standard normal c.d.f. and $(z_{0i}, z_{1i}, z_{2i}, z_{3i}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ having unit diagonal and 0.9 for all other elements. The noise level was set by either using the appropriate values of the distribution parameters or by multiplying the linear predictor by the appropriate scale factor as indicated in the second column of table 1 (the scale factor is denoted by d). The simulation settings and failure rates are given in table 1

F Some examples

This section presents some example applications which are all routine given the framework developed here, but would not have been without it. See appendix M for a brief description of the software used for this.

F.1 Kidney renal clear cell carcinoma: Cox survival modelling with smooth interactions

The left 2 panels of figure 5 show survival times of patients with kidney renal clear cell carcinoma, plotted against disease stage at diagnosis and age, with survival time data in red and censoring time data in black (available from `https://tcga-data.nci.nih.gov/tcga/`). Other recorded variables include race, previous history of malignancy and laterality (whether the left or right kidney is affected). A possible model for the survival times would be a Cox proportional hazards model with linear predictor dependent on parametric effects of the factor predictors and smooth effects of age and stage. Given the new methods this model can readily be estimated, as detailed in appendix G. A model with smooth main effects plus an interaction has a marginally lower AIC than the main effects only and the combined effect of age and stage is shown in the right panel of figure 5. Broadly it appears that both age and stage increase the hazard, except at relatively high stage where age matters little below ages in the mid sixties. Disease in the right kidney leads to significantly reduced hazard (p=.005) relative to disease in the

Simulation setting			Alternative	MSE/Biers diff.
Family	parameters	approx. r^2	% failure	p-value
nb	$\theta = 3, d = .12$	0.25	-(.3)	0.0015(0.0013)
	$\theta = 3, d = .2$	0.45	-(.7)	0.087($< 10^{-5}$)
	$\theta = 3, d = .4$	0.79	-(.3)	$< 10^{-5}$($< 10^{-5}$)
beta	$\theta = 0.02$	0.3	-(1.3)	0.40($< 10^{-5}$)
	$\theta = 0.01$	0.45	-(1.3)	0.16($< 10^{-5}$)
	$\theta = 0.001$	0.9	-(.7)	$< 10^{-5}$(0.044)
scat	$\nu = 5, \sigma = 2.5$	0.5	-(2)	.021($< 10^{-5}$)
	$\nu = 3, \sigma = 1.3$	0.7	.3(-)	$< 10^{-5}$($< 10^{-5}$)
	$\nu = 4, \sigma = 0.9$	0.85	-(.3)	$< 10^{-5}$(0.41)
zip	$\boldsymbol{\theta} = (-2, 0), d = 2$	0.5	2(3.3)	$< 10^{-5}$(0.004)
	$\boldsymbol{\theta} = (-2, 0), d = 2.5$	0.67	4.3(3.7)	$< 10^{-5}$(0.001)
	$\boldsymbol{\theta} = (-2, 0), d = 3$	0.8	8.3(4.3)	$< 10^{-5}$($< 10^{-5}$)
ocat	$\boldsymbol{\theta} = (-1, 0, 3), d = .3$	0.4	-	0.388($< 10^{-5}$)
	$\boldsymbol{\theta} = (-1, 0, 3), d = 1$	0.7	-	0.0025(0.0023)
	$\boldsymbol{\theta} = (-1, 0, 3), d = 2$	0.85	-	0.191(7.3×10^{-4})

Table 1: Simulation settings, failure rates and p-values for performance differences when comparing the new methods to existing software. The approximate r^2 column gives the approximate proportion of the variance explained by the linear predictor, for each scenario. The fit failure rates for the alternative procedure are also given (for the correlated covariate case in brackets): the new method produced no failures. The p-values for the difference between MSE or Briers scores between the methods are also reported. The new method had the better average scores in all cases that were significant at the 5% level, except for the zip model on uncorrelated data, where the GAMLSS methods achieved slightly lower MSE.

left kidney: the reduction on the linear predictor scale being 0.45. This effect is likely to relate to the asymmetry in arrangement of other internal organs. There was no evidence of an effect of race or previous history of malignancy.

F.2 Overdispersed Horse Mackerel eggs

Figure 6 shows data from a 2010 survey of Horse Mackerel eggs. The data are from the WGMEGS working group (`http://www.ices.dk/marine-data/data-portals/Pages/Eggs-and-larvae.aspx`). Egg surveys are commonly undertaken to help in fish stock assessment and are attractive because unbiased sampling of eggs is much easier than unbiased sampling of adult fish. The eggs are collected by ship based sampling and typically show over-dispersion relative to Poisson and a high proportion of zeroes. The high proportion of zeroes is often used to justify the use of zero inflated models, although reasoning based on the marginal distribution of eggs is clearly incorrect, and the zeroes are often highly clustered in space, suggesting a process with a spatial varying mean, rather than zero inflation.

The new methods make it straightforward to rapidly compare several possible models for the data, in particular Poisson, zero-inflated Poisson, Tweedie and negative binomial distributions. A common structure for the expected number of eggs, μ_i, (or Poisson parameter in the zero inflated case) was :

$$\log(\mu_i) = \log(\texttt{vol}_i) + b_{s(i)} + f_1(\texttt{lo}_i, \texttt{la}_i) + f_2(\texttt{T.20}_i) + f_3(\texttt{T.surf}_i) + f_4(\texttt{sal.20}_i)$$

where $\texttt{vol}_i$ is the volume of water sampled, $b_{s(i)}$ is an independent Gaussian random effect for the ship that obtained sample i, $\texttt{lo}_i$ and $\texttt{la}_i$ are longitude and latitude (actually converted onto a square grid for modelling), $\texttt{T.20}_i$ and $\texttt{T.surf}_i$ are water temperature at 20m depth and the surface, respectively and $\texttt{sal.20}_i$ is salinity at 20m depth. Univariate smooth effects were modelled using rank 10 thin plate regression splines, while the spatial effect was modelled using a rank 50 Duchon spline, with a first order derivative penalty and $s = 1/2$ (Duchon, 1977; Miller and Wood, 2014).

An initial Poisson fit of this model structure was very poor with clear over-dispersion. We therefore tried negative binomial, Tweedie and two varieties of zero inflated Poisson models. The details of the zero inflated

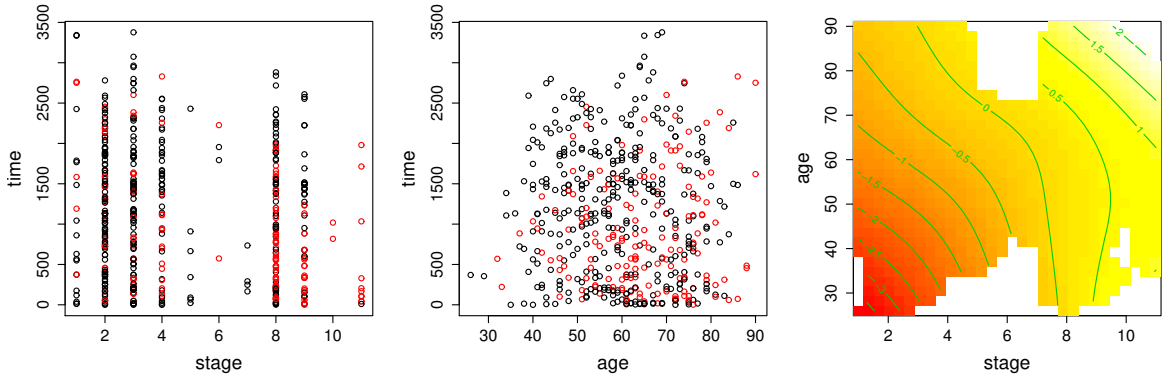

Figure 5: Left: Survival times (red) and censoring times (black) against disease stage for patients with kidney renal clear cell carcinoma. Middle: times against patient age. Right: the combined smooth effect of age and stage on the linear predictor scale from a Cox Proportional hazards survival model estimated by maximum penalized partial likelihood. Higher values indicate higher hazard resulting in shorter survival times.

