JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, 1548–1575, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1180986

Taylor & Francis
Taylor & Francis Group

Smoothing Parameter and Model Selection for General Smooth Models

Simon N. Wood[a], Natalya Pya[b], and Benjamin Säfken[c]

[a]School of Mathematics, University of Bristol, Bristol, UK; [b]School of Science and Technology, Nazarbayev University, Astana, Kazakhstan, and KIMEP University, Almaty, Kazakhstan; [c]Chairs of Statistics and Econometrics, Georg-August-Universität Göttingen, Germany

ABSTRACT

This article discusses a general framework for smoothing parameter estimation for models with regular likelihoods constructed in terms of unknown smooth functions of covariates. Gaussian random effects and parametric terms may also be present. By construction the method is numerically stable and convergent, and enables smoothing parameter uncertainty to be quantified. The latter enables us to fix a well known problem with AIC for such models, thereby improving the range of model selection tools available. The smooth functions are represented by reduced rank spline like smoothers, with associated quadratic penalties measuring function smoothness. Model estimation is by penalized likelihood maximization, where the smoothing parameters controlling the extent of penalization are estimated by Laplace approximate marginal likelihood. The methods cover, for example, generalized additive models for nonexponential family responses (e.g., beta, ordered categorical, scaled t distribution, negative binomial and Tweedie distributions), generalized additive models for location scale and shape (e.g., two stage zero inflation models, and Gaussian location-scale models), Cox proportional hazards models and multivariate additive models. The framework reduces the implementation of new model classes to the coding of some standard derivatives of the log-likelihood. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2015
Revised March 2016

KEYWORDS
Additive model; AIC; Distributional regression; GAM; Location scale and shape model; Ordered categorical regression; Penalized regression spline; REML; Smooth Cox model; Smoothing parameter uncertainty; Statistical algorithm; Tweedie distribution.

1. Introduction

This article is about smoothing parameter estimation and model selection in statistical models with a smooth regular likelihood, where the likelihood depends on smooth functions of covariates and these smooth functions are the targets of inference. Simple Gaussian random effects and parametric dependencies may also be present. When the likelihood (or a quasi-likelihood) decomposes into a sum of independent terms each contributed by a response variable from a single parameter exponential family distribution, then such a model is a generalized additive model (GAM, Hastie and Tibshirani 1986, 1990). GAMs are widely used in practice (see, e.g., Ruppert, Wand, and Carroll 2003; Fahrmeir et al. 2013) with their popularity resting in part on the availability of statistically well founded smoothing parameter estimation methods that are numerically efficient and robust (Wood 2000, 2011) and perform the important task of estimating how smooth the component functions of a model should be.

The purpose of this article is to provide a general method for smoothing parameter estimation when the model likelihood does not have the convenient exponential family (or quasi-likelihood) form. For the most part we have in mind regression models of some sort, but the proposed methods are not limited to this setting. The simplest examples of the extension are generalized additive models where the response distribution is not in the single parameter exponential family. For example, when the response has a Tweedie, negative binomial, beta, scaled t, or some sort of ordered categorical or zero inflated distribution. Examples of models with a less GAM like likelihood structure are Cox proportional hazard and Cox process models, scale-location models, such as the GAMLSS class of Rigby and Stasinopoulos (2005), and multivariate additive models (e.g., Yee and Wild 1996). Smooth function estimation for such models is not new: what is new here is the general approach to smoothing parameter estimation, and the wide variety of smooth model components that it admits.

The proposed method broadly follows the strategy of Wood (2011) that has proved successful for the GAM class. The smooth functions will be represented using reduced rank spline bases with associated smoothing penalties that are quadratic in the spline coefficients. There is now a substantial literature showing that the reduced rank approach is well-founded, and the basic issues are covered in an online Supplementary Appendix A (henceforth online "SA A"). More importantly, from an applied perspective, a wide range of spline and Gaussian process terms can be included as model components by adopting this approach (Figure 1). We propose to estimate smoothing parameters by Newton optimization of a Laplace approximate marginal likelihood criterion, with each Newton step requiring an inner Newton iteration to find maximum penalized likelihood estimates of

CONTACT Simon N. Wood simon.wood@bath.edu School of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

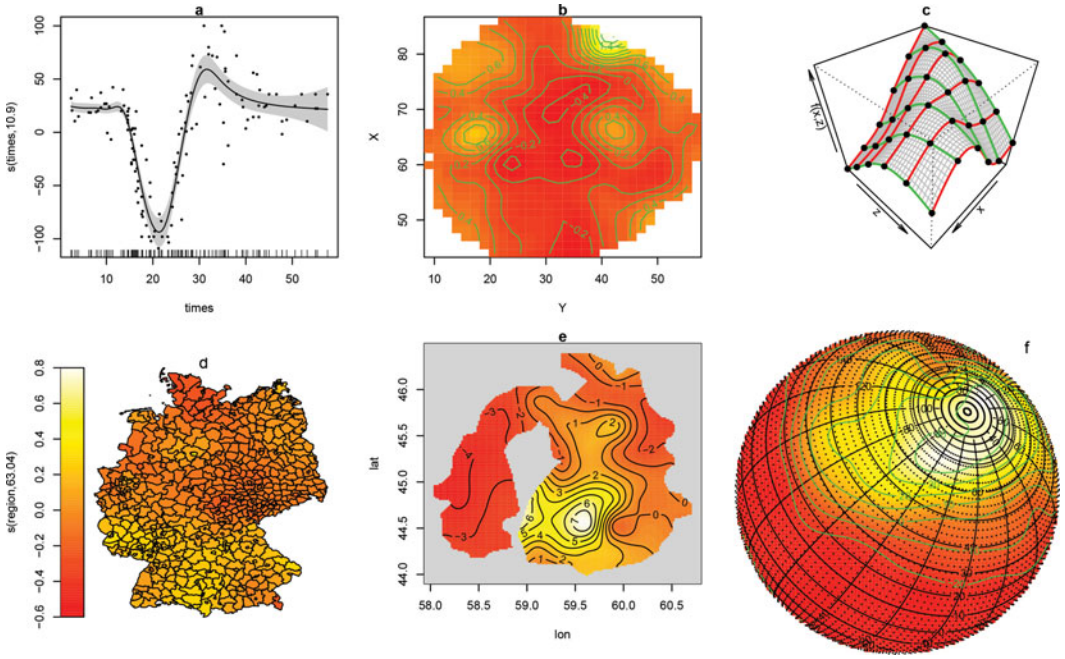

Figure 1. Examples of the rich variety of smooth model components that can be represented as reduced rank basis smoothers, with quadratic penalties and therefore can routinely be incorporated as components of a GAM. This article develops methods to allow their routine use in a much wider class of models. (a) One dimensional smooths such as cubic, P- and adaptive splines. (b) isotropic smooths of several variables, such as thin plate splines and Duchon splines. (c) Nonisotropic tensor product splines used to model smooth interactions. (d) Gaussian Markov random fields for data on discrete geographies. (e) Finite area smoothers, such as soap film smoothers. (f) Splines on the sphere. Another important class are simple Gaussian random effects.

the model coefficients. Implicit differentiation is used to obtain derivatives of the coefficients with respect to the smoothing parameters. This basic strategy works well in the GAM setting, but is substantially more complex when the simplifications of a GLM type likelihood no longer apply.

Our aim is to provide a general method that is as numerically efficient and robust as the GAM methods, such that (i) implementation of a model class requires only the coding of some standard derivatives of the log-likelihood for that class and (ii) much of the inferential machinery for working with such models can reuse GAM methods (e.g., interval estimation or p-value computations). An important consequence of our approach is that we are able to compute a simple correction to the conditional AIC for the models considered, which corrects for smoothing parameter estimation uncertainty and the consequent deficiencies in a conventionally computed conditional AIC (see Greven and Kneib 2010). This facilitates the part of model selection distinct from smoothing parameter estimation.

The article is structured as follows. Section 2 introduces the general modeling framework. Section 3 then covers smoothness selection methods for this framework, with Section 3.1 developing a general method, Section 3.2 illustrating its use for the special case of distributional regression, and Section 3.3 covering the simplified methods that can be used in the even more restricted case of models with a similar structure to generalized additive models. Section 4 then develops approximate distributional results accounting for smoothing parameter uncertainty which are applied in Section 5 to propose a corrected AIC suitable for the general model class. The remaining sections present simulation results and examples, while extensive further background, and details for particular models, are given in the supplementary appendices (referred to as online "SA A," "SA B," etc., below).

2. The General Framework

Consider a model for an n-vector of data, $\mathbf{y}$, constructed in terms of unknown parameters, $\boldsymbol{\theta}$, and some unknown functions, g_j, of covariates, x_j. Suppose that the log-likelihood for this model satisfies the Fisher regularity conditions, has four continuous derivatives, and can be written $l(\boldsymbol{\theta}, g_1, g_2, \ldots, g_M) = \log f(\mathbf{y}|\boldsymbol{\theta}, g_1, g_2, \ldots, g_M)$. In contrast to the usual GAM case, the likelihood need not be based on a single parameter exponential family distribution, and we do not assume that the log-likelihood can be written in terms of a single additive linear predictor. Now let the $g_j(x_j)$ be represented via basis expansions of modest rank (k_j),

$$g_j(x) = \sum_{i=1}^{k_j} \beta_{ji} b_{ji}(x),$$

where the β_{ji} are unknown coefficients and the $b_{ji}(x)$ are known basis functions such as splines, usually chosen to have good approximation theoretical properties. With each g_j is associated a smoothing penalty, which is quadratic in the basis coefficients and measures the complexity of g_j. Writing all the basis coefficients and $\boldsymbol{\theta}$ in one p-vector $\boldsymbol{\beta}$, then the jth smoothing penalty can be written as $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^j\boldsymbol{\beta}$, where $\mathbf{S}^j$ is a matrix of known coefficients, but generally has only a small nonzero block. The estimated model coefficients are then

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmax}} \left\{ l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j}^{M} \lambda_j \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^j\boldsymbol{\beta} \right\} \tag{1}$$

given M smoothing parameters, λ_j, controlling the extent of penalization. A slight extension is that the smoothing penalties may be such that several $\lambda_i\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^i\boldsymbol{\beta}$ are associated with one g_j, for

example when g_j is a nonisotropic function of several variables. Note also that the framework can incorporate Gaussian random effects, provided the corresponding precision matrices can be written as $\sum \lambda_i \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}^i \boldsymbol{\beta}$ (where the $\mathbf{S}^i$ are known).

From a Bayesian viewpoint $\hat{\boldsymbol{\beta}}$ is a posterior mode for $\boldsymbol{\beta}$. The Bayesian approach views the smooth functions as intrinsic Gaussian random fields with prior f_λ given by $N(\mathbf{0}, \mathbf{S}^{\lambda-})$ where $\mathbf{S}^{\lambda-}$ is a Moore–Penrose (or other suitable) pseudoinverse of $\sum_j \lambda_j \mathbf{S}^j$. Then the posterior modes are $\hat{\boldsymbol{\beta}}$ from (1), and in the large sample limit, assuming fixed smoothing parameter vector, $\boldsymbol{\lambda}$, we have $\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\mathcal{I} + \mathbf{S}^\lambda)^{-1})$, where $\mathcal{I}$ is the expected negative Hessian of the log-likelihood (or its observed version) at $\hat{\boldsymbol{\beta}}$. An empirical Bayesian approach is appealing here as it gives well calibrated inference for the g_j (Wahba 1983; Silverman 1985; Nychka 1988; Marra and Wood 2012) in a GAM context. Appropriate summations of the elements of $\text{diag}\{(\mathcal{I} + \mathbf{S}^\lambda)^{-1}\mathcal{I}\}$ provide estimates of the "effective degrees of freedom" of the whole model, or of individual smooths.

Under this Bayesian view, smoothing parameters can be estimated to maximize the log marginal likelihood

$$\mathcal{V}_r(\boldsymbol{\lambda}) = \log \int f(\mathbf{y}|\boldsymbol{\beta}) f_\lambda(\boldsymbol{\beta}) d\boldsymbol{\beta}, \tag{2}$$

or a Laplace approximate version of this (e.g., Wood 2011). In practice optimization is with respect to $\boldsymbol{\rho}$ where $\rho_i = \log \lambda_i$. Marginal likelihood estimation of smoothing parameters in a Gaussian context goes back to Anderssen and Bloomfield (1974) and Wahba (1985), while Shun and McCullagh (1995) showed that Laplace approximation of more general likelihoods is theoretically well founded. That marginal likelihood is equivalent to REML (in the sense of Laird and Ware 1982) supports its use when the model contains Gaussian random effects. Theoretical work by Reiss and Ogden (2009) also suggests practical advantages at finite sample sizes, in that marginal likelihood is less prone to multiple local minima than GCV (or AIC). Supplementary Appendix B (SA B) also demonstrates how Laplace approximate marginal likelihood (LAML) estimation of smoothing parameters maintains statistical consistency of reduced rank spline estimates. The use of Laplace approximation and demonstration of statistical consistency requires the assumption that $\dim(\boldsymbol{\beta}) = O(n^\alpha)$ where $\alpha < 1/3$.

3. Smoothness Selection Methods

This section describes the general smoothness selection method, and a simplified method for the special case in which the likelihood is a simple sum of terms for each observation of a univariate response, and there is a single GAM like linear predictor.

The nonlinear dependencies implied by employing a general smooth likelihood result in unwieldy expressions unless some care is taken to establish a compact notation. In the rest of this article, Greek subscripts denote partial differentiation with respect to the given variable, while Roman superscripts are indices associated with the derivatives. Hence, $D^{ij}_{\beta\theta} = \partial^2 D/\partial\beta_i \partial\theta_j$. Similarly $D^{ij}_{\beta\theta} = \partial^2 D/\partial\beta_i \partial\theta_j|_{\hat{\beta}}$. Roman subscripts denote vector or array element indices. For matrices the first Roman sub- or superscript denotes rows, the second columns. Roman superscripts without a corresponding Greek subscript are labels, for example $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ denote two separate vectors $\boldsymbol{\beta}$. For Hessian matrices only, $D^{\beta\theta}_{i\ j}$ is element i, j of the inverse of the matrix with elements $D^{i\ j}_{\beta\theta}$. If any Roman index appears in two or more multiplied terms, but the index is absent on the other side of the equation, then a summation over the product of the corresponding terms is indicated (the usual Einstein summation convention being somewhat unwieldy in this context). To aid readability, in this article summation indices will be highlighted in bold. For example, the equation $a_{ij}b_{ik}c^{il} + d_{jkl} = 0$ is equivalent to $\sum_i a_{ij}b_{ik}c^{il} + d_{jkl} = 0$. An indexed expression not in an equation is treated like an equation with no indices on the other side (so $a_{ij}b_{\boldsymbol{j}}$ is interpreted as $\sum_j a_{ij}b_j$).

3.1. General Model Estimation

Consider the general case in which the log-likelihood depends on several smooth functions of predictor variables, each represented via a basis expansion and each with one or more associated penalties. The likelihood may also depend on some strictly parametric model components. The log-likelihood is assumed to satisfy the Fisher regularity conditions and in addition we usually assume that it has 4 bounded continuous derivatives with respect to the parameters (with respect to $g_j(x)$ for any relevant fixed x in the case of a smooth, g_j). Let the model coefficients be $\boldsymbol{\beta}$ (recalling that this includes the vector $\boldsymbol{\theta}$ of parametric coefficients and nuisance parameters). The penalized log-likelihood is then

$$\mathcal{L}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\lambda_{\boldsymbol{j}} \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}^{\boldsymbol{j}} \boldsymbol{\beta},$$

and we assume that the model is well enough posed that this has a positive definite maximum (at least after dealing with any parameter redundancy issues that can be addressed by linear constraint). Let $\hat{\boldsymbol{\beta}}$ be the maximizer of $\mathcal{L}$ and let $\boldsymbol{\mathcal{H}}$ be the negative Hessian, with elements $-\mathcal{L}^{i\ j}_{\hat{\beta}\hat{\beta}}$. The log LAML (see online SA C) is

$$\mathcal{V}(\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\log|\mathbf{S}^\lambda|_+ - \frac{1}{2}\log|\boldsymbol{\mathcal{H}}| + \frac{M_p}{2}\log(2\pi),$$

where $\mathbf{S}^\lambda = \lambda_{\boldsymbol{j}}\mathbf{S}^{\boldsymbol{j}}$ and $|\mathbf{S}^\lambda|_+$ is the product of the positive eigenvalues of $\mathbf{S}^\lambda$. M_p is the number of zero eigenvalues of $\mathbf{S}^\lambda$, when all λ_j are strictly positive. The basic strategy is to optimize $\mathcal{V}$ with respect to $\boldsymbol{\rho} = \log(\boldsymbol{\lambda})$ via Newton's method. This requires $\hat{\boldsymbol{\beta}}$ to be obtained for each trial $\boldsymbol{\rho}$ via an inner Newton iteration, and derivatives of $\hat{\boldsymbol{\beta}}$ must be obtained by implicit differentiation. The log determinant computations have the potential to be computationally unstable, and reparameterization is needed to deal with this. The full Newton method based on computationally exact derivatives has the substantial practical advantage that it can readily be detected when $\mathcal{V}$ is indefinite with respect to a particular ρ_i, since then $\partial\mathcal{V}/\partial\rho_i = \partial^2\mathcal{V}/\partial\rho_i^2 \simeq 0$. Such indefiniteness occurs when a smoothing parameter, $\lambda_i, \to \infty$ or a variance component tends to zero, both of which are perfectly legitimate. Dropping a ρ_i from Newton update when such indefiniteness is detected ensures that it takes a value which can be treated as "working infinity" without overflowing. Methods

which use an approximate Hessian, or none, do not have this advantage.

The proposed general method consists of outer and inner iterations, as follows.