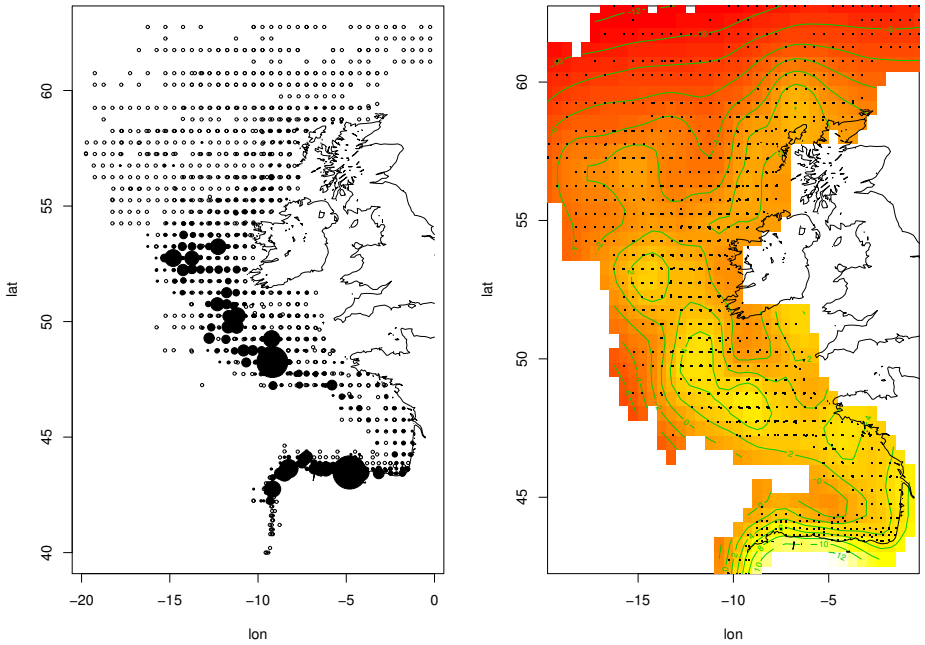

Figure 6: Left: 2010 Horse Mackerel egg survey data. Open circles are survey stations with no eggs, while solid symbols have area proportional to number of eggs sampled. Right: Fitted spatial effect from the best fit negative binomial based model.

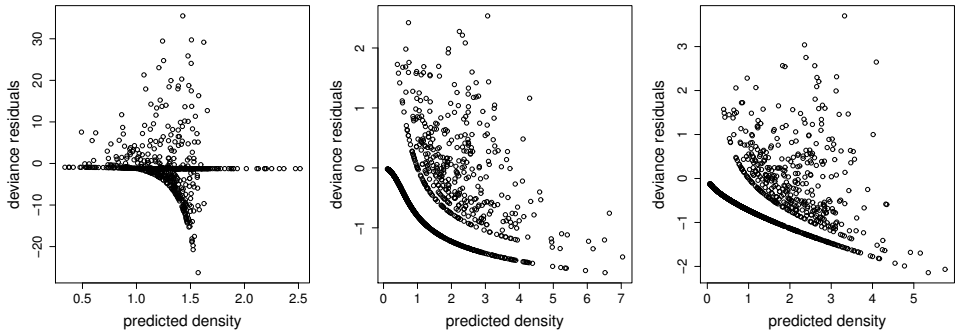

Figure 7: Residual plots for three Horse Mackerel egg models. Deviance residuals have been scaled by the scale parameter so that they should have unit variance for a good model. The fourth root of fitted values is used to best show the structure of the residuals. Left is for a zero inflated Poisson model: the zero inflation has served to reduce the variability in the fitted values, allowed substantial over prediction of a number of zeroes, and has not dealt with over-dispersion. Middle and right are for the equivalent negative binomial and Tweedie models. The right two are broadly acceptable, although there is some over-prediction of very low counts evident at the right of both plots.

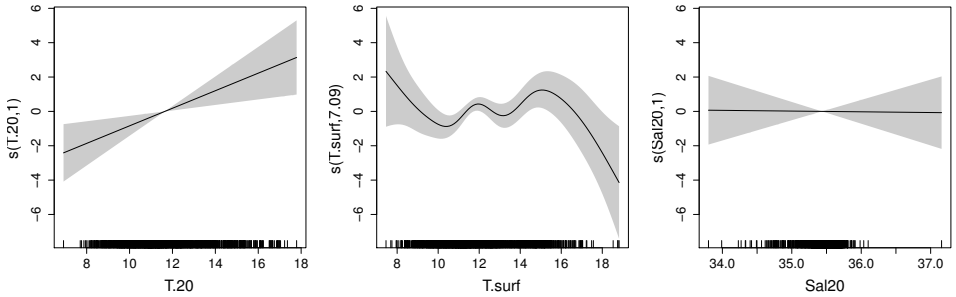

Figure 8: Horse Mackerel model univariate effect estimates.

Poisson models are given in Appendix I. The extended GAM version has the zero inflation rate depending on a logistic function of the linear predictor controlling the Poisson mean, with the restriction that zero inflation must be non-increasing with the Poisson mean. The more general GAMLSS formulation (section 3.2 and appendix I) has a linear predictor for the probability, p_i, of potential presence of eggs

$$\text{logit}(p_i) = f_1(\texttt{lo}_i, \texttt{la}_i) + f_2(\texttt{T.20}_i)$$

with the same model as above for the Poisson mean given potential presence.

Figure 7 shows simple plots of scaled deviance residuals against 4th root of fitted values. The plot for the extended GAM version of the zero inflated model is shown in the left panel and makes it clear that zero inflation is not the answer to the over-dispersion problem in the Poisson model; the GAMLSS zero inflated plot is no better. The negative binomial and Tweedie plots are substantially better, so that formal model selection makes sense in this case. The AIC of section 5 selects the negative binomial model with an AIC of 4482 against 4979 for the Tweedie (the Poisson based models have much higher AIC, of course).

Further model checking then suggested increasing the basis dimension of the spatial smooth and changing from a Duchon spline to a thin plate spline, so that the final model spatial effect estimate, plotted on the right hand side of figure 6, uses a thin plate regression spline with basis dimension 150 (although visually the broad structure of the effect estimates is very similar to the original fit). The remaining effect estimates are plotted in figure 8. Clearly there is no evidence for an effect of salinity, while the association of density with water temperature is real, but given the complexity of the surface affect, possibly acting as a surrogate for the causal variables here. The final fitted model explains around 70% of the deviance in egg count.

G Cox proportional hazards model

The Cox proportional Hazards model (Cox, 1972) is an important example of a general smooth model requiring the methods of section 3.1, at least if the computational cost is to be kept linear in the sample size, rather than quadratic. With some care in the structuring of the computations, the computational cost can be kept to $O(Mnp^2)$. Let the n data be of the form $(\tilde{t}_i, \mathbf{X}_i, \delta_i)$, i.e. an event time, model matrix row (there is no intercept in the Cox model) and an indicator of death (1) or censoring (0). Assume w.l.o.g. that the data are ordered so that the t_i are non-increasing with i. The time data can conveniently be replaced by a vector $\mathbf{t}$ of n_t unique decreasing event times, and an n vector of indices, r, such that $t_{r_i} = \tilde{t}_i$.

The log likelihood, as in Hastie and Tibshirani (1990), is

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{n_t} \left[\sum_{\{i:r_i=j\}} \delta_i \mathbf{X}_i \boldsymbol{\beta} - d_j \log \left\{ \sum_{\{i:r_i \leq j\}} \exp(\mathbf{X}_i \boldsymbol{\beta}) \right\} \right].$$

Now let $\eta_i \equiv \mathbf{X}_i \boldsymbol{\beta}$, $\gamma_i \equiv \exp(\eta_i)$ and $d_j = \sum_{\{i:r_i=j\}} \delta_i$ (i.e the count of deaths at this event time). Then

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{n_t} \left[\sum_{\{i:r_i=j\}} \delta_i \eta_i - d_j \log \left\{ \sum_{\{i:r_i \leq j\}} \gamma_i \right\} \right].$$

Further define $\gamma_j^+ = \sum_{\{i:r_i \leq j\}} \gamma_i$, so that we have the recursion

$$\gamma_j^+ = \gamma_{j-1}^+ + \sum_{\{i:r_i=j\}} \gamma_i$$

where $\gamma_0^+ = 0$. Then

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \eta_i - \sum_{j=1}^{n_t} d_j \log(\gamma_j^+).$$

Turning to the gradient $g_k = \partial l / \partial \beta_k$, we have

$$\mathbf{g} = \sum_{i=1}^{n} \delta_i \mathbf{X}_i - \sum_{j=1}^{n_t} d_j \mathbf{b}_j^+ / \gamma_j^+$$

where $\mathbf{b}_j^+ = \mathbf{b}_{j-1}^+ + \sum_{\{i:r_i=j\}} \mathbf{b}_i$, $\mathbf{b}_i = \gamma_i \mathbf{X}_i$. and $\mathbf{b}_0^+ = \mathbf{0}$. Finally the Hessian $H_{km} = \partial^2 l / \partial \beta_k \partial \beta_m$ is given by

$$\mathbf{H} = \sum_{j=1}^{n_t} d_j \mathbf{b}_j^+ \mathbf{b}_j^{+\mathsf{T}} / \gamma_j^{+2} - d_j \mathbf{A}_j^+ / \gamma_j^+$$

where $\mathbf{A}_j^+ = \mathbf{A}_{j-1}^+ + \sum_{\{i:r_i=j\}} \mathbf{A}_i$, $\mathbf{A}_i = \gamma_i \mathbf{X}_i \mathbf{X}_i^{\mathsf{T}}$ and $\mathbf{A}_0^+ = \mathbf{0}$.

Derivatives with respect to smoothing parameters

To obtain derivatives it will be necessary to obtain expressions for the derivatives of l and $\mathbf{H}$ with respect to $\rho_k = \log(\lambda_k)$. Firstly we have

$$\frac{\partial \eta_i}{\partial \rho_k} = \mathbf{X}_i \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_k}, \quad \frac{\partial \gamma_i}{\partial \rho_k} = \gamma_i \frac{\partial \eta_i}{\partial \rho_k}, \quad \frac{\partial \mathbf{b}_i}{\partial \rho_k} = \frac{\partial \gamma_i}{\partial \rho_k} \mathbf{X}_i \quad \text{and} \quad \frac{\partial \mathbf{A}_i}{\partial \rho_k} = \frac{\partial \gamma_i}{\partial \rho_k} \mathbf{X}_i \mathbf{X}_i^{\mathsf{T}}.$$

Similarly

$$\frac{\partial^2 \eta_i}{\partial \rho_k \partial \rho_m} = \mathbf{X}_i \frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_k \partial \rho_m}, \quad \frac{\partial^2 \gamma_i}{\partial \rho_k \partial \rho_m} = \gamma_i \frac{\partial \eta_i}{\partial \rho_k} \frac{\partial \eta_i}{\partial \rho_m} + \gamma_i \frac{\partial^2 \eta_i}{\partial \rho_k \partial \rho_m}, \quad \frac{\partial^2 \mathbf{b}_i}{\partial \rho_k \partial \rho_m} = \frac{\partial^2 \gamma_i}{\partial \rho_k \partial \rho_m} \mathbf{X}_i.$$

Derivatives sum in the same way as the terms they relate to.

$$\frac{\partial l}{\partial \rho_k} = \sum_{i=1}^{n} \delta_i \frac{\partial \eta_i}{\partial \rho_k} - \sum_{j=1}^{n_t} \frac{d_j}{\gamma_j^+} \frac{\partial \gamma_j^+}{\partial \rho_k},$$

and

$$\frac{\partial^2 l}{\partial \rho_k \partial \rho_m} = \sum_{i=1}^{n} \delta_i \frac{\partial^2 \eta_i}{\partial \rho_k \partial \rho_m} + \sum_{j=1}^{n_t} \left(\frac{d_j}{\gamma_j^{+2}} \frac{\partial \gamma_j^+}{\partial \rho_m} \frac{\partial \gamma_j^+}{\partial \rho_k} - \frac{d_j}{\gamma_j^+} \frac{\partial^2 \gamma_j^+}{\partial \rho_k \partial \rho_m} \right),$$

while

$$\frac{\partial \mathbf{H}}{\partial \rho_k} = \sum_{j=1}^{n_t} \frac{d_j}{\gamma_j^{+2}} \left\{ \mathbf{A}_j^+ \frac{\partial \gamma_j^+}{\partial \rho_k} + \frac{\partial \mathbf{b}^+}{\partial \rho_k} \mathbf{b}^{+\mathsf{T}} + \mathbf{b}^+ \frac{\partial \mathbf{b}^{+\mathsf{T}}}{\partial \rho_k} \right\} - \frac{d_j}{\gamma_j^+} \frac{\partial \mathbf{A}_j^+}{\partial \rho_k} - \frac{2d_j}{\gamma_j^{+3}} \mathbf{b}^+ \mathbf{b}^{+\mathsf{T}} \frac{\partial \gamma_j^+}{\partial \rho_k}$$

and

$$\begin{aligned}
\frac{\partial^2 \mathbf{H}}{\partial \rho_k \partial \rho_m} = \sum_{j=1}^{n_t} & \frac{-2d_j}{\gamma_j^{+3}} \frac{\partial \gamma_j^+}{\partial \rho_m} \left\{ \mathbf{A}_j^+ \frac{\partial \gamma_j^+}{\partial \rho_k} + \frac{\partial \mathbf{b}^+}{\partial \rho_k} \mathbf{b}^{+\mathsf{T}} + \mathbf{b}^+ \frac{\partial \mathbf{b}^{+\mathsf{T}}}{\partial \rho_k} \right\} + \frac{d_j}{\gamma_j^{+2}} \left\{ \frac{\partial \mathbf{A}_j^+}{\partial \rho_m} \frac{\partial \gamma_j^+}{\partial \rho_k} \right. \\
& \left. + \mathbf{A}_j^+ \frac{\partial^2 \gamma_j^+}{\partial \rho_k \partial \rho_m} + \frac{\partial^2 \mathbf{b}^+}{\partial \rho_k \partial \rho_m} \mathbf{b}^{+\mathsf{T}} + \frac{\partial \mathbf{b}^+}{\partial \rho_k} \frac{\partial \mathbf{b}^{+\mathsf{T}}}{\partial \rho_m} + \frac{\partial \mathbf{b}^+}{\partial \rho_m} \frac{\partial \mathbf{b}^{+\mathsf{T}}}{\partial \rho_k} + \mathbf{b}^+ \frac{\partial^2 \mathbf{b}^{+\mathsf{T}}}{\partial \rho_k \partial \rho_m} \right\} \\
& + \frac{d_j}{\gamma_j^{+2}} \frac{\partial \gamma_j^+}{\partial \rho_m} \frac{\partial \mathbf{A}_j^+}{\partial \rho_k} - \frac{d_j}{\gamma_j^+} \frac{\partial^2 \mathbf{A}_j^+}{\partial \rho_k \partial \rho_m} + \frac{6d_j}{\gamma_j^{+4}} \frac{\partial \gamma_j^+}{\partial \rho_m} \mathbf{b}^+ \mathbf{b}^{+\mathsf{T}} \frac{\partial \gamma_j^+}{\partial \rho_k} \\
& - \frac{2d_j}{\gamma_j^{+3}} \left\{ \frac{\partial \mathbf{b}^+}{\partial \rho_m} \mathbf{b}^{+\mathsf{T}} \frac{\partial \gamma_j^+}{\partial \rho_k} + \mathbf{b}^+ \frac{\partial \mathbf{b}^{+\mathsf{T}}}{\partial \rho_m} \frac{\partial \gamma_j^+}{\partial \rho_k} + \mathbf{b}^+ \mathbf{b}^{+\mathsf{T}} \frac{\partial^2 \gamma_j^+}{\partial \rho_k \partial \rho_m} \right\}.
\end{aligned}$$

In fact with suitable reparameterization it will only be necessary to obtain the second derivatives of the leading diagonal of $\mathbf{H}$, although the full first derivative of $\mathbf{H}$ matrices will be needed. All that is actually needed is $\mathrm{tr}\left(\boldsymbol{\mathcal{H}}^{-1} \partial^2 \mathbf{H} / \partial \rho_k \partial \rho_m\right)$. Consider the eigen-decomposition $\boldsymbol{\mathcal{H}}^{-1} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^{\mathsf{T}}$. We have

$$\mathrm{tr}\left(\boldsymbol{\mathcal{H}}^{-1} \frac{\partial \mathbf{H}}{\partial \theta} \right) = \mathrm{tr}\left(\boldsymbol{\Lambda} \frac{\partial \mathbf{V}^{\mathsf{T}} \mathbf{H} \mathbf{V}}{\partial \theta} \right), \quad \mathrm{tr}\left(\boldsymbol{\mathcal{H}}^{-1} \frac{\partial^2 \mathbf{H}}{\partial \theta_k \partial \theta_m} \right) = \mathrm{tr}\left(\boldsymbol{\Lambda} \frac{\partial^2 \mathbf{V}^{\mathsf{T}} \mathbf{H} \mathbf{V}}{\partial \theta_k \partial \theta_m} \right).$$

Since $\boldsymbol{\Lambda}$ is diagonal only the leading diagonal of the derivative of the reparameterized Hessian $\mathbf{V}^{\mathsf{T}} \mathbf{H} \mathbf{V}$ is required, and this can be efficiently computed by simply using the reparameterized model matrix $\mathbf{XV}$. So the total cost of all derivatives is kept to $O(Mnp^2)$.