Outer algorithm for $\boldsymbol{\rho}$

1. Obtain initial values for $\boldsymbol{\rho} = \log(\boldsymbol{\lambda})$, to ensure that the effective degrees of freedom of each smooth lies away from its maximum or minimum possible values.
2. Find initial $\hat{\boldsymbol{\beta}}$ guesstimates (model specific).
3. Perform the initial reparameterizations required in Section 3.1.1 to facilitate stable computation of $\log|\mathbf{S}^\lambda|_+$.
4. Repeat the following standard Newton iteration until convergence is detected at Step (c).
   (a) Find $\hat{\boldsymbol{\beta}}$, $\mathcal{V}_\rho^i$ and $\mathcal{V}_{\rho\rho}^{ij}$ by the inner algorithm.
   (b) Drop any $\mathcal{V}_\rho^i$, $\mathcal{V}_{\rho\rho}^{ij}$ and $\mathcal{V}_{\rho\rho}^{ji}$ for which $\mathcal{V}_\rho^i \simeq \mathcal{V}_{\rho\rho}^{ii} \simeq 0$. Let $\mathbb{I}$ denote the indices of the retained terms.
   (c) Test for convergence, that is, all $\mathcal{V}_\rho^i \simeq 0$ and the Hessian (elements $-\mathcal{V}_{\rho\rho}^{ji}$) is positive semidefinite.
   (d) If necessary perturb the Hessian (elements $-\mathcal{V}_{\rho\rho}^{ji}$) to make it positive definite (guaranteeing that the Newton step will be a descent direction).
   (e) Define $\boldsymbol{\Delta}_\mathbb{I}$ as the subvector of $\boldsymbol{\Delta}$ indexed by $\mathbb{I}$, with elements $-\mathcal{V}_{ij}^{\rho\rho}\mathcal{V}_\rho^j$, and set $\Delta_j = 0\ \forall\ j \notin \mathbb{I}$.
   (f) While $\mathcal{V}(\boldsymbol{\rho}+\boldsymbol{\Delta}) < \mathcal{V}(\boldsymbol{\rho})$ set $\boldsymbol{\Delta} \leftarrow \boldsymbol{\Delta}/2$.
   (g) Set $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho} + \boldsymbol{\Delta}$.
5. Reverse the Step 3 reparameterization.

The method for evaluating $\mathcal{V}$ and its gradient and Hessian with respect to $\boldsymbol{\rho}$ is as follows, where $\mathcal{L}^{\hat\beta\hat\beta}_{kj}$ denotes the inverse of $\mathcal{L}^{kj}_{\hat\beta\hat\beta}$.

Inner algorithm for $\boldsymbol{\beta}$

1. Reparameterize to deal with any "type 3" penalty blocks as described in Section 3.1.1 so that computation of $\log|\mathbf{S}^\lambda|_+$ is stable, and evaluate the derivatives of $\log|\mathbf{S}^\lambda|_+$.
2. Use Newton's method to find $\hat{\boldsymbol{\beta}}$, regularizing the Hessian, and applying step length control, to ensure convergence even when the Hessian is indefinite and/or $\hat{\boldsymbol{\beta}}$ is not identifiable, as described in Section 3.1.2.
3. Test for identifiability of $\hat{\boldsymbol{\beta}}$ at convergence by examining the rank of the $\mathcal{H}$ as described in Section 3.1.2. Drop unidentifiable coefficients.
4. If coefficients were dropped, find the reduced $\hat{\boldsymbol{\beta}}$ by further steps of Newton's method (Section 3.1.2).
5. Compute $\mathrm{d}\hat\beta_i/\mathrm{d}\rho_k = \mathcal{L}^{\hat\beta\hat\beta}_{ij}\lambda_k S^k_{jl}\hat\beta_l$ and hence $l^{ij}_{\hat\beta\hat\beta\rho} = l^{ijk}_{\hat\beta\hat\beta\hat\beta}\mathrm{d}\hat\beta_k/\mathrm{d}\rho_l$ (Section 3.1.3).
6. Compute $\mathrm{d}^2\hat\beta_i/\mathrm{d}\rho_k\mathrm{d}\rho_l = \mathcal{L}^{\hat\beta\hat\beta}_{ij}\{(-l^{jpl}_{\hat\beta\hat\beta\rho} + \lambda_l S^l_{jp})\mathrm{d}\hat\beta_p/\mathrm{d}\rho_k + \lambda_k S^k_{jp}\mathrm{d}\hat\beta_p/\mathrm{d}\rho_l\} + \delta^l_k\mathrm{d}\hat\beta_i/\mathrm{d}\rho_k$, (Section 3.1.3).
7. Compute $\mathcal{L}^{\hat\beta\hat\beta}_{kj}l^{jkpv}_{\hat\beta\hat\beta\rho\rho}$ (model specific). (3.1.3)
8. The derivatives of $\mathcal{V}$ can now be computed according to Section 3.1.4.
9. For each parameter dropped from $\hat{\boldsymbol{\beta}}$ during fitting, zeroes must be inserted in $\hat{\boldsymbol{\beta}}$, $\partial\hat{\boldsymbol{\beta}}/\partial\rho_j$ and the corresponding rows and columns of $\mathcal{L}^{\hat\beta\hat\beta}_{kj}$. The Step 1 reparameterization is then reversed.

The following subsections fill in the method details, but note that to implement a particular model in this class it is necessary to be able to compute, l, l^i_β and $l^{ij}_{\beta\beta}$, given $\boldsymbol{\beta}$, along with $l^{ijk}_{\hat\beta\hat\beta\rho}$ given $\mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\rho_k$, and $\mathcal{L}^{\hat\beta\hat\beta}_{kj}l^{jkpv}_{\hat\beta\hat\beta\rho\rho}$ given $\mathrm{d}^2\hat{\boldsymbol{\beta}}/\mathrm{d}\rho_k\mathrm{d}\rho_l$. The last of these is usually computable much more efficiently than if $l^{jkpv}_{\hat\beta\hat\beta\rho\rho}$ was computed explicitly.

3.1.1. Derivatives and Stable Evaluation of $\log|\mathbf{S}^\lambda|_+$

This section covers the details for outer Step 3 and inner Step 1. Stable evaluation of the log determinant terms is the key to stable computation with the LAML. The online SA C explains the issue. Wood (2011) proposed a solution which involves orthogonal transformation of the whole parameter vector $\boldsymbol{\beta}$, but in the general case the likelihood may depend on each smooth function separately and such a transformation is therefore untenable. It is necessary to develop a reparameterization strategy which does not combine coefficients from different smooths. This is possible if we recognize that $\mathbf{S}^\lambda$ is block diagonal, with different blocks relating to different smooths. For example, if $\mathbb{S}^j$ denotes the nonzero sub-block of $\mathbf{S}^j$,

$$\mathbf{S}^\lambda = \begin{pmatrix} \lambda_1\mathbb{S}^1 & . & . & . & . \\ . & \lambda_2\mathbb{S}^2 & . & . & . \\ . & . & \lambda_j\mathbb{S}^j & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix}.$$

That is, there are some blocks with single smoothing parameters, and others with a more complicated additive structure. There are usually also some zero blocks on the diagonal. The block structure means that the generalized determinant, its derivatives with respect to $\rho_k = \log\lambda_k$ and the matrix square root of $\mathbf{S}^\lambda$ can all be computed blockwise. So for the above example,

$$\log|\mathbf{S}^\lambda|_+ = \mathrm{rank}(\mathbb{S}^1)\log(\lambda_1) + \log|\mathbb{S}^1|_+ + \mathrm{rank}(\mathbb{S}^2)\log(\lambda_2) + \log|\mathbb{S}^2|_+ + \log|\lambda_j\mathbb{S}^j|_+ + \cdots$$

For any ρ_k relating to a single parameter block we have

$$\frac{\partial\log|\mathbf{S}^\lambda|_+}{\partial\rho_k} = \mathrm{rank}(\mathbb{S}^k)$$

and zero second derivatives. For multi-λ blocks there will generally be first and second derivatives to compute. There are no second derivatives "between-blocks."

In general, there are three block types, each requiring different preprocessing.

1. Single parameter diagonal blocks. A reparameterization can be used so that all nonzero elements are one, and the rank precomputed.
2. Single parameter dense blocks. An orthogonal reparameterization, based on the eigenvectors of the symmetric eigen-decomposition of the block, can be used to make these blocks look like the previous type (by similarity transform). Again the rank is computed.
3. Multi-λ blocks will require the reparameterization method of Wood (2011) appendix B to be applied for

each new $\boldsymbol{\rho}$ proposal, since the numerical problem that the reparameterization avoids is $\boldsymbol{\rho}$ dependent (see online SA C). Initially, before the smoothing parameter selection iteration, it is necessary to reparameterize to separate the parameters corresponding to the block into penalized and unpenalized subvectors. This initial reparameterization can be based on the eigenvectors of the symmetric eigen decomposition of the "balanced" version of the block penalty matrix, $\sum_j \mathbb{S}^j/\|\mathbb{S}^j\|_F$, where $\|\cdot\|_F$ is the Frobenious norm. The balanced penalty is used for maximal numerical stability, and is usable because formally the spaces for the penalized and unpenalized components do not change with the smoothing parameters.

The reparameterizations from each block type are applied to the model, usually to the model matrices $\mathbf{X}^j$ of the individual smooth terms. The reparameterization information must be stored so that we can return to the original parameterization at the end.

After the one off initial reparameterization just described, then step one of the inner algorithm requires only that the reparameterization method of Wood (2011) Appendix B be applied to the parameters corresponding to type 3 blocks, for each new set of smoothing parameters.

3.1.2. Newton Iteration for $\hat{\boldsymbol{\beta}}$

This section provides details for inner Steps 2–4. Newton iteration for $\hat{\boldsymbol{\beta}}$ requires the gradient vector, $\boldsymbol{\mathcal{G}}$, with elements $\mathcal{L}^i_\beta = l^i_\beta - \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}\beta_j$ and negative Hessian matrix $\boldsymbol{\mathcal{H}}$ with elements $-\mathcal{L}^{i\ j}_{\beta\beta} = -l^{i\ j}_{\beta\beta} + \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}$ (we will also use $\mathbf{H}$ to denote the Hessian of the negative unpenalized log-likelihood with elements $-l^{i\ j}_{\beta\beta}$). In principle Newton iteration proceeds by repeatedly setting $\boldsymbol{\beta}$ to $\boldsymbol{\beta} + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} = \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\mathcal{G}}$. In practice, Newton's method is only guaranteed to converge to a maximum of $\mathcal{L}$, provided (i) that the Hessian is perturbed to be positive definite if it is not, guaranteeing that the Newton direction is an ascent direction, (ii) that step reduction is used to ensure that the step taken actually increases $\mathcal{L}$ and (iii) that the computation of the step is numerically stable (see Nocedal and Wright 2006).

$\mathcal{L}$ may be indefinite away from a maximum, but even near the maximum there are two basic impediments to stability and positive definiteness. First, some elements of $\boldsymbol{\beta}$ may be unidentifiable. This issue will be dealt with by dropping parameters at convergence, as described shortly. The second issue is that some smoothing parameters may legitimately become very large during fitting, resulting in very large $\lambda_j \mathbf{S}^j$ components, poor scaling, poor conditioning and, hence, computational singularity. However, given the initial and Step 1 reparameterizations, such large elements can be dealt with by diagonal preconditioning of $\boldsymbol{\mathcal{H}}$. That is define diagonal matrix $\mathbf{D}$ such that $D_{ii} = |\mathcal{H}_{ii}|^{-1/2}$, and preconditioned Hessian $\boldsymbol{\mathcal{H}}' = \mathbf{D}\boldsymbol{\mathcal{H}}\mathbf{D}$. Then $\boldsymbol{\mathcal{H}}^{-1} = \mathbf{D}\boldsymbol{\mathcal{H}}'^{-1}\mathbf{D}$, with the right-hand side resulting in much better scaled computation. In the work reported here the pivoted Cholesky decomposition of the perturbed Hessian $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \boldsymbol{\mathcal{H}}' + \epsilon\mathbf{I}$ is repeated with increasing ϵ, starting from zero, until positive definiteness is obtained. The Newton step is then computed as $\boldsymbol{\Delta} = \mathbf{D}\mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\mathbf{D}\boldsymbol{\mathcal{G}}$. If the step to $\boldsymbol{\beta} + \boldsymbol{\Delta}$ fails to increase the likelihood then $\boldsymbol{\Delta}$ is repeatedly halved until it does. Note that the perturbation of the Hessian does not change the converged state of a Newton algorithm (although varying the perturbation strength can change the algorithm convergence rate).

At convergence $\boldsymbol{\mathcal{H}}$ can at worst be positive semi- definite, but it is necessary to test for the possibility that some parameters are unidentifiable. The test should not depend on the particular values of the smoothing parameters. This can be achieved by constructing the balanced penalty $\mathbf{S} = \sum_j \mathbf{S}^j/\|\mathbf{S}^j\|_F$ ($\|\cdot\|_F$ is the Frobenius norm, but another norm could equally well be used), and then forming the pivoted Cholesky decomposition $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{H}/\|\mathbf{H}\|_F + \mathbf{S}/\|\mathbf{S}\|_F$. The rank of $\mathbf{P}$ can then be estimated by making use of Cline et al. (1979). If this reveals rank deficiency of order q then the coefficients corresponding to the matrix rows and columns pivoted to the last q positions should be dropped from the analysis. The balanced penalty is used to avoid dropping parameters simply because some smoothing parameters are very large. Given the nonlinear setting it is necessary to repeat the Newton iteration to convergence with the reduced parameter set, in order that the remaining parameters adjust to the omission of those dropped.

3.1.3. Implicit Differentiation

This section provides the details for inner Steps 5–7. We obtain the derivatives of the identifiable elements of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\rho}$. All computations here are in the reduced parameter space, if parameters were dropped. At the maximum penalized likelihood estimate we have $\mathcal{L}^i_{\hat\beta} = l^i_{\hat\beta} - \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}\hat\beta_j = 0$ and differentiating with respect to $\rho_k = \log\lambda_k$ yields

$$\mathcal{L}^{i\ k}_{\hat\beta\rho} = l^{i\ j}_{\hat\beta\hat\beta}\frac{\mathrm{d}\hat\beta_j}{\mathrm{d}\rho_k} - \lambda_k S^k_{ij}\hat\beta_j - \lambda_{\boldsymbol{l}} S^{\boldsymbol{l}}_{ij}\frac{\mathrm{d}\hat\beta_j}{\mathrm{d}\rho_k} = 0 \text{ and rearranging,}$$

$$\frac{\mathrm{d}\hat\beta_i}{\mathrm{d}\rho_k} = \mathcal{L}^{\hat\beta\hat\beta}_{i\ j}\lambda_k S^k_{j\boldsymbol{l}}\hat\beta_{\boldsymbol{l}},$$

given which we can compute $l^{i\ j\ l}_{\hat\beta\hat\beta\rho} = l^{i\ j\ \boldsymbol{k}}_{\hat\beta\hat\beta\hat\beta}\mathrm{d}\hat\beta_{\boldsymbol{k}}/\mathrm{d}\rho_l$ from the model specification. $-l^{i\ j\ l}_{\hat\beta\hat\beta\rho} + \delta^l_{\boldsymbol{k}}\lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}$ are the elements of $\partial\boldsymbol{\mathcal{H}}/\partial\rho_l$, required in the next section (δ^l_k is 1 for $l = k$ and 0 otherwise). Then

$$\frac{\mathrm{d}^2\hat\beta_i}{\mathrm{d}\rho_k\mathrm{d}\rho_l} = \mathcal{L}^{\hat\beta\hat\beta}_{i\ \boldsymbol{j}}\left\{\left(-l^{\boldsymbol{j}\ \boldsymbol{p}\ l}_{\hat\beta\hat\beta\rho} + \lambda_l S^l_{\boldsymbol{jp}}\right)\frac{\mathrm{d}\hat\beta_{\boldsymbol{p}}}{\mathrm{d}\rho_k} + \lambda_k S^k_{\boldsymbol{jp}}\frac{\mathrm{d}\hat\beta_{\boldsymbol{p}}}{\mathrm{d}\rho_l}\right\} + \delta^l_k\frac{\mathrm{d}\hat\beta_i}{\mathrm{d}\rho_k},$$

which enables computations involving $\partial^2\boldsymbol{\mathcal{H}}/\partial\rho_k\partial\rho_l$, with elements $-l^{i\ j\ kl}_{\hat\beta\hat\beta\rho\rho} + \delta^l_{\boldsymbol{k}}\lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}$, and

$$l^{i\ j\ kl}_{\hat\beta\hat\beta\rho\rho} = l^{i\ j\ r\ t}_{\hat\beta\hat\beta\hat\beta\hat\beta}\frac{\mathrm{d}\hat\beta_{\boldsymbol{r}}}{\mathrm{d}\rho_k}\frac{\mathrm{d}\hat\beta_{\boldsymbol{t}}}{\mathrm{d}\rho_l} + \mathcal{L}^{i\ j\ r}_{\hat\beta\hat\beta\hat\beta}\frac{\mathrm{d}^2\hat\beta_{\boldsymbol{r}}}{\mathrm{d}\rho_k\mathrm{d}\rho_l}.$$

As mentioned in Section 3.1, it will generally be inefficient to form this last quantity explicitly, as it occurs only in the summations involved in computing the final trace in (3).