Prediction and the baseline hazard

Klein and Moeschberger (2003, pages 283, 359, 381) gives the details. Here we simply restate the required expressions in forms suitable for efficient computation, using the notation and assumptions of the previous sections.

1. The estimated cumulative baseline hazard is

$$H_0(t) = \begin{cases} h_j & t_j \le t < t_{j-1} \\ 0 & t < t_{n_t} \\ h_1 & t \ge t_1 \end{cases}$$

where the following back recursion defines h_j

$$h_j = h_{j+1} + \frac{d_j}{\gamma_j^+}, \quad h_{n_t} = \frac{d_{n_t}}{\gamma_{n_t}^+}.$$

2. The variance of the estimated cumulative hazard is given by the back recursion

$$q_j = q_{j+1} + \frac{d_j}{\gamma_j^{+2}}, \quad q_{n_t} = \frac{d_{n_t}}{\gamma_{n_t}^{+2}}.$$

3. The estimated survival function for time t, covariate vector $\mathbf{x}$, is

$$\hat{S}(t, \mathbf{x}) = \exp\{-H_0(t)\}^{\exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta})}$$

and consequently $\log \hat{S}(t, \mathbf{x}) = -H_0(t) \exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta})$. Let $\hat{S}_i$ denote the estimated version for the i^{th} subject, at their event time.

4. The estimated variance of $\hat{S}(t, \mathbf{x})$ is[2]

$$\exp(\mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}})\hat{S}(t, \mathbf{x})(q_i + \mathbf{v}_i^\mathsf{T}\mathbf{V}_\beta \mathbf{v}_i)^{1/2}, \quad \text{if} \quad t_i \leq t < t_{i-1}$$

where $\mathbf{v}_i = \mathbf{a}_i - \mathbf{x}h_i$, and the vector $\mathbf{a}_i$ is defined by the back recursion

$$\mathbf{a}_i = \mathbf{a}_{i+1} + \mathbf{b}_i^+ \frac{d_i}{\gamma_i^{+2}}, \quad \mathbf{a}_{n_t} = \mathbf{b}_{n_t}^+ \frac{d_{n_t}}{\gamma_{n_t}^{+2}}.$$

For efficient prediction with standard errors, there seems to be no choice but to compute the n_t, $\boldsymbol{a}_i$ vectors at the end of fitting and store them.

5. Martingale residuals are defined as

$$\hat{M}_j = \delta_j + \log \hat{S}_j,$$

and deviance residuals as

$$\hat{D}_j = \text{sign}(\hat{M}_j)[-2\{\hat{M}_j + \delta_j \log(-\log \hat{S}_j)\}]^{1/2}.$$

The latter also being useful for computing a deviance.

H Multivariate additive model example

Consider a model in which independent observations $\mathbf{y}$ are m dimensional multivariate Gaussian with precision matrix $\boldsymbol{\Sigma}^{-1} = \mathbf{R}^\mathsf{T}\mathbf{R}$, $\mathbf{R}$ being a Cholesky factor of the form

$$\mathbf{R} = \begin{pmatrix} e^{\theta_1} & \theta_2 & \cdot & \cdot \\ 0 & e^{\theta_{m+1}} & \theta_{m+2} & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Let $\mathbb{D}$ denote the set of θ_i's giving the diagonal elements of $\mathbf{R}$, with corresponding indicator function $\mathbb{I}_\mathbb{D}(i)$ taking value 1 if θ_i is in $\mathbb{D}$ and 0 otherwise. The mean vector $\boldsymbol{\mu}$ has elements $\mu_i = \mathbf{x}^i\boldsymbol{\beta}^i$, where $\mathbf{x}^i$ is a model matrix row for the i^{th} component with corresponding coefficient vector $\boldsymbol{\beta}^i$. In what follows it will help to define $\bar{\mathbf{x}}_i^l$ as an m vector of zeroes except for element l which is x_i^l .

Consider the log likelihood for a single $\mathbf{y}$

$$l = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{R}(\mathbf{y} - \boldsymbol{\mu}) + \sum_{\theta_i \in \mathbb{D}} \theta_i,$$

where $\sum_{\theta_i \in \mathbb{D}} \theta_i = \log |\mathbf{R}|$. For Newton estimation of the model coefficients we need gradients

$$l_\theta^i = -(\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{R}_\theta^i(\mathbf{y} - \boldsymbol{\mu}) + \mathbb{I}_\mathbb{D}(i)$$

[2] Klein and Moeschberger (2003) miss a term in their expression (8.8.5). The correct form used here can be found in Andersen et al. (1996) expression (10), for example.

and

$$l^i_{\beta^l} = \bar{\mathbf{x}}_i^{l\mathsf{T}}\mathbf{R}^\mathsf{T}\mathbf{R}(\mathbf{y}-\boldsymbol{\mu}).$$

Then we need Hessian blocks

$$l^{i\ j}_{\beta^l\beta^k} = -\bar{\mathbf{x}}_i^{l\mathsf{T}}\mathbf{R}^\mathsf{T}\mathbf{R}\bar{\mathbf{x}}_j^k,$$

$$l^{i\ j}_{\beta^l\theta} = \bar{\mathbf{x}}_i^{l\mathsf{T}}(\mathbf{R}_\theta^{j\mathsf{T}}\mathbf{R} + \mathbf{R}^\mathsf{T}\mathbf{R}_\theta^j)(\mathbf{y}-\boldsymbol{\mu}),$$

$$l^{ij}_{\theta\theta} = -(\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\mathbf{R}_\theta^{j\mathsf{T}}\mathbf{R}_\theta^i(\mathbf{y}-\boldsymbol{\mu}) - (\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{R}^{ij}_{\theta\theta}(\mathbf{y}-\boldsymbol{\mu}).$$

For optimization with respect to log smoothing parameters ρ we need further derivatives, but note that the third derivatives with respect to $\boldsymbol{\beta}^l$ are zero. The non-zero 3rd derivatives are

$$l^{i\ j\ k}_{\beta^l\beta^m\theta} = -\bar{\mathbf{x}}_i^{l\mathsf{T}}(\mathbf{R}_\theta^{k\mathsf{T}}\mathbf{R} + \mathbf{R}^\mathsf{T}\mathbf{R}_\theta^k)\bar{\mathbf{x}}_j^m,$$

$$l^{i\ jk}_{\beta^l\theta\theta} = \bar{\mathbf{x}}_i^{l\mathsf{T}}(\mathbf{R}^{jk\mathsf{T}}_{\theta\theta}\mathbf{R} + \mathbf{R}_\theta^{j\mathsf{T}}\mathbf{R}_\theta^k + \mathbf{R}_\theta^{k\mathsf{T}}\mathbf{R}_\theta^j + \mathbf{R}^\mathsf{T}\mathbf{R}^{jk}_{\theta\theta})(\mathbf{y}-\boldsymbol{\mu}),$$

$$l^{i\,jk}_{\theta\theta\theta} = -(\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}(\mathbf{R}^{jk\mathsf{T}}_{\theta\theta}\mathbf{R}^i_\theta + \mathbf{R}_\theta^{j\mathsf{T}}\mathbf{R}^{ik}_{\theta\theta} + \mathbf{R}_\theta^{k\mathsf{T}}\mathbf{R}^{ij}_{\theta\theta} + \mathbf{R}^\mathsf{T}\mathbf{R}^{ijk}_{\theta\theta\theta})(\mathbf{y}-\boldsymbol{\mu}).$$

These are useful for computing the following...

$$l^{i\ j\ k}_{\beta^l\beta^m\rho} = l^{i\ j\ q}_{\beta^l\beta^m\theta}\frac{\partial\hat{\theta}_q}{\partial\rho_k},$$

$$l^{i\,jk}_{\theta\theta\rho} = l^{i\,jq}_{\theta\theta\theta}\frac{\partial\hat{\theta}_q}{\partial\rho_k} + l^{q\ ij}_{\beta^l\theta\theta}\frac{\partial\hat{\beta}^l_q}{\partial\rho_k},$$

$$l^{i\ jk}_{\beta^l\theta\rho} = l^{i\ jq}_{\beta^l\theta\theta}\frac{\partial\hat{\theta}_q}{\partial\rho_k} + l^{i\ q\ j}_{\beta^l\beta^m\theta}\frac{\partial\hat{\beta}^m_q}{\partial\rho_k}.$$

This is sufficient for Quasi-Newton estimation of smoothing parameters.