3.1.4. The Remaining Derivatives

Recalling that $\mathcal{H}$ is the matrix with elements $-\mathcal{L}^{i\ j}_{\beta\beta} = -l^{i\ j}_{\beta\beta} + \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{ij}$, we require (inner Step 8)

$$\frac{\partial \mathcal{V}}{\partial \rho_k} = -\frac{\lambda_k}{2}\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}^k\hat{\boldsymbol{\beta}} + \frac{1}{2}\frac{\partial \log|\mathbf{S}^\lambda|_+}{\partial \rho_k} - \frac{1}{2}\frac{\partial \log|\mathcal{H}|}{\partial \rho_k}$$

and

$$\frac{\partial^2 \mathcal{V}}{\partial \rho_k \partial \rho_l} = -\delta^l_k\frac{\lambda_k}{2}\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}^k\hat{\boldsymbol{\beta}} - \frac{\mathrm{d}\hat{\boldsymbol{\beta}}^{\mathsf{T}}}{\mathrm{d}\rho_l}\mathcal{H}\frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\rho_k} + \frac{1}{2}\frac{\partial^2 \log|\mathbf{S}^\lambda|_+}{\partial \rho_k \partial \rho_l} - \frac{1}{2}\frac{\partial^2 \log|\mathcal{H}|}{\partial \rho_k \partial \rho_l},$$

where components involving $\mathcal{L}^j_{\hat{\beta}}$ are zero by definition of $\hat{\boldsymbol{\beta}}$. The components not covered so far are

$$\frac{\partial \log|\mathcal{H}|}{\partial \rho_k} = \mathrm{tr}\left(\mathcal{H}^{-1}\frac{\partial \mathcal{H}}{\partial \rho_k}\right) \text{ and}$$
$$\frac{\partial^2 \log|\mathcal{H}|}{\partial \rho_k \partial \rho_l} = -\mathrm{tr}\left(\mathcal{H}^{-1}\frac{\partial \mathcal{H}}{\partial \rho_k}\mathcal{H}^{-1}\frac{\partial \mathcal{H}}{\partial \rho_l}\right) + \mathrm{tr}\left(\mathcal{H}^{-1}\frac{\partial^2 \mathcal{H}}{\partial \rho_k \partial \rho_l}\right). \quad (3)$$

The final term above is expensive if computed naively by explicitly computing each term $\partial^2\mathcal{H}/\partial\rho_k\partial\rho_l$, but this is unnecessary and the computation of $\mathrm{tr}\left(\mathcal{H}^{-1}\partial^2\mathcal{H}/\partial\rho_k\partial\rho_l\right)$ can usually be performed efficiently as the final part of the model specification, keeping the total cost to $O(Mnp^2)$: see online SA G and Section 3.2 for illustrative examples.

The Cox (1972) proportional hazards model provides a straightforward application of the general method, and the requisite computations are set out in online SA G in a manner that maintains $O(Mnp^2)$ computational cost. Another example is the multivariate additive model, in which the means of a multivariate Gaussian response are given by separate linear predictors, which may optionally share terms. This model is covered in the online SA H and Section 8. Section 3.2 considers how another class of models falls into the general framework.

3.2. A Special Case: GAMLSS Models

The GAMLSS (or "distributional regression") models discussed by Rigby and Stasinopoulos (2005) (and also Yee and Wild 1996; Klein et al. 2014, 2015) fall within the scope of the general method. The idea is that we have independent univariate response observations, y_i, whose distributions depend on several unknown parameters, each of which is determined by its own linear predictor. The log-likelihood is a straightforward sum of contributions from each y_i (unlike the Cox models, e.g.), and the special structure can be exploited so that implementation of new models in this class requires only the supply of some derivatives of the log-likelihood terms with respect to the distribution parameters. Given the notational conventions established previously, the expressions facilitating this are rather compact (without such a notation they can easily become intractably complex).

Let the log-likelihood for the ith observation be $l(y_i, \eta^1_i, \eta^2_i, \ldots)$ where the $\eta^k = \mathbf{X}^k\boldsymbol{\beta}^k$ are K linear predictors. The Newton iteration for estimating $\boldsymbol{\beta} = (\boldsymbol{\beta}^{1\mathsf{T}}, \boldsymbol{\beta}^{2\mathsf{T}}, \ldots)^{\mathsf{T}}$ requires $l^j_{\beta^l} = l^{\boldsymbol{i}}_{\eta^l}X^l_{\boldsymbol{ij}}$ and $l^{j\ k}_{\beta^l\beta^m} = l^{\boldsymbol{i}\ \boldsymbol{i}}_{\eta^l\eta^m}X^l_{\boldsymbol{ij}}X^m_{\boldsymbol{ik}}$, which are also sufficient for first-order implicit differentiation.

LAML optimization also requires

$$l^{j\ k\ p}_{\hat{\beta}^l\hat{\beta}^m\rho} = l^{j\ k\ \boldsymbol{r}}_{\hat{\beta}^l\hat{\beta}^m\hat{\beta}^{\boldsymbol{q}}}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p} = l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}}X^l_{\boldsymbol{ij}}X^m_{\boldsymbol{ik}}X^{\boldsymbol{q}}_{\boldsymbol{ir}}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p}$$
$$= l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}}X^l_{\boldsymbol{ij}}X^m_{\boldsymbol{ik}}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p}.$$

Notice how this is just an inner product $\mathbf{X}^{\mathsf{T}}\mathbf{V}\mathbf{X}$, where the diagonal matrix $\mathbf{V}$ is the sum over q of some diagonal matrices. At this stage the second derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\rho}$ can be computed, after which we require only

$$l^{j\ k\ p\ v}_{\hat{\beta}^l\hat{\beta}^m\rho\rho} = l^{j\ k\ \boldsymbol{r}\ \boldsymbol{t}}_{\hat{\beta}^l\hat{\beta}^m\hat{\beta}^{\boldsymbol{q}}\hat{\beta}^{\boldsymbol{s}}}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{s}}_{\boldsymbol{t}}}{\mathrm{d}\rho_v} + l^{j\ k\ \boldsymbol{r}}_{\hat{\beta}^l\hat{\beta}^m\hat{\beta}^{\boldsymbol{q}}}\frac{\mathrm{d}^2\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p\mathrm{d}\rho_v}$$
$$= l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}\hat{\eta}^{\boldsymbol{s}}}X^l_{\boldsymbol{ij}}X^m_{\boldsymbol{ik}}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{s}}_{\boldsymbol{i}}}{\mathrm{d}\rho_v} + l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}}X^l_{\boldsymbol{ij}}X^m_{\boldsymbol{ik}}\frac{\mathrm{d}^2\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p\mathrm{d}\rho_v}.$$

So to implement a new family for GAMLSS estimation requires mixed derivatives up to fourth order with respect to the parameters of the likelihood. In most cases what would be conveniently available is, for example, $l^{i\ i\ i\ i}_{\hat{\mu}^l\hat{\mu}^m\hat{\mu}^q\hat{\mu}^s}$ rather than $l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}\hat{\eta}^{\boldsymbol{s}}}$, where μ^k is the kth parameter of the likelihood and is given by $h^k(\mu^k) = \eta^k$, h^k being a link function.

To get from the μ derivatives to the η derivatives, the rules (A.1)–(A.4) from Appendix A are used. This is straightforward for any derivative that is not mixed. For mixed derivatives containing at least one first-order derivative the transformation rule applying to the highest order derivative is applied first, followed by the transformations for the first-order derivatives. This leaves only the transformation of $l^{i\ i\ i\ i\ i}_{\hat{\mu}^j\hat{\mu}^j\hat{\mu}^k\hat{\mu}^k\hat{\mu}^k}$ as at all awkward, but we have

$$l^{i\ i\ i\ i\ i}_{\hat{\eta}^j\hat{\eta}^j\hat{\eta}^k\hat{\eta}^k\hat{\eta}^k} = (l^{i\ i\ i\ i\ i}_{\hat{\mu}^j\hat{\mu}^j\hat{\mu}^k\hat{\mu}^k\hat{\mu}^k}/h^{j\prime 2}_i - l^{i\ i\ i\ i}_{\hat{\mu}^j\hat{\mu}^k\hat{\mu}^k\hat{\mu}^k}h^{j\prime\prime}_i/h^{j\prime 3}_i)/$$
$$h^{k\prime 2}_i - (l^{i\ i\ i\ i}_{\hat{\mu}^j\hat{\mu}^j\hat{\mu}^k\hat{\mu}^k}/h^{j\prime 2}_i - l^{i\ i\ i}_{\hat{\mu}^j\hat{\mu}^k\hat{\mu}^k}h^{j\prime\prime}_i/h^{j\prime 3}_i)h^{k\prime\prime}_i/h^{k\prime 3}_i.$$

The general method requires $\mathcal{L}^{\hat{\beta}\hat{\beta}}_{\boldsymbol{k}\ \boldsymbol{j}}l^{\boldsymbol{j}\ \boldsymbol{k}\ p\ v}_{\hat{\beta}\hat{\beta}\rho\ \rho}$ to be computed, which would have $O\{M(M+1)n\mathcal{P}^2/2\}$ cost if the terms $l^{\boldsymbol{j}\ \boldsymbol{k}\ p\ v}_{\hat{\beta}\ \hat{\beta}\ \rho\ \rho}$ were computed explicitly for this purpose (where $\mathcal{P}$ is the dimension of combined $\boldsymbol{\beta}$). However, this can be reduced to $O(n\mathcal{P}^2)$ using a trick most easily explained by switching to a matrix representation. For simplicity of presentation assume $K = 2$, and define matrix $\mathbf{B}$ to be the inverse of the penalized Hessian, so that $B_{ij} = \mathcal{L}^{\hat{\beta}\hat{\beta}}_{i\ j}$. Defining

$$v^{lm}_i = l^{i\ i\ i\ i}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}\hat{\eta}^{\boldsymbol{s}}}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{s}}_{\boldsymbol{i}}}{\mathrm{d}\rho_v} + l^{\boldsymbol{i}\ \boldsymbol{i}\ \boldsymbol{i}}_{\hat{\eta}^l\hat{\eta}^m\hat{\eta}^{\boldsymbol{q}}}\frac{\mathrm{d}^2\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p\mathrm{d}\rho_v} \text{ and}$$

$\mathbf{V}^{lm} = \mathrm{diag}(v^{lm}_i)$ we have

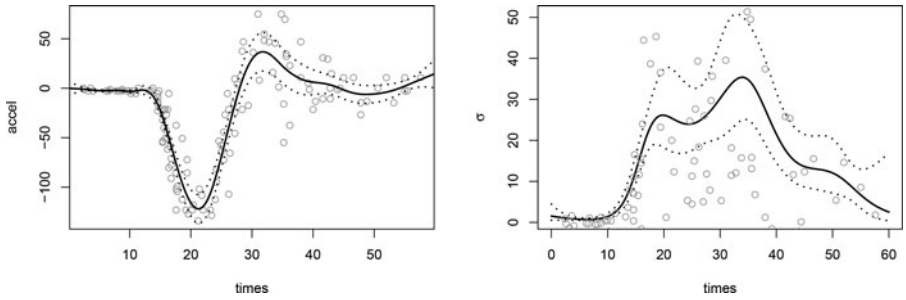

Figure 2. A smooth Gaussian location scale model fit to the motorcycle data from Silverman (1985), using the methods developed in Section 3.2. The left plot shows the raw data as open circles and an adaptive p-spline smoother for the mean overlaid. The right plot shows the simultaneous estimate of the standard deviation in the acceleration measurements, with the absolute values of the residuals as circles. Dotted curves are approximate 95% confidence intervals. The effective degrees of freedom of the smooths are 12.5 and 7.3 respectively.

$$\mathcal{L}^{\hat\beta\hat\beta}_{\boldsymbol{kj}} l^{\boldsymbol{j}\,\boldsymbol{k}\;p\;v}_{\hat\beta\hat\beta\;\rho\;\rho} = \mathrm{tr}\left\{\mathbf{B}\begin{pmatrix}\mathbf{X}^{1\mathsf{T}}\mathbf{V}^{11}\mathbf{X}^1 & \mathbf{X}^{1\mathsf{T}}\mathbf{V}^{12}\mathbf{X}^2\\ \mathbf{X}^{2\mathsf{T}}\mathbf{V}^{12}\mathbf{X}^1 & \mathbf{X}^{2\mathsf{T}}\mathbf{V}^{22}\mathbf{X}^2\end{pmatrix}\right\}$$
$$= \mathrm{tr}\left\{\mathbf{B}\begin{pmatrix}\mathbf{X}^1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^2\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\mathbf{V}^{11}\mathbf{X}^1 & \mathbf{V}^{12}\mathbf{X}^2\\ \mathbf{V}^{12}\mathbf{X}^1 & \mathbf{V}^{22}\mathbf{X}^2\end{pmatrix}\right\}. \quad (4)$$

Hence, following the one off formation of $\mathbf{B}\begin{pmatrix}\mathbf{X}^1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^2\end{pmatrix}^{\mathsf{T}}$ (which need only have $O(n\mathcal{P}^2)$ cost), each trace computation has $O(Mn\mathcal{P})$ cost (since $\mathrm{tr}(\mathbf{C}^{\mathsf{T}}\mathbf{D}) = D_{ij}C_{ij}$).

See online SA I where a zero inflated Poisson model provides an example of the details. Figure 2 shows estimates for the model $\texttt{accel}_i \sim N(f_1(t_i), \sigma_i^2)$ where $\log\sigma_i = f_2(t_i)$, f_1 is an adaptive P-spline and f_2 a cubic regression spline, while SA F.2 provides another application. Package `mgcv` also includes multinomial logistic regression implemented this way and further examples are under development. An interesting possibility with any model which has multiple linear predictors is that one or more of those predictors should depend on some of the same terms, and online SA H shows how this can be handled.

3.3. A More Special Case: Extended Generalized Additive Models

For models with a single linear predictor, in which the log-likelihood is a sum of contributions per y_i, it is possible to perform fitting by iterative weighted least squares, enabling profitable reuse of some components of standard GAM fitting methods, including the exploitation of very stable orthogonal methods for solving least squares problems. Specifically, consider observations y_i, and let the corresponding log-likelihood be of the form

$$l = \sum_i l_i(y_i, \mu_i, \boldsymbol{\theta}, \phi),$$

where the terms in the summation may also be written as l_i for short, and μ_i is often $\mathbb{E}(y_i)$, but may also be a latent variable (as in the ordered categorical model of SA K). Given h, a known link function, $h(\mu_i) = \eta_i$ where $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{o}$, $\mathbf{X}$ is a model matrix, $\boldsymbol{\beta}$ is a parameter vector and $\mathbf{o}$ is an offset (often simply 0). $\boldsymbol{\theta}$ is a parameter vector, containing the extra parameters of the likelihood, such as the p parameter of a Tweedie density (see online SA J), or the cut points of an ordered categorical model (see online SA K). Notice that in this case $\boldsymbol{\theta}$ is not treated as part of $\boldsymbol{\beta}$, since $\boldsymbol{\theta}$ can not always be estimated by straightforward iterative regression. Instead $\boldsymbol{\theta}$ will be estimated alongside the smoothing parameters. ϕ is a scale parameter, often fixed at one. Let $\tilde{l}_i = \max_{\mu_i} l_i(y_i, \mu_i, \boldsymbol{\theta}, \phi)$ denote the saturated log-likelihood. Define the *deviance* corresponding to y_i as $D_i = 2(\tilde{l}_i - l_i)\phi$, where ϕ is the scale parameter on which D_i does not depend. Working in terms of the deviance is convenient in a regression setting, where deviance residuals are a preferred method for model checking and the proportion deviance explained is a natural substitute for the r^2 statistic as a measure of goodness of fit (but see the final comment in online SA I).

In general the estimates of $\boldsymbol{\beta}$ will depend on some log smoothing parameter $\rho_j = \log\lambda_j$, and it is notationally expedient to consider these to be part of the vector $\boldsymbol{\theta}$, although it is to be understood that l does not actually depend on these elements of $\boldsymbol{\theta}$. Given $\boldsymbol{\theta}$, estimation of $\boldsymbol{\beta}$ is by minimization of the penalized deviance $\mathcal{D}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_i D_i(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_j \lambda_j \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^j\boldsymbol{\beta}$, with respect to $\boldsymbol{\beta}$. This can be achieved by penalized iteratively reweighted least squares (PIRLS), which consists of iterative minimization of $\sum_i w_i(z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \sum_j \lambda_j\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}^j\boldsymbol{\beta}$, where the pseudodata and weights are given by

$$z_i = \eta_i - o_i - \frac{1}{2w_i}\frac{\partial D_i}{\partial \eta_i}, \quad w_i = \frac{1}{2}\frac{\mathrm{d}^2 D_i}{\mathrm{d}\eta_i^2}.$$

Note that if $w_i = 0$ (or w_i is too close to 0), the penalized least squares estimate can be computed using only $w_i z_i$, which is then well defined and finite when z_i is not.

Estimation of $\boldsymbol{\theta}$, and possibly ϕ, is by LAML. Writing $\mathbf{W}$ as the diagonal matrix of w_i values, the log LAML is given by

$$\mathcal{V}(\boldsymbol{\theta}, \phi) = -\frac{\mathcal{D}(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta})}{2\phi} + \tilde{l}(\boldsymbol{\theta}, \phi) - \frac{\log|\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S}^\lambda| - \log|\mathbf{S}^\lambda|_+}{2}$$
$$+ \frac{M_p}{2}\log(2\pi\phi),$$

where $\mathbf{W}$ is evaluated at the $\hat{\boldsymbol{\beta}}$ implied by $\boldsymbol{\theta}$. To compute the derivatives of $\mathcal{V}$ with respect to $\boldsymbol{\theta}$ the derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\theta}$ are required. Note that $\mathcal{V}$ is a full Laplace approximation, rather than the "approximate" Laplace approximation used to justify PQL (Breslow and Clayton 1993), so that PQL's well known problems with binary and low count data are much reduced. In particular: (i) most PQL implementations estimate ϕ when fitting the working linear mixed model, even in the binomial and Poisson cases, where it is fixed at 1. For binary

and low count data this can give very poor results. (ii) PQL uses the expected Hessian rather than the Hessian, and these only coincide for the canonical link case. (iii) PQL is justified by an assumption that the iterative fitting weights only vary slowly with the smoothing parameters, an assumption that is not needed here.

The parameters $\boldsymbol{\theta}$ and ϕ can be estimated by maximizing $\mathcal{V}$ using Newton's method, or a quasi-Newton method. Notice that $\mathcal{V}$ depends directly on the elements of $\boldsymbol{\theta}$ via $\mathcal{D}$, $\tilde{l}$ and $\mathbf{S}^\lambda$, but also indirectly via the dependence of $\hat{\boldsymbol{\mu}}$ and $\mathbf{W}$ on $\hat{\boldsymbol{\beta}}$ and hence on $\boldsymbol{\theta}$. Hence, each trial $\boldsymbol{\theta}$, ϕ requires a PIRLS iteration to find the corresponding $\hat{\boldsymbol{\beta}}$, followed by implicit differentiation to find the derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\theta}$. Once these are obtained, the chain rule can be applied to find the derivatives of $\mathcal{V}$ with respect to $\boldsymbol{\theta}$ and ϕ.

As illustrated in SA C, there is scope for serious numerical instability in the evaluation of the determinant terms in $\mathcal{V}$, but for this case we can reuse the stabilization strategy from Wood (2011), namely for each trial $\boldsymbol{\theta}$ and ϕ:

1. Use the orthogonal reparameterization from Appendix B of Wood (2011) to ensure that $\log|\mathbf{S}^\lambda|_+$ can be computed in a stable manner.
2. Estimate $\hat{\boldsymbol{\beta}}$ by PIRLS using the stable least squares method for negatively weighted problems from Section 3.3 of Wood (2011), setting structurally unidentifiable coefficients to zero.
3. Using implicit differentiation, obtain the derivatives of $\mathcal{V}$ required for a Newton update.

Step 3 is substantially more complicated than in Wood (2011), and is covered in Appendix A.

3.3.1. Extended GAM New Model Implementation

The general formulation above assumes that various standard information is available for each distribution and link. What is needed depends on whether quasi-Newton or full Newton is used to find $\hat{\boldsymbol{\theta}}$. Here is a summary of what is needed for each distribution

1. For finding $\hat{\boldsymbol{\beta}}$. D^i_μ, $D^{i\ i}_{\mu\mu}$, h', and h''.
2. For $\hat{\boldsymbol{\rho}}$ via quasi-Newton. h''', $D^{i\ j}_{\mu\theta}$, D^j_θ, $D^{i\ i\ i}_{\mu\mu\mu}$, and $D^{i\ i\ j}_{\mu\mu\theta}$.
3. For $\hat{\boldsymbol{\rho}}$ via full Newton. h'''', $D^{i\ j}_{\theta\theta}$, $D^{i\ jk}_{\mu\theta\theta}$, $D^{i\ i\ i\ i}_{\mu\mu\mu\mu}$, $D^{i\ i\ i\ j}_{\mu\mu\mu\theta}$, and $D^{i\ i\ jk}_{\mu\mu\theta\theta}$.

In addition, first and second derivatives of $\tilde{l}$ with respect to its arguments are needed. All of these quantities can be obtained automatically using a computer algebra package. $\mathbb{E}D^{i\ i}_{\mu\mu}$ is also useful for further inference. If it is not readily computed then we can substitute $D^{i\ i}_{\mu\mu}$, but a complication of penalized modeling is that $D^{i\ i}_{\mu\mu}$ can fail to be positive definite at $\hat{\boldsymbol{\beta}}$. When this happens $\mathbb{E}D^{i\ i}_{\mu\mu}$ can be estimated as the nearest positive definite matrix to $D^{i\ i}_{\mu\mu}$.

We have implemented beta, negative binomial, scaled t models for heavy tailed data, simple zero inflated Poisson, ordered categorical and Tweedie additive models in this way. The first three were essentially automatic: the derivatives were computed by a symbolic algebra package and coded from the results. Some care is required in doing this, to avoid excessive cancellation error, underflow or overflow in the computations. Overly naive coding of derivatives can often lead to numerical problems: The online SA I on the zero inflated Poisson provides an example of the sort of issues that can be encountered. The ordered categorical and Tweedie models are slightly more complicated and details are therefore provided in the online SA J and K (including further examples of the need to avoid cancellation error).

4. Smoothing Parameter Uncertainty

Conventionally in a GAM context smoothing parameters have been treated as fixed when computing interval estimates for functions, or for other inferential tasks. In reality smoothing parameters must be estimated, and the uncertainty associated with this has generally been ignored except in fully Bayesian simulation approaches. Kass and Steffey (1989) proposed a simple first-order correction for this sort of uncertainty in the context of iid Gaussian random effects in a one way ANOVA type design. Some extra work is required to understand how their method works when applied to smooths. It turns out that the estimation methods described above provide the quantities required to correct for smoothing parameter uncertainty.

Assume we have several smooth model components, let $\rho_i = \log\lambda_i$ and $\mathbf{S}^\lambda = \sum_j \lambda_j \mathbf{S}^j$. Writing $\hat{\boldsymbol{\beta}}_\rho$ for $\hat{\boldsymbol{\beta}}$, to emphasize the dependence of $\hat{\boldsymbol{\beta}}$ on the smoothing parameters, we use the Bayesian large sample approximation (see SB.4)

$$\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\rho} \sim N(\hat{\boldsymbol{\beta}}_\rho, \mathbf{V}_\beta) \text{ where } \mathbf{V}_\beta = (\hat{\mathcal{I}} + \mathbf{S}^\lambda)^{-1} \tag{5}$$

which is exact in the Gaussian case, along with the large sample approximation

$$\boldsymbol{\rho}|\mathbf{y} \sim N(\hat{\boldsymbol{\rho}}, \mathbf{V}_\rho), \tag{6}$$

where $\mathbf{V}_\rho$ is the inverse of the Hessian of the negative log marginal likelihood with respect to $\boldsymbol{\rho}$. Since the approximation (6) applies in the interior of the parameter space, it is necessary to substitute a Moore-Penrose pseudoinverse of the Hessian if a smoothing parameter is effectively infinite, or otherwise to regularize the inversion (which is equivalent to placing a Gaussian prior on $\boldsymbol{\rho}$). Conventionally (5) is used with $\hat{\boldsymbol{\rho}}$ plugged in and the uncertainty in $\boldsymbol{\rho}$ neglected. To improve on this note that if (5) and (6) are correct, while $\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$ and independently $\boldsymbol{\rho}^* \sim N(\hat{\boldsymbol{\rho}}, \mathbf{V}_\rho)$, then $\boldsymbol{\beta}|\mathbf{y} \overset{d}{=} \hat{\boldsymbol{\beta}}_{\rho^*} + \mathbf{R}^\mathsf{T}_{\rho^*}\mathbf{z}$ where $\mathbf{R}^\mathsf{T}_{\rho^*}\mathbf{R}_{\rho^*} = \mathbf{V}_\beta$ (and $\mathbf{V}_\beta$ depends on $\boldsymbol{\rho}^*$). This provides a way of simulating from $\boldsymbol{\beta}|\mathbf{y}$, but it is computationally expensive as $\hat{\boldsymbol{\beta}}_{\rho^*}$ and $\mathbf{R}_{\rho^*}$ must be computed afresh for each sample. (The conventional approximation would simply set $\boldsymbol{\rho}^* = \hat{\boldsymbol{\rho}}$.) Alternatively consider a first-order Taylor expansion

$$\boldsymbol{\beta}|\mathbf{y} \overset{d}{=} \hat{\boldsymbol{\beta}}_{\hat\rho} + \mathbf{J}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}) + \mathbf{R}^\mathsf{T}_{\hat\rho}\mathbf{z} + \sum_k \left.\frac{\partial \mathbf{R}^\mathsf{T}_\rho \mathbf{z}}{\partial \rho_k}\right|_{\hat{\boldsymbol{\rho}}} (\rho_k - \hat\rho_k) + r,$$

where r is a lower order remainder term and $\mathbf{J} = \mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\boldsymbol{\rho}|_{\hat\rho}$. Dropping r, the expectation of the right-hand side is $\hat{\boldsymbol{\beta}}_{\hat\rho}$. Denoting the elements of $\mathbf{R}_\rho$ by R_{ij}, tedious but routine calculation shows that the three remaining random terms are uncorrelated with covariance matrix

$$\mathbf{V}'_\beta = \mathbf{V}_\beta + \mathbf{V}' + \mathbf{V}'', \text{ where } \mathbf{V}' = \mathbf{J}\mathbf{V}_\rho\mathbf{J}^\mathsf{T} \text{ and}$$

$$V''_{jm} = \sum_i^p \sum_l^M \sum_k^M \frac{\partial R_{ij}}{\partial \rho_k} V_{\rho,kl} \frac{\partial R_{im}}{\partial \rho_l}, \tag{7}$$

which is computable at $O(Mp^3)$ cost (see online SA D). Dropping $\mathbf{V}''$ we have the Kass and Steffey (1989) approximation $\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}_{\hat\rho}, \mathbf{V}^*_\beta)$ where $\mathbf{V}^*_\beta = \mathbf{V}_\beta + \mathbf{J}\mathbf{V}_\rho\mathbf{J}^\mathsf{T}$. (A first-order Taylor expansion of $\hat{\boldsymbol{\beta}}$ about $\boldsymbol{\rho}$ yields a similar correction for the frequentist covariance matrix of $\hat{\boldsymbol{\beta}}$: $\mathbf{V}^*_{\hat\beta} = (\hat{\mathcal{I}} + \mathbf{S}^\lambda)^{-1}\hat{\mathcal{I}}(\hat{\mathcal{I}} + \mathbf{S}^\lambda)^{-1} + \mathbf{J}\mathbf{V}_\rho\mathbf{J}^\mathsf{T}$, where $\hat{\mathcal{I}}$ is the negative Hessian of the log-likelihood).

The online SA D shows that in a Demmler-Reinsch like parameterization, for any penalized parameter β_i with posterior standard deviation σ_{β_i},

$$\frac{\mathrm{d}\hat\beta_i/\mathrm{d}\rho_j}{\mathrm{d}(\mathbf{R}^\mathsf{T}\mathbf{z})_i/\mathrm{d}\rho_j} \simeq \frac{\hat\beta_i}{z_i\sigma_{\beta_i}}.$$

So the $\mathbf{J}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}})$ correction is dominant for components that are strongly nonzero. This offers some justification for using the Kass and Steffey (1989) approximation, but not in a model selection context, where near zero model components are those of most interest: hence, in what follows we will use (7) without dropping $\mathbf{V}''$.

5. An Information Criterion for Smooth Model Selection

When viewing smoothing from a Bayesian perspective, the smooths have improper priors (or alternatively vague priors of convenience) corresponding to the null space of the smoothing penalties. This invalidates model selection via marginal likelihood comparison. An alternative is a frequentist AIC (Akaike 1973), based on the conditional likelihood of the model coefficients, rather than the marginal likelihood. In the exponential family GAM context, Hastie and Tibshirani (1990, §6.8.3) proposed a widely used version of this *conditional* AIC in which the effective degrees of freedom of the model, τ_0, is used in place of the number of model parameters (in the general setting $\tau_0 = \mathrm{tr}\{\mathbf{V}_\beta\hat{\mathcal{I}}\}$ is equivalent to the Hastie and Tibshirani (1990) proposal). But Greven and Kneib (2010) showed that this is overly likely to select complex models, especially when the model contains random effects: the difficulty arises because τ_0 neglects the fact that the smoothing parameters have been estimated and are, therefore, uncertain (a marginal AIC based on the frequentist marginal likelihood, in which unpenalized effects are not integrated out, is equally problematic, partly because of underestimation of variance components and consequent bias toward simple models). A heuristic alternative is to use $\tau_1 = \mathrm{tr}(2\hat{\mathcal{I}}\mathbf{V}_\beta - \hat{\mathcal{I}}\mathbf{V}_\beta\hat{\mathcal{I}}\mathbf{V}_\beta)$ as the effective degrees of freedom, motivated by considering the number of unpenalized parameters required to optimally approximate a bias corrected version of the model, but the resulting AIC is too conservative (see, Section 6, e.g.). Greven and Kneib (2010) show how to exactly compute an effective modified AIC for the Gaussian additive model case based on defining the effective degrees of freedom as $\sum_i \partial\hat{y}_i/\partial y_i$ (as proposed by Liang et al. 2008). Yu and Yau (2012) and Säfken et al. (2014) considered extensions to generalized linear mixed models. The novel contribution of this section is to use the results of the previous section to avoid the problematic neglect of smoothing parameter uncertainty in the conditional AIC computation in a manner that is easily computed and applicable to the general model class considered in this article.

The derivation of AIC (see, e.g., Davison 2003, sec. 4.7) with the MLE replaced by the penalized MLE is identical up to the point at which the AIC score is represented as

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2\mathbb{E}\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^\mathsf{T}\mathcal{I}_d(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)\right\} \tag{8}$$

$$= -2l(\hat{\boldsymbol{\beta}}) + 2\mathrm{tr}\left[\mathbb{E}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^\mathsf{T}\}\mathcal{I}_d\right], \tag{9}$$

where $\boldsymbol{\beta}_d$ is the coefficient vector minimizing the KL divergence and $\mathcal{I}_d$ is the corresponding expected negative Hessian of the log-likelihood. In an unpenalized setting $\mathbb{E}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^\mathsf{T}\}$ is estimated as the observed inverse information matrix $\hat{\mathcal{I}}^{-1}$ and $\tau' = \mathrm{tr}\{\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^\mathsf{T}\mathcal{I}_d\}$ is estimated as $\mathrm{tr}(\hat{\mathcal{I}}^{-1}\hat{\mathcal{I}}) = k$. Penalization means that the expected inverse covariance matrix of $\hat{\boldsymbol{\beta}}$ is no longer well approximated by $\hat{\mathcal{I}}$, and there are then two ways of proceeding.

The first is to view $\boldsymbol{\beta}$ as a frequentist random effect, with predicted values $\hat{\boldsymbol{\beta}}$. In that case the covariance matrix for the predictions, $\hat{\boldsymbol{\beta}}$, corresponds to the posterior covariance matrix obtained when taking the Bayesian view of the smoothing process, so we have the conventional estimate $\tau = \mathrm{tr}\{\mathbf{V}_\beta\hat{\mathcal{I}}\}$ if we neglect smoothing parameter uncertainty, or $\tau = \mathrm{tr}(\mathbf{V}'_\beta\hat{\mathcal{I}})$ accounting for it using (7).

The frequentist random effects formulation is not a completely natural way to view smooths, since we do not usually expect the smooth components of a model to be resampled from the prior with each replication of the data. However in the smoothing context $\mathbf{V}_\beta$ has the interpretation of being the frequentist covariance matrix for $\hat{\boldsymbol{\beta}}$ plus an estimate of the prior expectation of the squared smoothing bias (matrix), which offers some justification for using the same τ estimate as in the strict random effects case. To see this consider the decomposition

$$\mathbb{E}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^\mathsf{T}\} = \mathbb{E}\{(\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}})^\mathsf{T}\} + \boldsymbol{\Delta}_\beta\boldsymbol{\Delta}_\beta^\mathsf{T},$$

where $\boldsymbol{\Delta}_\beta$ is the smoothing bias in $\hat{\boldsymbol{\beta}}$. The first term on the right-hand side, above, can be replaced by the standard frequentist estimate $\mathbf{V}_{\hat\beta} = (\hat{\mathcal{I}} + \mathbf{S}^\lambda)^{-1}\hat{\mathcal{I}}(\hat{\mathcal{I}} + \mathbf{S}^\lambda)^{-1}$. Now expand the penalized log-likelihood around $\boldsymbol{\beta}_d$:

$$l_p(\boldsymbol{\beta}') \simeq l(\boldsymbol{\beta}_d) + \frac{\partial l}{\partial \boldsymbol{\beta}^\mathsf{T}}(\boldsymbol{\beta}' - \boldsymbol{\beta}_d) - \frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta}_d)^\mathsf{T}\mathcal{I}_d(\boldsymbol{\beta}' - \boldsymbol{\beta}_d)$$
$$-\frac{1}{2}\boldsymbol{\beta}'^\mathsf{T}\mathbf{S}^\lambda\boldsymbol{\beta}'.$$

Differentiating with respect to $\boldsymbol{\beta}'$ and equating to zero we obtain the approximation

$$\hat{\boldsymbol{\beta}} \simeq (\mathcal{I}_d + \mathbf{S}^\lambda)^{-1}\left(\mathcal{I}_d\boldsymbol{\beta}_d + \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\beta_d}\right).$$

$\mathbb{E}\mathrm{d}l/\mathrm{d}\boldsymbol{\beta}|_{\beta_d} = 0$ by definition of $\boldsymbol{\beta}_d$, so taking expectations of both sides we have $\mathbb{E}(\hat{\boldsymbol{\beta}}) \simeq (\boldsymbol{\mathcal{I}}_d + \mathbf{S}^\lambda)^{-1}\boldsymbol{\mathcal{I}}_d\boldsymbol{\beta}_d$. Hence estimating $\boldsymbol{\mathcal{I}}_d$ by $\hat{\boldsymbol{\mathcal{I}}}$ we have $\tilde{\boldsymbol{\Delta}}_\beta \simeq \{(\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}^\lambda)^{-1}\hat{\boldsymbol{\mathcal{I}}} - \mathbf{I}\}\boldsymbol{\beta}_d$. Considering the expected value of $\tilde{\boldsymbol{\Delta}}_\beta\tilde{\boldsymbol{\Delta}}_\beta^{\mathsf{T}}$ according to the prior mean and variance assumptions of the model, we have the following.

Lemma 1. Let the setup be as above and let $\mathbb{E}_\pi$ denote expectation assuming the prior mean and covariance for $\boldsymbol{\beta}$. Treating $\hat{\boldsymbol{\mathcal{I}}}$ as fixed, then $\mathbf{V}_{\hat{\beta}} + \mathbb{E}_\pi(\tilde{\boldsymbol{\Delta}}_\beta\tilde{\boldsymbol{\Delta}}_\beta^{\mathsf{T}}) = \mathbf{V}_\beta$.

For proof see online SA D. This offers some justification for again using $\tau = \mathrm{tr}\{\mathbf{V}_\beta\hat{\boldsymbol{\mathcal{I}}}\}$, or $\tau = \mathrm{tr}(\mathbf{V}'_\beta\hat{\boldsymbol{\mathcal{I}}})$ accounting for $\boldsymbol{\rho}$ uncertainty. So both the frequentist random effects perspective and the prior expected smoothing bias approach result in

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2\mathrm{tr}(\hat{\boldsymbol{\mathcal{I}}}\mathbf{V}'_\beta). \tag{10}$$

This is the conventional Hastie and Tibshirani (1990) conditional AIC with an additive correction $2\mathrm{tr}\{\hat{\boldsymbol{\mathcal{I}}}(\mathbf{V}' + \mathbf{V}'')\}$, accounting for smoothing parameter uncertainty. The correction is readily computed for any model considered here, provided only that the derivatives of $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}_\beta$ can be computed: the methods of Section 3 provide these. Section 6 provides an illustration of the efficacy of (10).