Sometimes models with multiple linear predictors should share some terms across predictors. In this case the general fitting and smoothing parameter methods should work with the vector of unique coefficients, $\bar{\boldsymbol{\beta}}$, say, to which corresponds a model matrix $\bar{\mathbf{X}}$. The likelihood derivative computations on the other hand can operate as if each linear predictor had unique coefficients, with the derivatives then being summed over the copies of each unique parameter. Specifically, let i_{kj} indicate which column of $\bar{\mathbf{X}}$ gives column j of $\mathbf{X}^k$, and let $\mathcal{J}_i = \{k, j : i_{kj} = i\}$, i.e. the set of k, j pairs identifying the replicates of column i of $\bar{X}$ among the $\mathbf{X}^k$, and the replicates of $\bar{\beta}_i$ among the $\boldsymbol{\beta}^k$. Define a '$\mathcal{J}$ contraction over x^k' to be an operation of the form

$$\bar{x}_i = \sum_{k,j\in\mathcal{J}_i} x^k_j \quad \forall i.$$

Then the derivative vectors with respect to $\bar{\boldsymbol{\beta}}$ are obtained by a $\mathcal{J}$ contraction over the derivative vectors with respect to $\boldsymbol{\beta}^k$. Similarly the Hessian with respect to $\bar{\boldsymbol{\beta}}$ is obtained by consecutive $\mathcal{J}$ contractions over the rows and columns of the Hessian with respect to the $\boldsymbol{\beta}^k$. For the computation (4) in section 3.2 we would apply $\mathcal{J}$ contractions to the columns of the two matrices in round brackets on the right hand side of (4) ($\mathbf{B}$ would already be of the correct dimension, of course). The notion of $\mathcal{J}$ contraction simplifies derivation and coding in the case when different predictors reuse terms, but note that computationally it is simpler and more efficient to implement $\mathcal{J}$ contraction based only on the index vector i_{kj}, rather than by explicitly forming the $\mathcal{J}_i$.

I Zero inflated Poisson models

Zero inflated Poisson models are popular in ecological abundance studies when one process determines whether a species is (or could be) present, and the number observed, given presence (suitability), is a Poisson random variable. Several alternatives are possible, but the following 'hurdle model' tends to minimise identifiability problems.

$$f(y) = \begin{cases} 1-p & y = 0 \\ p\lambda^y/\{(e^\lambda - 1)y!\} & \text{otherwise.} \end{cases}$$

So observations greater than zero follow a zero truncated Poisson. Now adopt the unconstrained parameterization, $\gamma = \log\lambda$ and $\eta = \log\{-\log(1-p)\}$ (i.e. using log and complementary log-log links). If $\gamma = \eta$ this recovers an un-inflated Poisson model. The log likelihood is now

$$l = \begin{cases} -e^{\eta} & y = 0 \\ \log(1-e^{-e^{\eta}}) + y\gamma - \log(e^{e^{\gamma}}-1) - \log y! & y > 0. \end{cases}$$

Some care is required to evaluate this without unnecessary overflow, since it is easy for the $1-e^{-e^{\eta}}$ and $e^{e^{\gamma}}-1$ to evaluate as zero in finite precision arithmetic. Hence the limiting results $\log(1-e^{-e^{\eta}}) \to \log(e^{\eta} - e^{2\eta}/2 + e^{3\eta})/6) \to \eta$ as $\eta \to -\infty$ and $\log(e^{e^{\gamma}}-1) \to \log(e^{\gamma} + e^{2\gamma}/2 + e^{3\gamma})/6) \to \gamma$ as $\gamma \to -\infty$ can be used. The first pair of limits is useful as the arguments of the logs becomes too close to 1 and the second pair as the exponential of η or γ approaches underflow to zero. (The log gamma function of $y+1$ computes $\log y!$)

The derivatives for this model are straightforward as all the mixed derivatives are zero. For the $y > 0$ part,

$$l_\eta = \frac{e^{\eta}}{e^{e^{\eta}}-1},\ \ l_\gamma = y - \alpha,\ \ \text{where}\ \ \alpha = \frac{e^{\gamma}}{1-e^{-e^{\gamma}}},\ \ l_{\eta\eta} = (1-e^{\eta})l_\eta - l_\eta^2,\ \ l_{\gamma\gamma} = \alpha^2 - (e^{\gamma}+1)\alpha,$$

$$l_{\eta\eta\eta} = -e^{\eta}l_\eta + (1-e^{\eta})^2 l_\eta - 3(1-e^{\eta})l_\eta^2 + 2l_\eta^3,\ \ l_{\gamma\gamma\gamma} = -2\alpha^3 + 3(e^{\gamma}+1)\alpha^2 - e^{\gamma}\alpha - (e^{\gamma}+1)^2\alpha,$$

$$l_{\eta\eta\eta\eta} = (3e^{\eta}-4)e^{\eta}l_\eta + 4e^{\eta}l_\eta^2 + (1-e^{\eta})^3 l_\eta - 7(1-e^{\eta})^2 l_\eta^2 + 12(1-e^{\eta})l_\eta^3 - 6l_\eta^4$$

$$\text{and}\ \ l_{\gamma\gamma\gamma\gamma} = 6\alpha^4 - 12(e^{\gamma}+1)\alpha^3 + 4e^{\gamma}\alpha^2 + 7(e^{\gamma}+1)^2\alpha^2 - (4+3e^{\gamma})e^{\gamma}\alpha - (e^{\gamma}+1)^3\alpha.$$

As with l itself, some care is required to ensure that the derivatives evaluate accurately and without overflow over as wide a range of γ and η as possible. To this end note that as $\eta \to \infty$ all derivatives with respect to η tend to zero, while as $\gamma \to \infty$, $l_{\gamma\gamma\gamma} \to l_{\gamma\gamma\gamma\gamma} \to -e^{\gamma}$. As $\eta \to -\infty$ accurate evaluation of the derivatives with respect to η rests on $l_\eta \to 1 - e^{\eta}/2 - e^{2\eta}/12$. Substituting this into the derivative expressions, the terms of $O(1)$ can be cancelled analytically: the remaining terms then evaluate the derivatives without cancellation error problems. For $\gamma \to -\infty$ the equivalent approach uses $\alpha \to 1 + e^{\gamma}/2 + e^{\gamma}/12$.

An extended GAM version of this model is also possible, in which η is a function of γ and extra parameters, θ, for example via $\eta = \theta_1 + e^{\theta_2}\gamma$. The idea is that the degree of zero inflation is a non-increasing function of γ, with $\theta_1 = \theta_2 = 0$ recovering the Poisson model. The likelihood expressions are obtained by transformation. Let $\bar{l}_\gamma$ denote the total derivative with respect to γ in such a model.

$$\bar{l}_\gamma = l_\gamma + l_\eta\eta_\gamma,\ \ \bar{l}_{\gamma\gamma} = l_{\gamma\gamma} + l_{\eta\eta}\eta_\gamma\eta_\gamma + l_\eta\eta_{\gamma\gamma},\ \ \bar{l}_{\theta_i} = l_\eta\eta_{\theta_i},\ \ \bar{l}_{\gamma\theta_i} = l_{\eta\eta}\eta_{\theta_i}\eta_\gamma + l_\eta\eta_{\theta\gamma}$$

$$\bar{l}_{\gamma\gamma\theta_i} = l_{\eta\eta\eta}\eta_{\theta_i}\eta_\gamma^2 + l_{\eta\eta}(2\eta_{\gamma\theta_i}\eta_\gamma + \eta_{\gamma\gamma}\eta_{\theta_i}) + l_\eta\eta_{\gamma\gamma\theta_i},\ \ \bar{l}_{\gamma\gamma\gamma} = l_{\gamma\gamma\gamma} + l_{\eta\eta\eta}\eta_\gamma^3 + 3l_{\eta\eta}\eta_\gamma\eta_{\gamma\gamma} + l_\eta\eta_{\gamma\gamma\gamma}$$

$$\bar{l}_{\theta_i\theta_j} = l_{\eta\eta}\eta_{\theta_i}\eta_{\theta_j} + l_\eta\eta_{\theta_i\theta_j},\ \ \ \bar{l}_{\gamma\theta_i\theta_j} = l_{\eta\eta\eta}\eta_{\theta_i}\eta_{\theta_j}\eta_\gamma + l_{\eta\eta}(\eta_{\theta_i\theta_j}\eta_\gamma + \eta_{\theta_i\gamma}\eta_{\theta_j} + \eta_{\theta_j\gamma}\eta_{\theta_i}) + l_\eta\eta_{\theta_i\theta_j\gamma}$$