6. Simulation Results

The improvement resulting from using the corrected AIC of Section 5 can be illustrated by simulation. Simulations were conducted for additive models with true expected values given by $\eta = f_0(x_0) + f_1(x_1) + f_2(x_2) + f_3(x_3)$, where the f_j are shown in the online SA E, and the x covariates are all independent $U(0, 1)$ deviates. Two model comparisons were considered. In the first a 40 level Gaussian random effect was added to η, with the random effect standard deviation being varied from 0 (no effect) to 1. AIC was then used to select between models with or without the random effect included, but where all smooth terms were modeled using penalized regression splines. In the second case models with and without f_0 were compared, with the true model being based on cf_0 in place of f_0, where the effect strength c was varied from 0 (no effect) to 1. Model selection was based on (i) conventional conditional generalized AIC using τ_0 from Section 5, (ii) the corrected AIC of Section 5, (iii) a version of AIC in which the degrees of freedom penalty is based on τ_1 from Section 5, (iv) AIC based on the marginal likelihood with the number of parameters given by the number of smoothing parameters and variance components plus the number of unpenalized coefficients in the model, and (v) The Greven and Kneib (2010) corrected AIC for the Gaussian response case. The marginal likelihood in case (iv) is a version in which unpenalized coefficients are not integrated out (to avoid the usual problems with fixed effect differences and REML, or improper priors and marginal likelihood).

Results are shown in the top row of Figure 3 for a sample size of 500 with Gaussian sampling error and standard deviation of 2. For the random effect comparison, conventional conditional AIC is heavily biased toward the more complex model, selecting it on over 70% of occasions. The ML based AIC is too conservative for an AIC criterion with 3.5% selection of the larger model when it is not correct, as against the roughly 16% one might expect from AIC comparison of models differing in 1 parameter. The known underestimation of variance components estimated by this sort of marginal likelihood is partly to blame. The AIC based on τ_1 from Section 5 also lacks power, performing even less well than the ML based version. By contrast, the new corrected AIC performs well, and in this example is a slight improvement on Greven and Kneib (2010). For the smooth comparison the different calculations differ much less, although the alternatives are slightly less biased

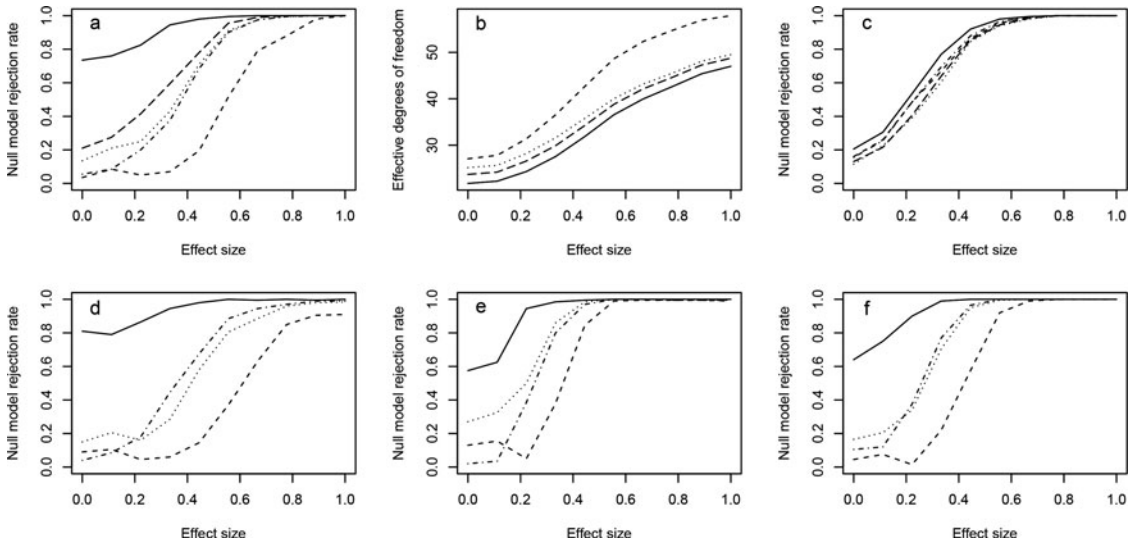

Figure 3. Simulation based illustration of the problems with previous AIC type model selection criteria and the relatively good performance of the Section 5 version. In all panels: (i) the solid curves are for conventional conditional AIC, (ii) the dotted curves are for the Section 5 version, (iii) the middle length dashed curves are for AIC based on the heuristic upper bound degrees of freedom, (iv) the dashed dot curves are for the marginal likelihood based AIC and (v) the long dashed curves are for the Greven and Kneib (2010) corrected AIC (top row only). (a) Observed probability of selecting the larger model as the effect strength of the differing term is increased from zero, for a 40 level random effect and Gaussian likelihood. (b) whole model effective degrees of freedom used in the alternative conditional AIC scores for the left hand panel as effect size increases. (c) Same as (a), but where the term differing between the two models was a smooth curve. (d) As (a) but for a Bernoulli likelihood. (e) As (a) for a beta likelihood. (f) As (a) for a Cox proportional hazards partial likelihood.

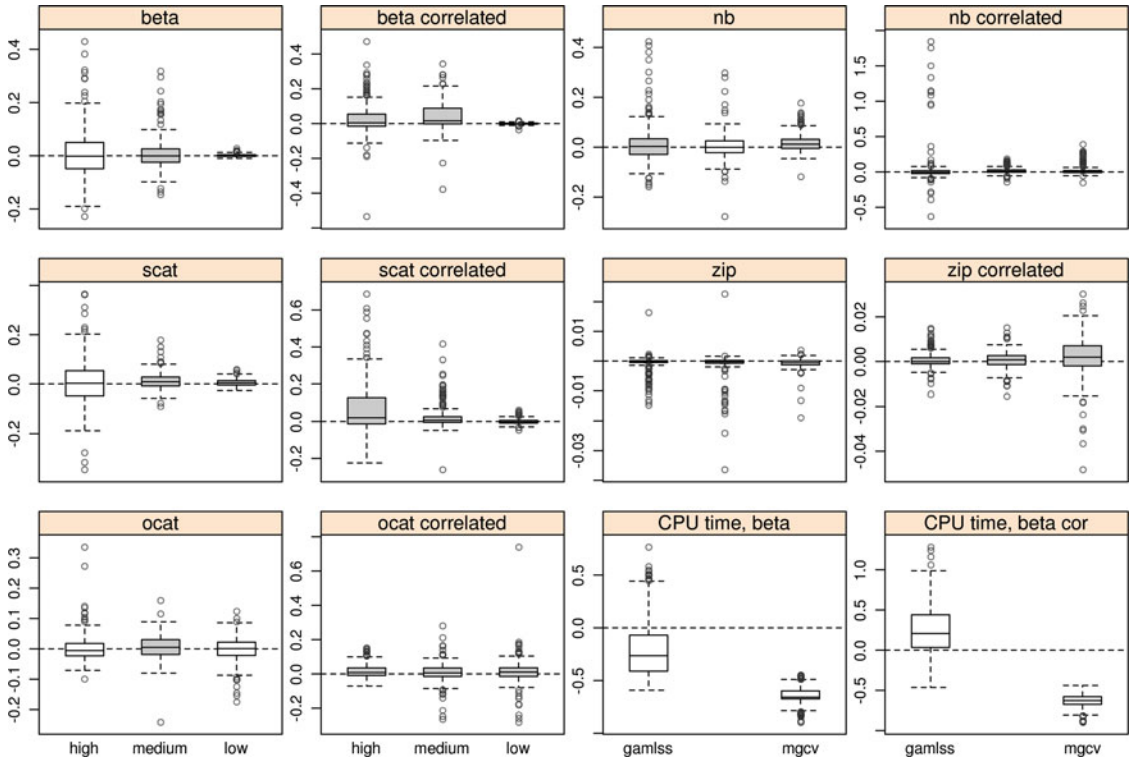

Figure 4. Results of simulation comparison with `gamlss` (beta, nb, scat, zip) and `BayesX` (ocat) packages for one dimensional P-spline models. The two plots at lower right show comparisons of $\log_{10}$ computing times for the case with the smallest time advantage for the new method — Beta regression. The remaining panels show boxplots of replicate by replicate difference in MSE/Brier's score each standardized by the average MSE or Brier's score for the particular simulation comparison. Each panel shows three box plots, one for each noise to signal level. Positive values indicate that the new method is doing better than the alternative. Boxplots are shaded grey when the difference is significant at the 5% level (all three for nb correlated should be gray). In all cases where the difference is significant at 5% the new method is better than the alternative, except for the zero inflated Poisson with uncorrelated data, where the alternative method is better at all noise levels.

toward the more complex model than the conventional conditional generalized AIC, with the corrected Section 5 version showing the smallest bias. The lower row of Figure 3 shows equivalent power plots for the same Gaussian random effect and linear predictor η, but with Bernoulli, beta and Cox proportional hazard (partial) likelihoods (the first two using logit links).

The purpose of this article is to develop methods to allow the rich variety of smoothers illustrated in Figure 1 to be used in models beyond the exponential family, a task for which general methods were not previously available. However, for the special case of univariate P-splines (Eilers and Marx 1996; Marx and Eilers 1998) some comparison with existing methods is possible, in particular using R package `gamlss` (Rigby and Stasinopoulos 2005, 2014) and the BayesX package (Fahrmeir and Lang 2001; Fahrmeir, Kneib, and Lang 2004; Brezger and Lang 2006; Umlauf et al. 2015; Belitz et al. 2015, *www.bayesx.org*). For this special case both packages implement models using essentially the same penalized likelihoods used by the new method, but they optimize localized marginal likelihood scores within the penalized likelihood optimization algorithm to estimate the smoothing parameters.

The comparison was performed using data simulated from models with the linear predictor given above (but without any random effect terms). Comparison of the new method with GAMLSS was only possible for negative binomial, beta, scaled t and simple zero inflated Poisson families, and with BayesX was only possible for the ordered categorical model (BayesX has a negative binomial family, but it is currently insufficiently stable for a sensible comparison to be made). Simulations with both uncorrelated and correlated covariates were considered. Three hundred replicates of the sample size 400 were produced for each considered family at three levels of noise (see SA E for further details). Models were estimated using the correct link and additive structure, and using P-splines with basis dimensions of 10, 10, 15, and 8 which were chosen to avoid any possibility of forced oversmoothing, while keeping down computational time.

Model performance for the negative binomial (nb), beta, scaled t (scat), and zero inflated Poisson (zip) families was compared via MSE, $n^{-1}\sum_{i=1}^{n}\left\{\hat{\eta}(\mathbf{x}_i)-\eta_t(\mathbf{x}_i)\right\}^2$, on the additive predictor scale. The Brier score, $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{R}(p_{ij}-\hat{p}_{ij})^2$, was used to measure the performance for the ordered categorical (ocat) family, where R is a number of categories, p_{ij} are true category probabilities and $\hat{p}_{ij}$ their estimated values. In addition, the computational performance (CPU time) of the alternative methods was recorded. Figure 4 summarizes the results. In general, the new method provides a small improvement in statistical performance, which is slightly larger when covariates are correlated. The correlated covariate setting is the one in which local approximate smoothness selection methods would be expected to perform less well, relative to "whole model" criteria. In terms of speed and reliability the new method is an improvement, especially for correlated covariates, which tend to lead to reduced numerical stability, leading the alternative methods to fail in up to 4% of cases.

7. Example: Predicting Prostate Cancer

This section and the next provide example applications of the new methods, while the online SA F provides further examples

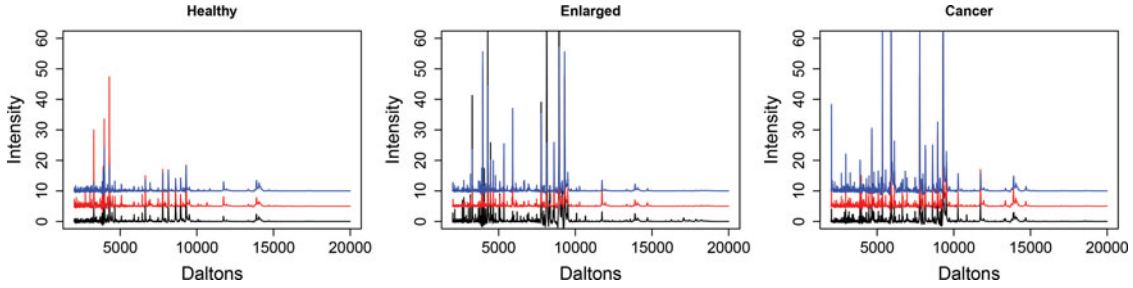

Figure 5. Three representative protein mass spectra (centered and normalized) from serum taken from patients with apparently healthy prostate, enlarged prostate, and prostate cancer. It would be useful to be able to predict disease status from the spectra. The red and blue spectra have been shifted upward by 5 and 10 units, respectively.

in survival analysis and animal distribution modeling. Figure 5 shows representative protein mass spectra from serum taken from patients with a healthy prostate, relatively benign prostate enlargement and prostate cancer (see Adam et al. 2002). To avoid the need for intrusive biopsy there is substantial interest in developing noninvasive screening tests to distinguish cancer, healthy and more benign conditions. One possible model is an ordered categorical signal regression in which the mean of a logistically distributed latent variable z is given by

$$\mu_i = \alpha + \int f(D)v_i(D)dD,$$

where $f(D)$ is an unknown smooth function of mass D (in Daltons) and $v_i(D)$ is the ith spectrum. The probability of the patient lying in category 1, 2, or 3 corresponding to "healthy," "benign enlargement" and "cancer" is then given by the probability of z_i lying in the range $(-\infty, -1]$, $(-1, \theta]$ or (θ, ∞), respectively (see online SA K).

Given the methods developed in this article, estimation of this model is routine, as is the exploration of whether an adaptive smooth should be used for f, given the irregularity of the spectra. Figure 6 shows some results of model fitting. The estimated $f(D)$ is based on a rank 100 thin plate regression spline. Its effective degrees of freedom is 29. An adaptive smooth gives almost identical results. The right panel shows a QQ-plot of ordered deviance residuals against simulated theoretical quantiles (Augustin, Sauleau, and Wood 2012). There is modest deviation in the lower tail. The middle panel shows boxplots of the probability of cancer according to the model for the three observed categories. Cancer and healthy are quite well separated, but cancer and benign enlargement less so. For cases with cancer, the model gave cancer a higher probability than normal prostate in 92% of cases, and a higher probability that either other category in 83% of cases. For healthy patients the model gave the normal category higher probability than cancer in 85% of cases and the highest probability in 77% of cases. These results are somewhat worse than those reported by Adam et al. (2002) for a relatively complex machine learning method which involved first preprocessing the spectra to identify peaks believed to be discriminating. On the other hand the signal regression model here would allow the straightforward inclusion of further covariates, and does automatically supply uncertainty estimates.

8. Multivariate Additive Modeling of Fuel Efficiency

Figure 7 shows part of a dataset on the fuel efficiency of 207 U.S. car models, along with their characteristics (Bache and Lichman 2013). Two efficiency measures were taken: miles per gallon (MPG) in city driving, and the same for highway driving. One possible model might be a bivariate additive model, as detailed in the online SA H, where the two `mpg` measurements are modeled as bivariate Gaussian, with means given by separate linear predictors for the two components. A priori, it might be expected that city efficiency would be highly influenced by weight and highway efficiency by air resistance and, hence, by frontal area or some other combination of height and width of the car.

The linear predictors for the two components were based on the additive fixed effects of factors "fuel type" (petrol or diesel), "style" of car (hatchback, sedan, etc.) and "drive" (all-, front- or rear-wheel). In addition i.i.d. Gaussian random effects of the 22 car manufacturers were included, as well as smooth additive effects of car weight and horsepower. Additive and tensor product smooths of height and width were tried as well as a smooth of the product of height and width, but there was no evidence to justify their inclusion-term selection penalties (Marra and

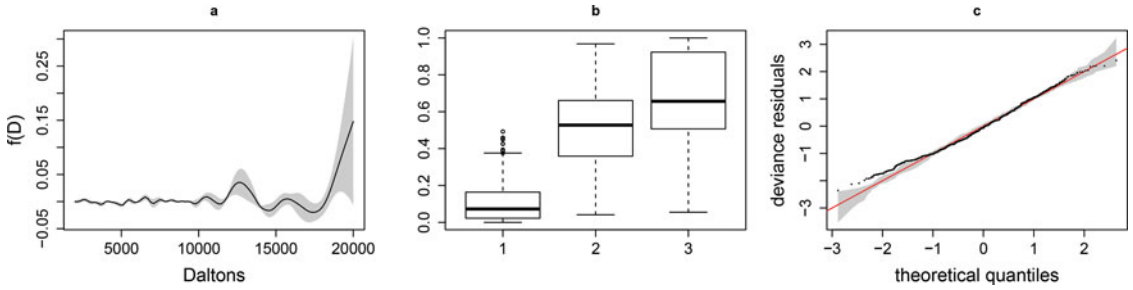

Figure 6. Results from the ordered categorical prostate model fit. (a) The estimated coefficient function $f(D)$ with 95% confidence interval. (b) Boxplots of the model probability of cancer, for the 3 observed states (1, healthy, 2, enlarged and 3, cancer). (c) QQ-plot of ordered deviance residuals against simulated theoretical quantiles, indicating some mismatch in the lower tail.

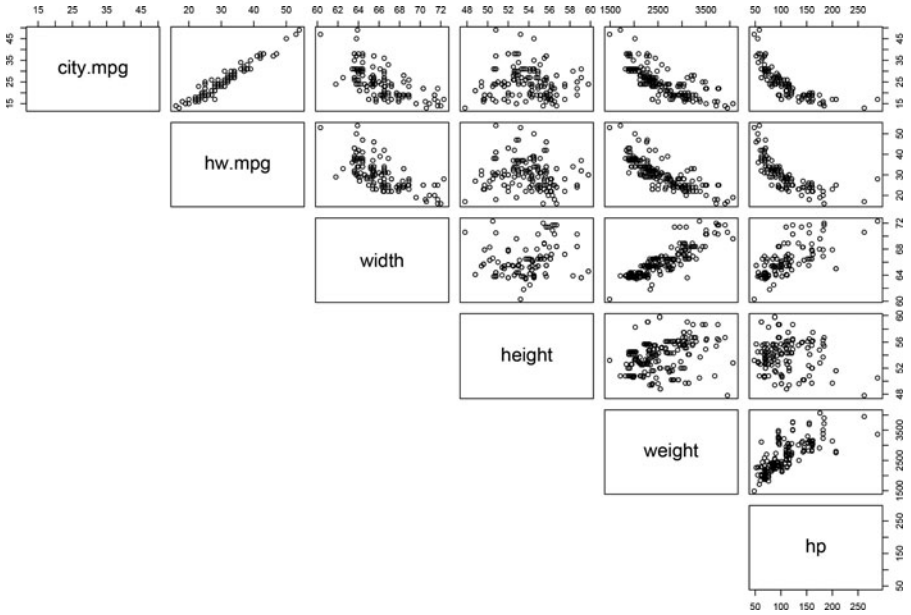

Figure 7. Part of a dataset from the USA on fuel efficiency of cars.

Wood 2011) remove them, p-values indicate they are not significant and AIC suggests that they are better dropped.

The possibility of smooth interactions between weight and horsepower were also considered, using smooth main effects plus smooth interaction formulations of the form $f_1(h) + f_2(w) + f_3(h, w)$. The smooth interaction term f_3 can readily be constructed in a way that excludes the main effects of w and h, by constructing its basis using the usual tensor product construction (e.g., Wood 2006), but based on marginal bases into which the constraints $\sum_i f_1(h_i) = 0$ and $\sum_i f_2(w_i) = 0$ have already been absorbed by linear reparameterization. The marginal smoothing penalties and, hence, the induced tensor product smoothing penalties are unaffected by the marginal constraint absorption. This construction is the obvious generalization of the construction of parametric interactions in linear models, and is simpler than the various schemes proposed in the literature.

The interactions again appear to add nothing useful to the model fit, and we end up with a model in which the important smooth effects are horse power (hp) and weight, while the important fixed effects are fuel type and drive, with diesel giving lower fuel consumption than petrol and all wheel drive giving higher consumption than the two-wheel drives. These effects were important for both city and highway, whereas the random effect of manufacturer was only important for the city. Figure 8 shows the smooth and random effects for the city and highway linear predictors. Notice the surprising similarity between the effects although the city smooth effects are generally slightly less

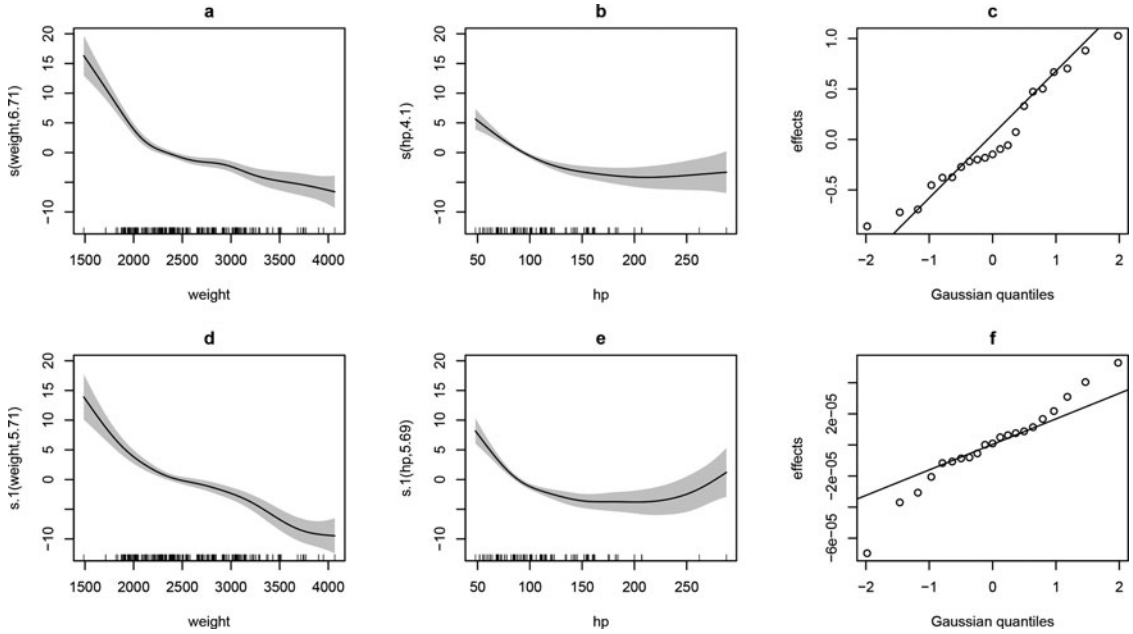

Figure 8. Fitted smooth and random effects for final car fuel efficiency model. Panels (a)–(c) relate to the city fuel consumption, while (d)–(f) are for the highway. (c) and (f) are normal QQ-plots of the predicted random effects for manufacturer, which in the case of highway MPG are effectively zero.

pronounced than those for the highway. The overall r^2 for the model is 85% but with the city and highway error MPG standard deviation estimated as 1.9 and 2.3 MPG respectively. The estimated correlation coefficient is 0.88.

9. Discussion

This article has outlined a practical framework for smooth regression modeling with reduced rank smoothers, for likelihoods beyond the exponential family. The methods build seamlessly on the existing framework for generalized additive modeling, so that practical application of any of the models implemented as part of this work is immediately accessible to anyone familiar with GAMs via penalized regression splines. The key novel components contributed here are (i) general, reliable and efficient smoothing parameter estimation methods based on maximized Laplace approximate marginal likelihood, (ii) a corrected AIC and distributional results incorporating smoothing parameter uncertainty to aid model selection and further inference, and (iii) demonstration of the framework's practical utility by provision of the details for some practically important models. The proposed methods should be widely applicable in situations in which effects are really smooth, and the methods scale well with the number of smooth model terms. In situations in which some component functions are high rank random fields, then the INLA approach of Rue, Martino, and Chopin (2009) will be much more efficient; however, there are trade-offs between efficiency and stability in this case, since pivoting, used by our method to preserve stability, has instead to be employed to preserve sparsity in the INLA method (see online SA K).

The methods are implemented in R package mgcv from version 1.8 (see online SA M).

Appendix A: Implicit Differentiation in the Extended Gam Case

Let $\mathcal{D}^{\hat\beta\hat\beta}_{ij}$ denote elements of the inverse of the Hessian matrix $(\mathbf{X}^{\mathsf T}\mathbf{W}\mathbf{X}+\mathbf{S}^\lambda)$ with elements $\mathcal{D}^{ij}_{\hat\beta\hat\beta}$, and note that $\hat{\boldsymbol\beta}$ is the solution of $\mathcal{D}^{i}_{\hat\beta}=0$. Finding the total derivative with respect to $\boldsymbol\theta$ of both sides of this we have

$$\mathcal{D}^{i\,\boldsymbol k}_{\hat\beta\hat\beta}\frac{\mathrm d\hat\beta_{\boldsymbol k}}{\mathrm d\theta_j}+\mathcal{D}^{ij}_{\hat\beta\theta}=0,\ \text{implying that}\ \frac{\mathrm d\hat\beta_k}{\mathrm d\theta_j}=-\mathcal{D}^{\hat\beta\hat\beta}_{k\,\boldsymbol i}\mathcal{D}^{\boldsymbol i\,j}_{\hat\beta\theta}$$

Differentiating once more yields

$$\frac{\mathrm d^2\hat\beta_i}{\mathrm d\theta_j\mathrm d\theta_k}=-\mathcal{D}^{\hat\beta\hat\beta}_{i\,\boldsymbol l}\left(\mathcal{D}^{\boldsymbol l\,\boldsymbol p\boldsymbol q}_{\hat\beta\hat\beta\hat\beta}\frac{\mathrm d\hat\beta_{\boldsymbol q}}{\mathrm d\theta_j}\frac{\mathrm d\hat\beta_{\boldsymbol p}}{\mathrm d\theta_k}+\mathcal{D}^{\boldsymbol l\,\boldsymbol p\ j}_{\hat\beta\hat\beta\ \theta}\frac{\mathrm d\hat\beta_{\boldsymbol p}}{\mathrm d\theta_k}+\mathcal{D}^{\boldsymbol l\,\boldsymbol p\,k}_{\hat\beta\hat\beta\ \theta}\frac{\mathrm d\hat\beta_{\boldsymbol p}}{\mathrm d\theta_j}+\mathcal{D}^{\boldsymbol l\ jk}_{\hat\beta\ \theta\theta}\right).$$

The required partials are obtained from those generically available for the distribution and link used and by differentiation of the penalty. Generically we can obtain derivatives of D_i w.r.t μ_i and $\boldsymbol\theta$.

The preceding expressions hold whether θ_j is a parameter of the likelihood or a log smoothing parameter. Suppose Λ denotes the set of log smoothing parameters, then

$$\mathcal{D}^{ij}_{\beta\theta}=\begin{cases}2\exp(\theta_j)S^j_{i\boldsymbol k}\beta_{\boldsymbol k} & \theta_j\in\Lambda\\ D^{ij}_{\beta\theta} & \text{otherwise,}\end{cases}$$

where $\mathbf S^j$ here denotes the penalty matrix associated with θ_j. Similarly

$$\mathcal{D}^{l\,pj}_{\beta\beta\theta}=\begin{cases}2\exp(\theta_j)S^j_{lp} & \theta_j\in\Lambda\\ D^{l\,pj}_{\beta\beta\theta} & \text{otherwise}\end{cases}\quad\text{while}$$

$$\mathcal{D}^{l\,jk}_{\beta\theta\theta}=\begin{cases}2\exp(\theta_j)S^j_{l\boldsymbol m}\beta_{\boldsymbol m} & j=k;\ \theta_j,\theta_k\in\Lambda\\ D^{l\,jk}_{\beta\theta\theta} & \theta_j,\theta_k\notin\Lambda\\ 0 & \text{otherwise.}\end{cases}$$

Derivatives with respect to $\boldsymbol\eta$ are obtained by standard transformations

$$D^i_\eta=D^i_\mu/h'_i,\tag{A.1}$$

where $h'_i=h'(\mu_i)$ and more primes indicate higher derivatives. Furthermore,

$$D^{i\,i}_{\eta\eta}=D^{i\,i}_{\mu\mu}/h'^2_i-D^i_\mu h''_i/h'^3_i,\tag{A.2}$$

where the expectation of the second term on the right-hand side is zero at the true parameter values.

$$\begin{aligned}\text{Also } D^{i\,i\,i}_{\eta\eta\eta}&=D^{i\,i\,i}_{\mu\mu\mu}/h'^3_i-3D^{i\,i}_{\mu\mu}h''_i/h'^4_i\\&\quad+D^i_\mu\left(3h''^2_i/h'^5_i-h'''_i/h'^4_i\right),\ \text{and}\end{aligned}\tag{A.3}$$

$$\begin{aligned}D^{i\,i\,i\,i}_{\eta\eta\eta\eta}&=D^{i\,i\,i\,i}_{\mu\mu\mu\mu}/h'^4_i-6D^{i\,i\,i}_{\mu\mu\mu}h''_i/h'^5_i+D^{i\,i}_{\mu\mu}(15h''^2_i/h'^6_i\\&\quad-4h'''_i/h'^5_i)-D^i_\mu(15h''^3_i/h'^7_i-10h''_ih'''_i/h'^6_i+h''''_i/h'^5_i).\end{aligned}\tag{A.4}$$

Mixed partial derivatives with respect to $\boldsymbol\eta/\boldsymbol\mu$ and $\boldsymbol\theta$ transform in the same way, the formula to use depending on the number of η subscripts. The rules relating the derivatives w.r.t $\boldsymbol\eta$ to those with respect to $\boldsymbol\beta$ are much easier: $D^i_\beta=D^{\boldsymbol k}_\eta X_{\boldsymbol k i}$, $D^{i\,j}_{\beta\beta}=D^{\boldsymbol{kk}}_{\eta\eta}X_{\boldsymbol k i}X_{\boldsymbol k j}$, $D^{i\,j\,k}_{\beta\beta\beta}=D^{\boldsymbol{lll}}_{\eta\eta\eta}X_{\boldsymbol l i}X_{\boldsymbol l j}X_{\boldsymbol l k}$. Again mixed partials follow the rule appropriate for the number of β subscripts present. It is usually more efficient to compute using the definitions, rather than forming the arrays explicitly.

The ingredients so far are sufficient to compute $\hat{\boldsymbol\beta}$ and its derivatives with respect to $\boldsymbol\theta$. We now need to consider the derivatives of $\mathcal V$ with respect to $\boldsymbol\theta$. Considering $\mathcal D$ first, the components relating to the penalties are straightforward. The deviance components are then

$$\frac{\mathrm dD}{\mathrm d\theta_i}=D^{\boldsymbol j}_{\hat\eta}\frac{\mathrm d\hat\eta_{\boldsymbol j}}{\mathrm d\theta_i}+D^i_{\hat\theta}\ \text{and}\ \frac{\mathrm d^2D}{\mathrm d\theta_i\mathrm d\theta_j}=D^{\boldsymbol{kk}}_{\hat\eta\hat\eta}\frac{\mathrm d\hat\eta_{\boldsymbol k}}{\mathrm d\theta_i}\frac{\mathrm d\hat\eta_{\boldsymbol k}}{\mathrm d\theta_j}+D^{\boldsymbol k}_{\hat\eta}\frac{\mathrm d^2\hat\eta_{\boldsymbol k}}{\mathrm d\theta_i\mathrm d\theta_j}+D^{\boldsymbol k\,j}_{\hat\eta\,\theta}\frac{\mathrm d\hat\eta_{\boldsymbol k}}{\mathrm d\theta_i}+D^{\boldsymbol k\,i}_{\hat\eta\,\theta}\frac{\mathrm d\hat\eta_{\boldsymbol k}}{\mathrm d\theta_j}+D^{ij}_{\theta\,\theta},$$

where the derivatives of $\hat{\boldsymbol\eta}$ are simply $\mathbf X$ multiplied by the derivatives of $\hat{\boldsymbol\beta}$. The partials of $\tilde l$ are distribution specific. The derivatives

of the determinant terms are obtainable using Wood (2011) once derivatives of w_i with respect to $\boldsymbol{\theta}$ have been obtained. These are

$$\frac{\mathrm{d}w_i}{\mathrm{d}\theta_j} = \frac{1}{2}D^{i\,i\,i}_{\hat\eta\hat\eta\hat\eta}\frac{\mathrm{d}\hat\eta_i}{\mathrm{d}\theta_j} + \frac{1}{2}D^{i\,i\,j}_{\eta\,\hat\eta\,\theta},$$

$$\frac{\mathrm{d}^2w_i}{\mathrm{d}\theta_j\mathrm{d}\theta_k} = \frac{1}{2}D^{i\,i\,i\,i}_{\hat\eta\hat\eta\hat\eta\hat\eta}\frac{\mathrm{d}\hat\eta_i}{\mathrm{d}\theta_j}\frac{\mathrm{d}\hat\eta_i}{\mathrm{d}\theta_k} + \frac{1}{2}D^{i\,i\,i}_{\hat\eta\hat\eta\hat\eta}\frac{\mathrm{d}^2\hat\eta_i}{\mathrm{d}\theta_j\mathrm{d}\theta_k} + \frac{1}{2}D^{i\,i\,i\,k}_{\hat\eta\hat\eta\hat\eta\theta}\frac{\mathrm{d}\hat\eta_i}{\mathrm{d}\theta_j} + \frac{1}{2}D^{i\,i\,i\,j}_{\hat\eta\hat\eta\hat\eta\theta}\frac{\mathrm{d}\hat\eta_i}{\mathrm{d}\theta_k} + \frac{1}{2}D^{i\,i\,j\,k}_{\hat\eta\hat\eta\theta\theta}.$$

Supplementary Materials

The online supplementary materials contain additional appendices for the article.

Acknowledgment

We thank the anonymous referees for a large number of very helpful comments that substantially improved the paper and Phil Reiss for spotting an embarrassing error in Supplementary Appendix A.

Funding

SNW and NP were funded by EPSRC grant EP/K005251/1 and NP was also funded by EPSRC grant EP/I000917/1. BS was funded by the German Research Association (DFG) Research Training Group "Scaling Problems in Statistics" (RTG 1644). SNW is grateful to Carsten Dorman and his research group at the University of Freiburg, where the extended GAM part of this work was carried out.