$$\begin{aligned}\bar{l}_{\gamma\gamma\theta_i\theta_j} = {} & l_{\eta\eta\eta\eta}\eta_{\theta_i}\eta_{\theta_j}\eta_\gamma^2 + l_{\eta\eta\eta}(\eta_{\theta_i\theta_j}\eta_\gamma^2 + 2\eta_{\theta_i}\eta_\gamma\eta_{\theta_j\gamma} + 2\eta_{\theta_j}\eta_\gamma\eta_{\theta_i\gamma} + \eta_{\theta_i}\eta_{\theta_j}\eta_{\gamma\gamma}) \\ & + l_{\eta\eta}(2\eta_{\gamma\theta_i}\eta_{\gamma\theta_j} + 2\eta_\gamma\eta_{\gamma\theta_i\theta_j} + \eta_{\theta_i}\eta_{\gamma\gamma\theta_j} + \eta_{\theta_j}\eta_{\gamma\gamma\theta_i} + \eta_{\theta_i\theta_j}\eta_{\gamma\gamma}) + l_\eta\eta_{\gamma\gamma\theta_i\theta_j}\end{aligned}$$

$$\bar{l}_{\gamma\gamma\gamma\theta_i} = l_{\eta\eta\eta\eta}\eta_{\theta_i}\eta_\gamma^3 + 3l_{\eta\eta\eta}(\eta_\gamma^2\eta_{\theta_i\gamma} + \eta_{\theta_i}\eta_\gamma\eta_{\gamma\gamma}) + l_{\eta\eta}(3\eta_{\theta_i\gamma}\eta_{\gamma\gamma} + 3\eta_\gamma\eta_{\gamma\gamma\theta_i} + \eta_{\theta_i}\eta_{\gamma\gamma\gamma}) + l_\eta\eta_{\gamma\gamma\gamma\theta_i}$$

$$\bar{l}_{\gamma\gamma\gamma\gamma} = l_{\gamma\gamma\gamma\gamma} + l_{\eta\eta\eta\eta}\eta_\gamma^4 + 6l_{\eta\eta\eta}\eta_\gamma^2\eta_{\gamma\gamma} + l_{\eta\eta}(3\eta_{\gamma\gamma}^2 + 4\eta_\gamma\eta_{\gamma\gamma\gamma}) + l_\eta\eta_{\gamma\gamma\gamma\gamma}$$

If $\eta = \theta_1 + e^{\theta_2}\gamma$ then $\eta_{\theta_1} = 1$, $\eta_{\gamma\theta_2\theta_2} = \eta_{\gamma\theta_2} = \eta_\gamma = e^{\theta_2}$, $\eta_\gamma = \eta_{\theta_2\theta_2} = e^{\gamma}e^{\theta_2}$: other required derivatives are 0.

Computationally it makes sense to define the deviance as $-2l$ and the saturated log likelihood as $\tilde{l} = 0$ during model estimation, and only to compute the true $\tilde{l}$ and deviance at the end of fitting, since there is no closed form for $\tilde{l}$ in this case (the same is true for beta regression).

J Tweedie model details

This example illustrates an extended GAM case where the likelihood is not available in closed form. The Tweedie distribution (Tweedie, 1984) has a single θ parameter, p, and a scale parameter, ϕ. We have $V(\mu) = \mu^p$, and a density.

$$f(y) = a(y, \phi, p) \exp\{\mu^{1-p}(y/(1-p) - \mu/(2-p))/\phi\}.$$

We only consider p in (1,2). The difficulty is that

$$a(y, \phi, p) = \frac{1}{y} \sum_{j=1}^{\infty} W_j$$

where, defining $\alpha = (2-p)/(1-p)$,

$$\log W_j = j\,\{\alpha \log(p-1) - \log(\phi)/(p-1) - \log(2-p)\} - \log\Gamma(j+1) - \log\Gamma(-j\alpha) - j\alpha \log y.$$

The sum is interesting in that the early terms are near zero, as are the later terms, so that it has to be summed-from-the-middle, which can be a bit involved: Dunn and Smyth (2005), give the details, but basically they show that the series maximum is around $j_{\max} = y^{2-p}/\{\phi(2-p)\}$.

Let $\omega = \sum_{j=1}^{\infty} W_j$. We need derivatives of $\log\omega$ with respect to $\rho = \log\phi$ and p, or possibly θ where

$$p = \{a + b\exp(\theta)\}/\{1 + \exp(\theta)\}$$

and $1 < a < b < 2$. For optimization this transformation is necessary since the density becomes discontinuous at $p = 1$ and the series length becomes infinite at $p = 2$. It is very easy to produce derivative schemes that overflow, underflow or have cancellation error problems, but the following avoids the worst of these issues. We use the identities

$$\frac{\partial \log\omega}{\partial x} = \text{sign}\left(\frac{\partial\omega}{\partial x}\right) \exp\left(\log\left|\frac{\partial\omega}{\partial x}\right| - \log\omega\right)$$

and

$$\frac{\partial^2 \log\omega}{\partial x \partial z} = \text{sign}\left(\frac{\partial^2\omega}{\partial x\partial z}\right) \exp\left(\log\left|\frac{\partial^2\omega}{\partial x\partial z}\right| - \log\omega\right) - \text{sign}\left(\frac{\partial\omega}{\partial x}\right)\text{sign}\left(\frac{\partial\omega}{\partial z}\right) \exp\left(\log\left|\frac{\partial\omega}{\partial x}\right| + \log\left|\frac{\partial\omega}{\partial x}\right| - 2\log\omega\right).$$

Now

$$\frac{\partial\omega}{\partial x} = \sum_i W_i \frac{\partial \log W_i}{\partial x}$$

while

$$\frac{\partial^2\omega}{\partial x\partial z} = \sum_i W_i \frac{\partial \log W_i}{\partial x}\frac{\partial \log W_i}{\partial z} + W_i \frac{\partial^2 \log W_i}{\partial z \partial x},$$

but note that to avoid over or underflow we can use $W_i' = W_i - \max(W_i)$ in place of W_i in these computations, without changing $\partial\log\omega/\partial x$ or $\partial^2\log\omega/\partial x\partial z$. Note also that

$$\log\omega = \log(\sum_i W_i') + \max(W_i).$$

All that remains is to find the actual derivatives of the $\log W_j$ terms.

$$\frac{\partial \log W_j}{\partial\rho} = \frac{-j}{p-1} \quad \text{and} \quad \frac{\partial^2 \log W_j}{\partial\rho^2} = 0.$$

It is simplest to find the derivatives with respect to p and then transform to those with respect to θ:

$$\frac{\partial^2 \log W_j}{\partial\rho\partial\theta} = \frac{\partial p}{\partial\theta}\frac{j}{(p-1)^2}.$$

The remaining derivatives are a little more complicated

$$\frac{\partial \log W_j}{\partial p} = j\left\{\frac{\log(p-1)+\log\phi}{(1-p)^2} + \frac{\alpha}{p-1} + \frac{1}{2-p}\right\} + \frac{j\psi_0(-j\alpha)}{(1-p)^2} - \frac{j\log y}{(1-p)^2}$$

and

$$\frac{\partial^2 \log W_i}{\partial p^2} = j\left\{\frac{2\log(p-1)+2\log\phi}{(1-p)^3} - \frac{(3\alpha-2)}{(1-p)^2} + \frac{1}{(2-p)^2}\right\} + \frac{2j\psi_0(-j\alpha)}{(1-p)^3} - \frac{j^2\psi_1(-j\alpha)}{(1-p)^4} - \frac{2j\log y}{(1-p)^3}$$

where ψ_0 and ψ_1 are digamma and trigamma functions respectively. These then transform according to

$$\frac{\partial \log W_j}{\partial \theta} = \frac{\partial p}{\partial \theta}\frac{\partial \log W_j}{\partial p} \quad \text{and} \quad \frac{\partial^2 \log W_j}{\partial \theta^2} = \frac{\partial^2 p}{\partial \theta^2}\frac{\partial \log W_j}{\partial p} + \left(\frac{\partial p}{\partial \theta}\right)^2 \frac{\partial^2 \log W_j}{\partial p^2}.$$

The required transform derivatives are

$$\frac{\partial p}{\partial \theta} = \frac{e^\theta(b-a)}{(e^\theta+1)^2} \quad \text{and} \quad \frac{\partial^2 p}{\partial \theta^2} = \frac{e^{2\theta}(a-b)+e^\theta(b-a)}{(e^\theta+1)^3}.$$

K Ordered categorical model details

This example provides a useful illustration of an extended GAM model where the number of θ parameters varies from model to model. The basic model is that y takes a value from $r = 1, \ldots, R$, these being ordered category labels. Given $-\infty = \alpha_0 < \alpha_1, \ldots < \alpha_R = \infty$ we have that $y = r$ if a latent variable $u = \mu + \epsilon$ is such that $\alpha_{r-1} < u \leq \alpha_r$, which occurs with probability

$$\Pr(Y = r) = F(\alpha_r - \mu) - F(\alpha_{r-1} - \mu)$$

where F is the c.d.f. of ϵ. See Kneib and Fahrmeir (2006) and (Fahrmeir et al., 2013, section 6.3.1) for a particularly clear exposition.

$$F(x) = \exp(x)/(1+\exp(x))$$

is usual. For identifiability reasons $\alpha_1 = -1$, or any other constant, so there are $R-2$ free parameters to choose to control the thresholds. Generically we let

$$\alpha_r = \alpha_1 + \sum_{i=1}^{r-1} \exp(\theta_i), \quad 1 < r < R.$$

Note that the cut points in this model *can* be treated as linear parameters in a GLM PIRLS iteration, but this is not a good approach if smoothing parameter estimates are required. The problem is that the cut points are then not forced to be correctly ordered, which means that the PIRLS iteration has to check for this as part of step length control. Worse still, if a category is missing from the data then the derivative of the likelihood with respect to the cut points can be non zero at the best fit, causing implicit differentiation to fail.

Direct differentiation of the $\log \Pr(Y = r) = F(\alpha_r - \mu) - F(\alpha_{r-1} - \mu)$ in terms of θ_i is ugly, and it is better to work with derivatives with respect to the α_r and use the chain rule. The saturated log likelihood can then be expressed as

$$\tilde{l} = \log[F\{(\alpha_r - \alpha_{r-1})/2\} - F\{(\alpha_{r-1} - \alpha_r)/2\}]$$

while the deviance is

$$D = 2[l_s - \log\{F(\alpha_r - \mu) - F(\alpha_{r-1} - \mu)\}].$$

Define $f_1 = F(\alpha_r - \mu)$ and $f_0 = F(\alpha_{r-1} - \mu)$, $f = f_1 - f_0$. Similarly

$$a_1 = f_1^2 - f_1, \quad a_0 = f_0^2 - f_0, \quad a = a_1 - a_0,$$

$$b_j = f_j - 3f_j^2 + 2f_j^3, \quad b = b_1 - b_0$$

$$c_j = -f_j + 7f_j^2 - 12f_j^3 + 6f_j^4, \quad c = c_1 - c_0$$

and

$$d_j = f_j - 15f_j^2 + 50f_j^3 - 60f_j^4 + 24f_j^5, \quad d = d_1 - d_0.$$

The sharp eyed reader will have noticed that all these expressions are prone to severe cancellation error problems as $f_j \to 1$. Stable expressions are required. For f, note that if $b > a$

$$\frac{e^b}{1+e^b} - \frac{e^a}{1+e^a} = \frac{e^{-a} - e^{-b}}{(e^{-b}+1)(e^{-a}+1)} = \frac{1 - e^{a-b}}{(e^{-b}+1)(1+e^a)}$$

The first is used if $0 > b > a$, the second if $b > a > 0$ and the last if $b > 0 > a$. Now writing x as a generic argument of F, we have

$$a_j = \frac{-e^x}{(1+e^x)^2} = \frac{-e^{-x}}{(e^{-x}+1)^2}, \quad b_j = \frac{e^x - e^{2x}}{(1+e^x)^3} = \frac{e^{-2x} - e^{-x}}{(e^{-x}+1)^3}, \quad c_j = \frac{-e^{3x} + 4e^{2x} - e^x}{(1+e^x)^4}$$
$$= \frac{-e^{-x} + 4e^{-2x} - e^{-3x}}{(e^{-x}+1)^4}, \quad d_j = \frac{-e^{4x} + 11e^{3x} - 11e^{2x} + e^x}{(1+e^x)^5} = \frac{-e^{-x} + 11e^{-2x} - 11e^{-3x} + e^{-4x}}{(e^{-x}+1)^5}$$

These are useful by virtue of involving terms of order 0, rather than 1.

Then

$$D_\mu = -2a/f, \quad D_{\mu\mu} = 2a^2/f^2 - 2b/f, \quad D_{\mu\mu\mu} = -2c/f - 4a^3/f^3 + 6ab/f^2.$$

Note that $D_{\mu\mu} \geq 0$.

$$D_{\mu\mu\mu\mu} = 6b^2/f^2 + 8ac/f^2 + 12a^4/f^4 - 24a^2b/f^3 - 2d/f,$$
$$D_{\mu\alpha_{r-1}} = 2a_0a/f^2 - 2b_0/f, \quad D_{\mu\alpha_r} = -2a_1a/f^2 + 2b_1/f,$$
$$D_{\mu\mu\alpha_{r-1}} = -2c_0/f + 4b_0a/f^2 - 4a_0a^2/f^3 + 2a_0b/f^2, \quad D_{\mu\mu\alpha_r} = 2c_1/f - 4b_1a/f^2 + 4a_1a^2/f^3 - 2a_1b/f^2,$$
$$D_{\mu\mu\mu\alpha_{r-1}} = -2d_0/f + 2a_0c/f^2 + 6c_0a/f^2 - 12b_0a^2/f^3 + 12a_0a^3/f^4 + 6b_0b/f^2 - 12a_0ab/f^3,$$
$$D_{\mu\mu\mu\alpha_r} = 2d_1/f - 2a_1c/f^2 - 6c_1a/f^2 + 12b_1a^2/f^3 - 12a_1a^3/f^4 - 6b_1b/f^2 + 12a_1ab/f^3.$$

Furthermore,

$$D_{\mu\alpha_{r-1}\alpha_{r-1}} = 2c_0/f - 2b_0a/f^2 + 4a_0b_0/f^2 - 4a_0^2a/f^3, \quad D_{\mu\alpha_r\alpha_r} = -2c_1/f + 2b_1a/f^2 + 4a_1b_1/f^2 - 4a_1^2a/f^3,$$
$$D_{\mu\alpha_{r-1}\alpha_r} = -2a_0b_1/f^2 - 2b_0a_1/f^2 + 4a_0a_1a/f^3,$$

while

$$D_{\mu\mu\alpha_{r-1}\alpha_{r-1}} = 2d_0/f - 4c_0a/f^2 + 4b_0^2/f^2 + 4a_0c_0/f^2 + 4b_0a^2/f^3 - 16a_0b_0a/f^3 + 12a_0^2a^2/f^4 - 2b_0b/f^2 - 4a_0^2b/f^3,$$
$$D_{\mu\mu\alpha_r\alpha_r} = -2d_1/f + 4c_1a/f^2 + 4b_1^2/f^2 + 4a_1c_1/f^2 - 4b_1a^2/f^3 - 16a_1b_1a/f^3 + 12a_1^2a^2/f^4 + 2b_1b/f^2 - 4a_1^2b/f^3,$$
$$D_{\mu\mu\alpha_{r-1}\alpha_r} = 0.$$

Finally there are some derivatives not involving μ and hence involving the terms in $\tilde{l}$. First define

$$\bar{\alpha} = (\alpha_r - \alpha_{r-1})/2, \quad \gamma_1 = F(\bar{\alpha}), \quad \gamma_0 = F(-\bar{\alpha}),$$
$$A = \gamma_1 - \gamma_0, \quad B = \gamma_1^2 - \gamma_1 + \gamma_0^2 - \gamma_0, \quad C = 2\gamma_1^3 - 3\gamma_1^2 + \gamma_1 - 2\gamma_0^3 + 3\gamma_0^2 - \gamma_0.$$

Then

$$D_{\alpha_{r-1}} = B/A - 2a_0/f, \quad D_{\alpha_r} = -B/A + 2a_1/f$$

and

$$D_{\alpha_{r-1}\alpha_{r-1}} = 2b_0/f + 2a_0^2/f^2 + C/(2A) - B^2/(2A^2), \quad D_{\alpha_r\alpha_r} = -2b_1/f + 2a_1^2/f^2 + C/(2A) - B^2/(2A^2),$$
$$D_{\alpha_r\alpha_{r-1}} = -2a_0a_1/f^2 - C/(2A) + B^2/(2A^2).$$

The derivatives of $\tilde{l}$ can be read from these expressions.