References

Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z., et al. (2002), "Serum Protein Fingerprinting Coupled With a Pattern-Matching Algorithm Distinguishes Prostate Cancer From Benign Prostate Hyperplasia and Healthy Men," *Cancer Research*, 62, 3609–3614. [1559]

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *International Symposium on Information Theory*, eds. B. Petran, and F. Csaaki, Budapest: Akadeemiai Kiadi, pp. 267–281. [1556]

Anderssen, R., and Bloomfield, P. (1974), "A Time Series Approach to Numerical Differentiation," *Technometrics*, 16, 69–75. [1550]

Augustin, N. H., Sauleau, E.-A., and Wood, S. N. (2012), "On Quantile Quantile Plots for Generalized Linear Models," *Computational Statistics & Data Analysis*, 56, 2404–2409. [1559]

Bache, K., and Lichman, M. (2013), "UCI Machine Learning Repository," available at *http://archive.ics.uci.edu/ml/*. [1559]

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2015), "Bayesx: Software for Bayesian Inference in Structured Additive Regression Models," available at *http://www.statistik.lmu.de/~bayesx/bayesx.html*. [1558]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [1554]

Brezger, A., and Lang, S. (2006), "Generalized Structured Additive Regression Based on Bayesian p-Splines," *Computational Statistics & Data Analysis*, 50, 967–991. [1558]

Cline, A. K., Moler, C. B., Stewart, G. W., and Wilkinson, J. H. (1979), "An Estimate for the Condition Number of a Matrix," *SIAM Journal on Numerical Analysis*, 16, 368–375. [1552]

Cox, D. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, Series B, 34, 187–220. [1553]

Davison, A. C. (2003), *Statistical Models*, Cambridge, UK: Cambridge University Press. [1556]

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science* 11, 89–121. [1558]

Fahrmeir, L., Kneib, T., and Lang, S. (2004), "Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective," *Statistica Sinica*, 14, 731–761. [1558]

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013), "*Regression Models*, New York: Springer. [1548]

Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Applied Statistics*, 50, 201–220. [1558]

Greven, S., and Kneib, T. (2010), "On the Behaviour of Marginal and Conditional AIC in Linear Mixed Models," *Biometrika*, 97, 773–789. [1549,1556,1557]

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models" (with discussion), *Statistical Science*, 1, 297–318. [1548]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman & Hall. [1548,1556,1557]

Kass, R. E., and Steffey, D. (1989), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," *Journal of the American Statistical Association*, 84, 717–726. [1555,1556]

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2014), "Bayesian Structured Additive Distributional Regression for Multivariate Responses," *Journal of the Royal Statistical Society*, Series C, 64, 569–591. [1553]

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015), "Bayesian Structured Additive Distributional Regression With an Application to Regional Income Inequality in Germany," *Annals of Applied Statistics*, 9, 1024–1052. [1553]

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974. [1550]

Liang, H., Wu, H., and Zou, G. (2008), "A Note on Conditional AIC for Linear Mixed-Effects Models," *Biometrika*, 95, 773–778. [1556]

Marra, G., and Wood, S. N. (2011), "Practical Variable Selection for Generalized Additive Models," *Computational Statistics & Data Analysis*, 55, 2372–2387. [1560]

Marra, G., and Wood, S. N. (2012), "Coverage Properties of Confidence Intervals for Generalized Additive Model Components," *Scandinavian Journal of Statistics*, 39, 53–74. [1550]

Marx, B. D., and Eilers, P. H. (1998), "Direct Generalized Additive Modeling With Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209. [1558]

Nocedal, J., and Wright, S. (2006), "*Numerical Optimization* (2nd ed.), New York: Springer Verlag. [1552]

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143. [1550]

Reiss, P. T., and Ogden, T. R. (2009), "Smoothing Parameter Selection for a Class of Semiparametric Linear Models," *Journal of the Royal Statistical Society*, Series B, 71, 505–523. [1550]

Rigby, R., and Stasinopoulos, D. M. (2005), "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society*, Series C, 54, 507–554. [1548,1553,1558]

Rigby, R. A., and Stasinopoulos, D. M. (2014), "Automatic Smoothing Parameter Selection in GAMLSS With an Application to Centile Estimation," *Statistical Methods in Medical Research*, 23, 318–332. [1558]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [1561]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [1548]

Säfken, B., Kneib, T., van Waveren, C.-S., and Greven, S. (2014), "A Unifying Approach to the Estimation of the Conditional Akaike Information in Generalized Linear Mixed Models," *Electronic Journal of Statistics*, 8, 201–225. [1556]

Shun, Z., and McCullagh, P. (1995), "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society*, Series B, 57, 749–760. [1550]

Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting," *Journal of the Royal Statistical Society*, Series B, 47, 1–53. [1550]

Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015), "Structured Additive Regression Models: An r Interface to Bayesx," *Journal of Statistical Software*, 63, 1–46. [1558]

Wahba, G. (1983), "Bayesian Confidence Intervals for the Cross Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Series B, 45, 133–150. [1550]

——— (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, pp. 1378–1402. [1550]

Wood, S. N. (2000), "Modelling and Smoothing Parameter Estimation With Multiple Quadratic Penalties," *Journal of the Royal Statistical Society*, Series B, 62, 413–428. [1548]

——— (2006), "Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models," *Biometrics*, 62, 1025–1036. [1560]

——— (2011), "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society*, Series B, 73, 3–36. [1548,1550,1551,1552,1555,1562]

Yee, T. W., and Wild, C. (1996), "Vector Generalized Additive Models," *Journal of the Royal Statistical Society*, Series B, 481–493. [1548,1553]

Yu, D., and Yau, K. K. (2012), "Conditional Akaike Information Criterion for Generalized Linear Mixed Models," *Computational Statistics & Data Analysis*, 56, 629–644. [1556]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1250576

Comment

Thomas Kneib

Chair of Statistics, Georg August University Göttingen, Göttingen, Germany

1. Introduction

It probably does not come as a surprise that I enjoyed reading the article under discussion with its developments for flexible regression modeling beyond the standard class of generalized additive models with responses originating from the simple exponential family. My research interests always had a strong overlap with the ones of Simon and his group, albeit with a stronger focus on Bayesian formulations. In the current article, Simon Wood, Natalya Pya, and Benjamin Säfken develop stable and versatile statistical methodology for what they call "general smooth models" and what we call "structured additive distributional regression models" (Klein et al. 2015b, 2015a). While there are certain subtle differences in the model structures supported by the one or the other approach, both share the same idea that relies on the following model structure:

- As a distributional assumption for the response, general types of distributions not necessarily from the simple exponential family are permitted. The only requirement is that the densities are smooth enough in the parameters to allow for the evaluation of a certain number of derivatives.
- In contrast to mean regression where a regression predictor is assumed for the (transformed) expectation of the response, a regression predictor is supplemented to potentially all parameters of the response distribution.
- The predictor is additively decomposed into a number of nonlinear components.
- These components are expanded in suitable basis functions and are associated with quadratic penalties/Gaussian priors to enforce specific properties such as smoothness or shrinkage.

The main contributions of the current article are (from my perspective)

- The detailed development of a stable and general inferential scheme that allows us to estimate a variety of distributional regression specifications with predictors of considerable complexity.
- The proposition of a novel Akaike information criterion (AIC) for general smooth models that takes uncertainty in the selection of smoothing parameters into account.
- The development of several results on the asymptotic behavior of penalized cubic splines.

2. Multivariate Regression Models

Although multivariate regression models are included in the article by Wood, Pya and Säfken, I would like to further emphasize the value of combining distributional regression ideas with multivariate response structures. Wood, Pya and Säfken followed the idea of seemingly unrelated regression (SUR, Smith and Kohn 2000; Lang et al. 2003) by assuming a multivariate normal specification for the responses with a fixed correlation structure. While this has the advantage of allowing for a basically arbitrary number of response components, it has the disadvantage that both the variances and (more importantly) the dependence parameters are not allowed to be modified by covariate values. The main difficulty in doing the latter is to obtain an interpretable and simple parameterization

CONTACT Thomas Kneib *tkneib@uni-goettingen.de* Department of Statistics and Econometrics, Georg-August-Universität Göttingen, Göttingen 37073, Germany.

of the covariance matrix that allows us to provide link functions between the parameters of the covariance matrix and the regression predictors.

For bivariate responses, making the correlation covariate-dependent is straightforward since the only restrictions to take into account are the nonnegativity of the variances and the restriction of the correlation coefficient to the interval $[-1, 1]$, which can for example be dealt with using the exponential and Fisher's z-transformation (Klein et al. 2015a). For the truly multivariate case, things are getting more difficult since positive definiteness of the covariance matrix has to be ensured. The most promising avenue for achieving this seems to be the log-Cholesky parameterization of the precision matrix that also links to interpretation via the conditional independence encoded by zero elements in the precision matrix (Pourahmadi 2011). Still it is a special challenge to provide simple interpretational parameterizations of dependence in the multivariate case.

More flexibility in the dependence structure can be gained by considering copula specifications (Joe 1997; Nelson 2006). For bivariate responses, there is a multitude of copulas to pick from, which, for example, enable lower or upper tail dependence or other deviations from linear correlation (Smith and Khaled 2012; Radice, Marra, and Wojtys 2016; Klein and Kneib 2016) but again it seems challenging to go beyond the bivariate case (see Joe 2014, for some suggestions in this direction using vine copulas).

3. Akaike Information Criterion

Constructing an appropriate AIC for additive models has been a challenge due to the necessity of incorporating uncertainty concerning the selection of smoothing parameters. This has been notoriously difficult in particular for cases, where the smoothing parameter is large, that is, the complex model is close to reducing to a simple alternative defined by the null space of the penalty matrix. This situation contradicts standard asymptotic considerations where parameters have to be bounded away from the boundaries of their parameter space. Unfortunately, this situation is also the most interesting case for model selection since one would then be interested in whether the model can safely be simplified or not.

It is therefore extremely valuable that the article by Wood, Pya and Säfken provides a solution to this issue. I would guess that the AIC developed in the article is actually also applicable in various types of mixed models with complex hierarchical structure but would be interested in hearing the opinion of the authors on this issue and whether there may be any difficulties due to the more complex structure of the resulting covariance matrix. In addition, I was wondering whether the AIC is indeed invariant under reparameterization of the smoothing parameters, that is, would we obtain the same results when considering, for example, the inverses of the smoothing parameters?

4. Frequentist vs. Bayes

Wood, Pya, and Säfken provided a comparison with Bayesian estimates derived from the implementation in BayesX (www.bayesx.org, Belitz et al. 2015) and found their implementation to be much more stable and competitive in terms of estimation results. Numerical (in)stability is indeed still a drawback of the current BayesX implementation that sometimes strongly depends on the scaling of the responses and the size of effects. Still, I would like to add some additional, more theoretical points to the comparison between the methods developed by Wood, Pya, and Säfken and their Bayesian counterparts:

- While the Bayesian estimates require only the existence of second derivatives of the log-likelihood with respect to the predictors, the proposed methodology uses up to fourth derivatives and also cross-derivatives. This certainly contributes to the numerical stability but also is computationally more demanding both in terms of implementation when setting up a new model and when actually performing estimation for a given dataset.
- While certain asymptotic results can probably be safely assumed for the regression coefficients of a general smooth model, transferring these results to quantities derived from the model seems much more challenging. For example, in Klein et al. (2015b), we derive the Gini inequality index from an estimated Dagum distribution for incomes in Germany. In this case, it is straightforward to perform a Bayesian assessment of uncertainty for the Gini index based on the samples provided by Markov chain Monte Carlo simulations, while it seems rather challenging to derive corresponding frequentist results. Since it is often the case that the function estimates themselves are not easy to interpret in general smooth models, such uncertainty assessments are of huge applied relevance.
- In general, I would be interested in getting to know the opinion of Wood, Pya and Säfken about how to best interpret the results achieved with a general smooth model where a covariate enters the regression specification for multiple parameters of the response distribution.

References

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015), "BayesX: Software for Bayesian Inference in Structured Additive Regression Models," Version 3.0.2, 2015. Available at *http://www.bayesx.org*. [1564]

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, London: Chapman & Hall. [1564]

——— (2014), *Dependence Modeling With Copulas*, London: Chapman & Hall. [1564]

Klein, N., and Kneib, T. (2016), "Simultaneous Estimation in Structured Additive Conditional Copula Regression Models: A Unifying Bayesian Approach," *Statistics and Computing*, 26, 841–860. [1564]

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015a), "Bayesian Structured Additive Distributional Regression for Multivariate Responses," *Journal of the Royal Statistical Society*, Series C, 64, 569–591. doi/10.1111/rssc.12090. [1563,1564]

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015b), "Bayesian Structured Additive Distributional Regression With an Application to Regional Income Inequality in Germany," *Annals of Applied Statistics*, 9, 1024–1052. [1563]

Lang, S., Adebayo, S. B., Fahrmeir, L., and Steiner, W. J. (2003), "Bayesian Geoadditive Seemingly Unrelated Regression," *Computational Statistics*, 18, 263–292. [1563]

Nelson, R. (2006), *An Introduction to Copulas* (2nd ed.), New York: Springer. [1564]

Pourahmadi, M. (2011), "Covariance Estimation: The GLM and Regularization Perspectives," *Statistical Science*, 26, 369–387. [1564]
Radice, R., Marra, G., and Wojtys, M. (2016), "Copula Regression Spline Models for Binary Outcomes," *Statistics and Computing*, 26, 981–995. [1564]
Smith, M. S., and Khaled, M. A. (2012), "Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation," *Journal of the American Statistical Association*, 107, 290–303. [1564]
Smith, M. S., and Kohn, R. (2000), "Nonparametric Seemingly Unrelated Regression," *Journal of Econometrics*, 98, 257–281. [1563]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1250579

Comment

Thomas W. Yee

Department of Statistics, University of Auckland, Auckland, New Zealand

The authors are congratulated for this significant and detailed contribution, which extends the exceedingly popular GAM technique to higher realms of generality and functionality. It is interesting to see how GAMs have developed over the last two decades or so. Not surprisingly, as authors have become more ambitious, the level of sophistication has escalated and this article is no exception. For a given model, the article conveys the substantial amount of work needed to implement automatic smoothing parameter selection based on LAML—these involve fourth-order derivatives, unremitting attention to the numerical analysis of computations, and careful programming—so that the code runs robustly and efficiently on real data. Inference and asymptotic properties are also considered here, as well as a framework that allows a rich variety of smoothing based on low-rank penalized splines. It is unfortunate that the vast majority of researchers in the statistical sciences are still content with pen and paper (or word processor), and do not offer the "full service" of developing new methodology from beginning to end, which includes a user-friendly software implementation that people can use straightaway.

Early work starting in the mid-1980s led by Hastie and Tibshirani developed GAMs based on backfitting, and were largely confined to the exponential family. Smoothing parameter selection was difficult for this. Over the last $1\frac{1}{2}$ decades, the first author has led the charge of automating smoothing parameter selection (and dispensing of backfitting), resulting in several methods such as UBRE/GCV optimization in conjunction with PIRLS, ML- and REML-based methods, and now refined LAML-based methods. However, these works were also largely confined to the exponential family, bar the present article. In my own work, I have, in the large part, had the strong conviction of breaking out of the shackles of the exponential family from the outset, and the handling of multiple linear predictors. Doing so really does open one up to a lot more of the statistical universe. Current work with C. Somchit and C. Wild on developing automatic smoothing parameter selection for the vector generalized additive model (VGAM) class, which is very general, is very briefly described in Section 1.1 and a working implementation should hopefully appear within the next 12 months. This means that we have attempted to reach one of the goals of the present article (hereafter referred to as "WPS"), albeit, approaching from a different direction.

The comments below are shared between the article and some selected supplementary appendices.

1. General Comments

The authors have done a valiant job developing LAML estimation for general settings and implementing several important regression models. Some computational tricks and inferential by-products have been found along the way. With the possibility to handle multiple additive predictors η_j now, is it possible to constrain some of these linearly? For example, in the case of two of them, can one fit $\eta_1 = a + b\,\eta_2$, where a and b are estimated too? There are several reasons for pursuing this idea. This is a special case of a reduced-rank regression, where the overall regression coefficients are subject to a rank restriction. There are several benefits, such as a lower computational cost, increased parsimony of parameters, interpretation in terms of latent variables, and low-dimensional views in the case of a rank-2 model. Some particularly useful regression models arise as special cases, such as a negative binomial regression with variance function $\mu + \delta_1 \mu^{\delta_2}$ (known as the NB-P model in the count literature). In the case of the multinomial logit model, when the number of classes and number of explanatory variables is even moderate then the number of regression coefficients becomes sizeable, hence some form of dimension reduction becomes warranted. Such a model was called the stereotype model by Anderson (1984). Additionally, it would be very useful if $\eta_j = f_j(\nu)$ were developed for regular models, where $\nu = \boldsymbol{c}^T \boldsymbol{x}$ is an optimal linear combination of the explanatory variables. For this, the exponential family case is referred to as a single-index model and it is a special case of a generalized additive index model (GAIM; Chen and Samworth 2016). Some of these ideas are described in Yee (2015).

CONTACT Thomas W. Yee t.yee@auckland.ac.nz Department of Statistics, University of Auckland, Auckland 1142, New Zealand.

Section 6 raises the issue of cost–benefit when some simulation results show that occasionally the new method gives only a small improvement in statistical performance relative to some less complicated methods. I have found that some prudence is required when deciding whether a certain complex procedure is worth implementing, especially when it involves a large expenditure of effort. There is a parallel to be seen from the 1980s and 1990s when smoothing was a very active field and yet it is probably safe to say that only a very small fraction of this work is seen to be used nowadays. My opinion is that mgcv would have a greater impact per unit effort if more families were added rather than developing further techniques for automatic smoothing parameter selection. An exception to this would be techniques that are surprisingly simple, such as the Fellner–Schall method (Wood and Fasiolo 2016) [personal communication] which only requires the first two derivatives. This promising method needs full development.

The example of Section 7 is known as the proportional odds model, or cumulative logit model, where there are parallelism constraints applied to the η_j's for each explanatory variable bar the intercept. It is a special case of a nonparallel cumulative logit model whereby $\eta_j = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}$. Would an unwarranted parallelism assumption explain the less than perfect behavior in the lower tail of Figure 6? Alternatively, it might be remedied by choosing a link function with a heavier tail or allowing asymmetry, such as a cauchit or complementary log–log link.

In the fuel efficiency example of Section 8, one might wish the smooths to be monotonic unless it is strongly believed that some interaction exists—this applies specifically for Figure 8(e). Would monotonic P-splines be more suitable? And how easy would it be to constrain the smooths of Figures 8(a) and (d) to be equal so that their difference is some constant? In Section 1.1, we analyze these data using some new methodology.

For a general additive model, is it possible to have smoothing parameters from, for example, $f_4(x_4)$ and $f_6(x_6)$ constrained to be equal? An application of this might be to smooth two variables whose support are equal.