Having expressed things this way, it is necessary to transform to derivatives with respect to $\boldsymbol{\theta}$.

$$\frac{\partial D}{\partial \theta_k} = \begin{cases} 0 & r \leq k \\ \exp(\theta_k)\partial D/\partial \alpha_r & r = k+1 \\ \exp(\theta_k)(\partial D/\partial \alpha_r + \partial D/\partial \alpha_{r-1}) & r > k+1 \\ \exp(\theta_k)\partial D/\partial \alpha_{r-1} & r = R. \end{cases}$$

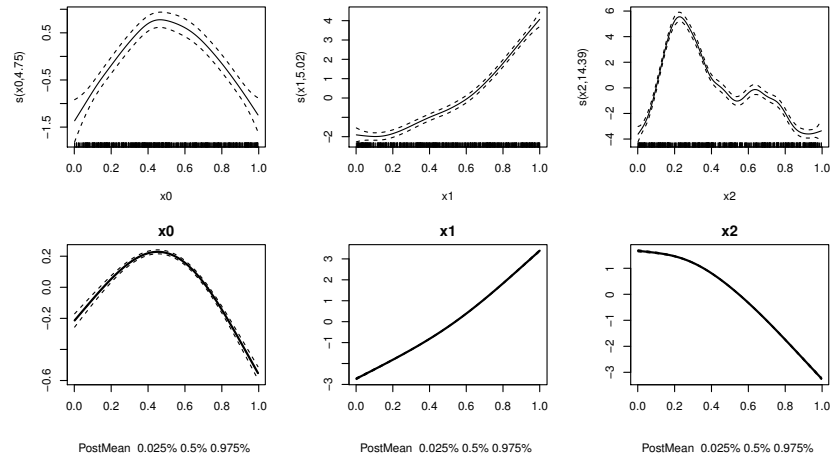

Figure 9: `mgcv` (top row) and INLA (lower row) fits (with 95% credible intervals) to the simple 3 term additive model simulated in Appendix L. Each row is supposed to be reconstructing the same true function, which in reality looks like the estimate in the upper row. On this occasion INLA encounters numerical stability problems and the estimates are poor.

L Example comparison with INLA and JAGS

As mentioned in the main text, for models that require high rank random fields INLA offers a clearly superior approach to the methods proposed here, but at the cost of requiring sparse matrix methods, which preclude stabilizing reparameterization or pivoting for stability. On occasion this has noticeable effects on inference. For example the following code is adapted from the first example in the `gam` helpfile in R package `mgcv`.

```
library(mgcv);library(INLA)
n <- 500; set.seed(0) ## simulate some data...
dat <- gamSim(1,n=n,dist="normal",scale=1)
k=20;m=2
b <- gam(y~s(x0,k=k,m=m)+s(x1,k=k,m=m)+s(x2,k=k,m=m),
                  data=dat,method="REML")
md <- "rw2"
b2 <- inla(y~f(x0,model=md)+ f(x1,model=md)+
             f(x2,model=md),data=dat,verbose=TRUE)})
```

On the same dual core laptop computer the `gam` fit took 0.2 seconds and the INLA fit 40.9 seconds. Figure 9 compares the function estimates: INLA has encountered numerical stability problems and the reconstructions, which should look like those on the top row of the figure, are poor. Replicate simulations often give INLA results close to the truth, indistinguishable from the `mgcv` results and computed in less than 1 second, but the shown example is not unusual. For this example we can fix the problem by binning the covariates, in which case the estimates and intervals are almost indistinguishable from the `gam` estimates. However the necessity of doing this does emphasise that the use of sparse matrix methods precludes the use of pivoting to alleviate the effects of poor model conditioning.

The same example can be coded in JAGS, for example using the function `jagam` from `mgcv` to auto-generate the JAGS model specification and starting values. To obtain samples giving comparable results to the top row of figure 9 took about 16 seconds, emphasising that simulation is relatively expensive for these models.

M Software implementation

We have implemented the proposed framework in R (R Core Team, 2014), by extending package `mgcv` (from version 1.8-0), so that the `gam` function can estimate all the models mentioned in this paper, in a manner that is

intuitively straightforward for anyone familiar with GAMs for exponential family distributions. Implementation was greatly facilitated by use of Bravington (2013). For the beta, Tweedie, negative binomial, scaled t, ordered categorical, simple zero inflated Poisson and Cox proportional hazards models, the user simply supplies one of the families `betar`, `tw`, `nb`, `scat`, `ocat`, `ziP` or `cox.ph` to `gam` as the `family` argument, in place of the usual exponential family `family`. For example the call to fit a Cox proportional hazards model is something like.

```
gam(time ~ s(x) + s(z),family=cox.ph,weights=censor)
```

where `censor` would contain a 0 for a censored observation, and a 1 otherwise. Model `summary` and `plot` functions work exactly as for any GAM, while `predict` allows for prediction on the survival scale.

Linear functionals of smooths are incorporated as model components using a simple summation convention. Suppose that `X` and `L` are matrices. Then a model term `S(X,by=L)` indicates a contribution to the i^{th} row of the linear predictor of the form $\sum_j f(X_{ij})L_{ij}$. This is the way that the section 8 model is estimated.

For models with multiple linear predictors `gam` accepts a list of model formulae. For example a GAMLSS style zero inflated Poisson model would be estimated with something like

```
gam(list(y ~ s(x) + s(z),~ s(v)+s(w)),family=ziplss)
```

where the first formula specifies the response and the linear predictor the Poisson parameter given presence, while the second, one sided, formula specifies the linear predictor for presence. `gaulss` and `multinom` families provide further examples.

Similarly a multivariate normal model is fit with something like

```
gam(list(y1 ~ s(x) + s(z),y2 ~ s(v)+s(w)),family=mvn(d=2))
```

where each formula now specifies one component of the multivariate response and the linear predictor for its mean. There are also facilities to allow terms to be shared by different linear predictors, for example

```
gam(list(y1 ~ s(x),y2 ~ s(v),y3 ~ 1 ,1 + 3 ~ s(z) - 1),family=mvn(d=3))
```

specifies a multivariate normal model in which the linear predictors for the first (`y1`) and third (`y3`) components of the response share the same dependence on a smooth of z.

The software is general and can accept an arbitrary number of formulae as well as dealing with the identifiability issues that can arise between parametric components when linear predictors share terms. Summary and plotting functions label model terms by component, and prediction produces matrix predictions when appropriate.

References

Agarwal, G. G. and W. Studden (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics*, 1307–1325.

Andersen, P. K., M. W. Bentzon, and J. P. Klein (1996). Estimating the survival function in the proportional hazards regression model: a study of the small sample size properties. *Scandinavian journal of statistics*, 1–12.

Bravington, M. V. (2013). *debug: MVB's debugger for R*. R package version 1.3.1.

Claeskens, G., T. Krivobokova, and J. D. Opsomer (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Cox, D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. D. (1983). Asymptotics for m-type smoothing splines. *The Annals of Statistics*, 530–551.

de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). New York: Springer.

Demmler, A. and C. Reinsch (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik 24*(5), 375–382.

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

Dunn, P. K. and G. K. Smyth (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing 15*(4), 267–280.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression Models*. Springer.

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

Golub, G. H. and C. F. Van Loan (2013). *Matrix computations* (4th ed.). Baltimore: Johns Hopkins University Press.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

Gu, C. and Y. J. Kim (2002). Penalized likelihood regression: general approximation and efficient approximation. *Canadian Journal of Statistics 34*(4), 619–628.

Gu, C. and G. Wahba (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing 12*(2), 383–398.

Hall, C. A. and W. W. Meyer (1976). Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory 16*(2), 105–122.

Hall, P. and J. D. Opsomer (2005). Theory for penalised spline regression. *Biometrika 92*(1), 105–118.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall.

Kauermann, G., T. Krivobokova, and L. Fahrmeir (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 487–503.

Klein, J. and M. Moeschberger (2003). *Survival analysis: techniques for censored and truncated data* (2nd ed.). New York: Springer.

Kneib, T. and L. Fahrmeir (2006). Structured additive regression for categorical space–time data: A mixed model approach. *Biometrics 62*(1), 109–118.

Lancaster, P. and K. Šalkauskas (1986). *Curve and Surface Fitting: An Introduction*. London: Academic Press.

Miller, D. L. and S. N. Wood (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*, 1–17.

Nychka, D. and D. Cummins (1996). Comment on 'Flexible smoothing with B-splines and penalties' by PHC Eilers and BD Marx. *Statist. Sci 89*, 104–5. Demmler Reinsch basis and P-splines.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shun, Z. and P. McCullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(4), 749–760.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, 970–983.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053.

Tweedie, M. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions: Proc. Indian Statistical Institute Golden Jubilee International Conference*, pp. 579–604.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 1378–1402.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B 65*, 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Yoshida, T. and K. Naito (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics 26*(2), 269–289.

Zhou, S. and D. A. Wolfe (2000). On derivative estimation in spline regression. *Statistica Sinica 10*(1), 93–108.