1.1. VGAMS with P-splines

Although WPS convey automatic additive smoothing to potentially a very large class of models, a simpler alternative for most models described in the Appendix is to consider them within the VGAM framework and use Wood (2004). Here are some sketch details, which makes use of the notation of Yee (2015). For the VGAM class with constraint matrices $\mathbf{H}_k$,

$$\boldsymbol{\eta}_i = \mathbf{H}_1 \, \boldsymbol{\beta}^*_{(1)} + \sum_{k=2}^{d} \mathbf{H}_k \, \boldsymbol{f}^*_k(x_{ik}), \tag{1}$$

where $\boldsymbol{f}^*_k(x_k) = (f^*_{(1)k}(x_k), \ldots, f^*_{(\mathcal{R}_k)k}(x_k))^T$ is a $\mathcal{R}_k$-vector of smooth functions of x_k to be estimated. Each vector of component functions in (1) generates several columns of the model matrix since they are linear combinations of B-spline basis functions, therefore,

$$\boldsymbol{\eta}_i = \mathbf{H}_1 \, \boldsymbol{\beta}^*_{(1)} + \sum_{k=2}^{d} \mathbf{H}_k \, \mathbf{X}^*_{[ik]} \, \boldsymbol{\beta}^*_{[k]} \tag{2}$$

for some submatrix $\mathbf{X}^*_{[ik]}$. Equation (2) can be further bolstered by allowing terms of the form $\mathbf{H}_k(\ldots, f^*_{(j)k}(x_{ikj})\ldots)$ in $\boldsymbol{\eta}_i$, that is, η_j-specific values of a covariate (known as the $\boldsymbol{x}_{ij}$ or `xij` facility), but details are not given here. We are maximizing the penalized log-likelihood

$$\ell(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \ell_i\{\eta_1(\boldsymbol{x}_i), \ldots, \eta_M(\boldsymbol{x}_i)\} - J(\boldsymbol{\lambda})$$

for a suitably regular model, where the penalty term is

$$J(\boldsymbol{\lambda}) = \sum_{k=2}^{d} \boldsymbol{\beta}^{*T}_{[k]} \left\{\mathbf{P}^*_k \otimes \operatorname{diag}(\lambda_{(1)k}, \ldots, \lambda_{(\mathcal{R}_k)k})\right\} \boldsymbol{\beta}^*_{[k]} = \boldsymbol{\beta}^{*T} \, \mathbf{P}^* \, \boldsymbol{\beta}^*$$

with $\mathbf{P}^*_k = \mathbf{D}^{*T}_k \mathbf{D}^*_k$, and $\mathbf{D}_k$ is the matrix representation of the δth-order differencing operator Δ^δ applied to the B-spline coefficients since the knots for x_k are equidistant.

For a response $\boldsymbol{y}$, the computations are performed by augmenting $\boldsymbol{y}$, the large model matrix $\mathbf{X}_{\text{VAM}}$ comprising blocks of $\mathbf{X}^*_{[ik]}$ and the weight matrices $\mathbf{W} = \operatorname{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$:

$$\boldsymbol{y}' = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0}_\varphi \end{pmatrix}, \quad \mathbf{X}_{\text{PVAM}} = \begin{pmatrix} \mathbf{X}_{\text{VAM}} \\ \widetilde{\mathbf{X}} \end{pmatrix}, \quad \mathbf{W}' = \operatorname{diag}(\mathbf{W}, \, \mathbf{I}_\varphi), \tag{3}$$

for some dimension φ and where $\mathbf{P}^* = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}$ with $\widetilde{\mathbf{X}} =$

$$\Big(\mathbf{O}, \; \operatorname{diag}\Big(\mathbf{D}^*_2 \otimes \operatorname{diag}(\lambda^{1/2}_{(1)2}, \ldots, \lambda^{1/2}_{(\mathcal{R}_2)2}), \ldots, \mathbf{D}^*_d \otimes \operatorname{diag}(\lambda^{1/2}_{(1)d}, \ldots, \lambda^{1/2}_{(\mathcal{R}_d)d})\Big)\Big). \tag{4}$$

For general responses, the above can be embedded within a PIRLS algorithm that uses working responses and working weight matrices, and then the GCV/UBRE is minimized. This could be performed by performance-oriented iteration or by outer iteration, as described by Wood (2006). The advantage of the above approach over WPS is its relative simplicity and fewer requirements such as only needing second-order derivatives.

Applying an implementation of this in the VGAM R package to the fuel efficiency data gives Figure 1. The family function `binormal(zero = NULL)` was used, which specifies $\eta_1 = \mu_1, \eta_2 = \mu_2, \eta_3 = \log\sigma_{11}, \eta_4 = \log\sigma_{22}, \eta_5 = \log\{(1+\rho)/(1-\rho)\}$, so that the covariances can be modelled with covariates (`weight` and `hp` here). There is substantial agreement between the fitted means and Figures 8(a), (b), (d), and (e), but with the functions decreasing monotonically here as expected. However, the plots (c)–(e), (h)–(j) suggest that the intercept-only assumption of the covariances made by WPS is in doubt; the fitted component functions appear nonlinear and, for example, the variability decreases with increasing `hp`.

In Figure 1, it would be straightforward using VGAM to constrain plots (a), (b) to differ by a constant (known or unknown), and similarly for plots (f), (g), for example,

$$\mathbf{H}_1 = \mathbf{I}_5, \quad \mathbf{H}_2 = \mathbf{H}_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{I}_3 \end{pmatrix}$$

in the unknown case.

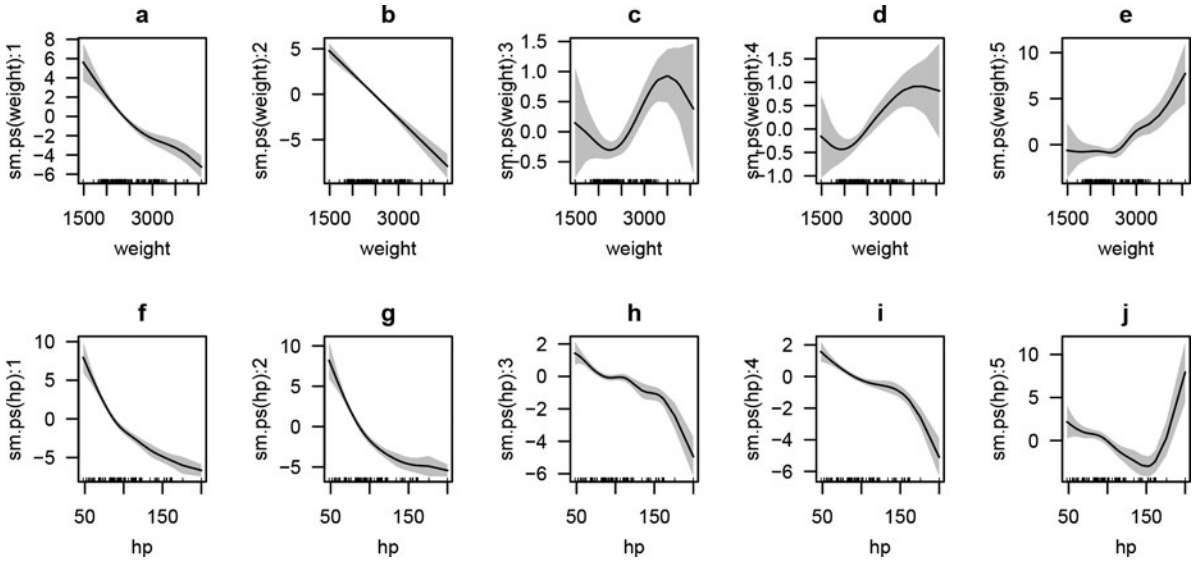

Figure 1. Fitted bivariate normal regression applied to the fuel efficiency data.

2. On Some Appendices

2.1. On Tweedie Models (Appendix J)

Some series expansions could be exploited, for example, for large y the digamma function $\psi(y) = \log y - 1/(2y) - \sum_{k=1}^{\infty} B_{2k}/(2ky^{2k}) \sim \log y - (2y)^{-1} - (12y^2)^{-1}$, where B_k is the kth Bernoulli number. Then $\partial \log W_j / \partial p$ involves the difference between two terms which can make use of this series. Likewise for the $j^2\,\psi'(-j\alpha)$ term involving the trigamma function.

2.2. On Ordered Categorical Models (Appendix K)

While only the logit link is currently implemented for an underlying logistic distribution, there is an argument `link` taking on the value `"identity,"` which is potentially confusing for the user. Ideally the chain rule could be applied more generally so that a variety of link functions could be catered for. Being based on the multinomial distribution, the ordered categorical model should share some of the computational particulars that the multinomial distribution has.

A nonparallel cumulative link model would be worthwhile but quite challenging to implement, and exacerbated by possible intersecting $\eta_j(\boldsymbol{x}_i)$ that results in out-of-range probabilities. Although the $\mathcal{J}$ contraction over $\boldsymbol{x}^k$ technique would be used much to handle the parallel case, it probably would be rather inefficient when the number of levels of the response becomes even moderate.

2.3. On Software Implementation (Appendix M)

The first author's mgcv R package is the most advanced state-of-the-art software for GAMs and the author must be commended for writing and enhancing this over many years. WPS have made a very good start by identifying a few of the most strategic models and implementing those first, such as the Cox model and zero-altered Poisson distribution. As an author of another large GAM-like R package, I can identify with the mgcv developers on the never-ending task of extending and maintaining the software. This explains why there are currently some limitations in a few family functions (which should hopefully be addressed in due course), for example, the negative binomial has an index parameter (called `theta` in the software) that is restricted to intercept-only and/or required to have a known value. Another example is the multivariate normal whose elements of the variance–covariance matrix are also intercept-only (Section 1.1). Section 3.3 mentions offsets for η_j, however mgcv currently seems unable to handle these when there are multiple linear predictors.

At the risk of being branded as an irritant, here are some suggestions that may be useful.

1. Prediction involving nested data-dependent terms fails, for example, `s(scale(x))`. Here, a limitation of *safe* prediction is exposed and one solution is *smart* prediction (Yee, 2015, Sections 8.2.5, 18.6).
2. The ability for family functions to handle multiple responses can be useful, for example, `gam(cbind(y1, y2) ~ s(x2), family = poisson, pdata)`. However, this would entail a considerable amount of work to convert other functions to handle this feature.
3. For `multinom()` and `ocat()` objects, the `fitted()` methods function returns the same result as `predict()`. The fitted probabilities for each class are a more natural type of fitted value.
4. Much attention is given to preserving numerical stability. There are instances where the use of `expm1()` is preferable, as is for `log1p()` too.
5. A specific distribution worth investigating in terms of robustness for handling singularities is the skew normal (Azzalini 1985) where for $Y = \lambda_1 + \lambda_2 Z$, with $0 < \lambda_2$ and $Z \sim \text{SN}(\lambda)$ (whose density is $f(z;\lambda) = 2\,\phi(z)\,\Phi(\lambda z)$ for real z and λ), the expected information matrix as a function of $(\lambda_1, \lambda_2, \lambda)$ is singular as $\lambda \to 0$ even though all three parameters remain identifiable.
6. A final question or two: using the syntax of `gam(list(y1 ~ s(x), y2 ~ s(v), y3 ~ 1, 1+3 ~ s(z)-1), family=mvn(d=3))`

how might one constraint the intercepts of η_1 and η_2 to be equal? Likewise, how might one constraint the intercepts of η_1 to be twice the value of the intercept of η_2?

References

Anderson, J. A. (1984), "Regression and Ordered Categorical Variables," *Journal of the Royal Statistical Society*, Series B, 46, 1–30(with discussion). [1565]

Azzalini, A. A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178. [1567]

Chen, Y., and Samworth, R. J. (2016), "Generalized Additive and Index Models With Shape Constraints," *Journal of the Royal Statistical Society*, Series B, 78, 729–754. [1565]

Wood, S. N. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686. [1566]

——— (2006), *Generalized Additive Models: An Introduction with R*, London: Chapman and Hall. [1566]

Wood, S. N., and Fasiolo, M. (2016), "A Generalized Fellner–Schall Method for Smoothing Parameter Estimation With Application to Tweedie Location, Scale and Shape Models," arXiv:1606.04802. [1566]

Yee, T. W. (2015), *Vector Generalized Linear and Additive Models: With an Implementation in R*, New York: Springer. [1565,1566]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1250580

Comment

Sonja Greven and Fabian Scheipl

Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany

1. Introduction

In their article "Smoothing Parameter and Model Selection for General Smooth Models," Wood, Pya, and Säfken make a significant and impressive contribution to the development of general smooth models and specifically to estimation, inference and model selection for these models. Given the popularity of the R package `mgcv` (Wood 2011) to date, the newly developed methods and their implementation in `mgcv` are likely to shape the routine use of smooth models—for more general model classes than were previously available—in the future.

The developed framework is very flexible and can be used even beyond the models covered in the article. For example, by shifting the functional structure to an appropriate smooth additive predictor (Scheipl, Staicu, and Greven 2015), we have used methods for smooth models (Wood 2011) to develop flexible functional additive mixed models for functional data. The new extensions in the discussed article then allowed us to extend this approach to "generalized" functional data (Greven and Scheipl 2016; Scheipl, Gertheiss, and Greven 2016), where the observations can be thought of as coming from some non-Gaussian process with underlying smoothness assumption and for example negative binomial, t- or Beta marginal distributions. It is a large advantage for researchers and users of flexible regression models that the `mgcv` package provides such a highly performant, innovative and well-documented implementation of state-of-the-art methodology.

While many aspects in the article are worthy of attention, we focus on smoothing parameter uncertainty and applaud the authors' effort to develop methods taking it into account for statistical inference. As covariances and confidence intervals for the regression coefficients are motivated from a Bayesian viewpoint, we initially focus on the implicit prior assumptions and their effects on the posteriors (Section 2) before coming back to the effects of smoothing parameter uncertainty on coefficient uncertainty in Section 3.

To facilitate discussion and intuition, we use a running example of a nonparametric regression model

$$y_i = m(x_i) + \varepsilon_i, \;\; \varepsilon_i \text{ i.i.d.} \, \mathcal{N}(0, \sigma^2), \;\; i = 1, \ldots, n. \qquad (1)$$

In simulations, we take x_i to be equidistant in $[0, 1]$ and the true $m(x_i) = d \exp(2x_i)$ for some d. Let $m(\cdot)$ be parameterized using a penalized spline basis expansion such that the coefficient vector $\boldsymbol{\beta}$ projected into the penalty null-space corresponds to a straight line for $m(\cdot)$, while the projection into the span of the penalty matrix captures deviations from linearity. This example has only one log-smoothing parameter $\rho_1 \equiv \rho$ controlling smoothness of $m(\cdot)$, with $\rho \to \infty$ leading to a linear estimate for $m(\cdot)$ and $\rho \to -\infty$ yielding an unpenalized least squares fit.

2. Prior Assumptions and Posteriors

The authors motivate their approach for smoothing parameter uncertainty and model selection using a Bayesian viewpoint. We thus think it informative to take a look at the implicit prior assumptions of the approach and their effects on the posteriors.

Results in the article are with respect to $\boldsymbol{\rho}$, where each entry is a log-smoothing parameter $\rho_j = \log(\lambda_j)$. We can view the

CONTACT Sonja Greven sonja.greven@stat.uni-muenchen.de Department of Statistics, Ludwig-Maximilians-Universität München, 80539 Munich, Germany.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

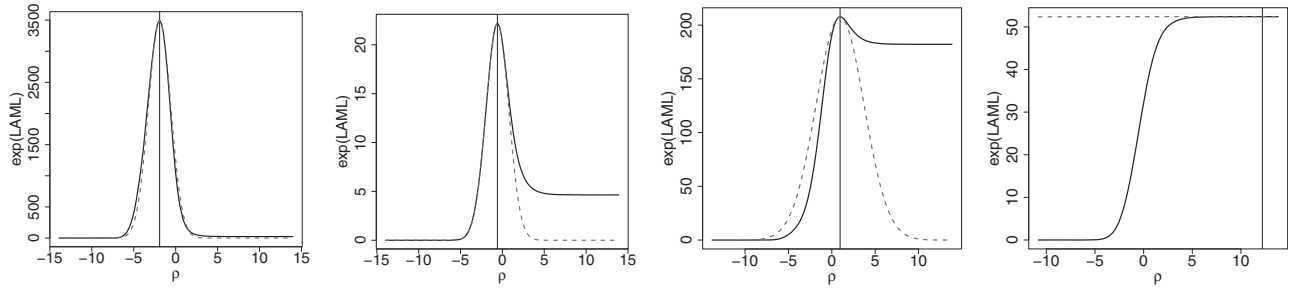

Figure 1. The LAML criterion $\exp(\mathcal{V}(\rho))$ (in black), as a function of ρ for four settings with increasing $\widehat{\rho}$ (vertical lines, as returned by `mgcv`). In the right panel, the LAML is maximized for $\rho \to \infty$. Red dashed lines indicate the normal posterior density for ρ as given by (6) in the article, scaled to have the same maximum value and using $\widehat{\rho}$ and $\boldsymbol{V}_\rho$ as returned by `mgcv` as mean and variance, respectively.

penalized log-likelihood in equation (1) of the article as the joint or the PQL-approximate log-likelihood (Ruppert, Wand, and Carroll 2003, pp. 204–216) of a mixed model with (usually improper) normal prior density $f(\boldsymbol{\beta}|\boldsymbol{\rho}) \propto \exp(-\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}^\lambda \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}}$ (as maximizer of (1)) as posterior mode maximizing $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\rho})$ for a given $\boldsymbol{\rho}$. If there is only one penalty associated with each subvector $\boldsymbol{\beta}_j$ of $\boldsymbol{\beta}$, that is, different blocks are nonzero for different $\boldsymbol{S}^j$, λ_j in $\boldsymbol{S}^\lambda = \sum_j \lambda_j \boldsymbol{S}^j$ corresponds to the inverse of a variance parameter τ_j^2 associated with the random effect $\boldsymbol{\beta}_j$ and controlling deviations of $g_j(\cdot)$ from the null space of the penalty. Estimated model coefficients $\widehat{\boldsymbol{\beta}}$ converge to estimates in the null space of the penalty matrix $\boldsymbol{S}^j$ if $\rho_j \to \infty$ and thus $\lambda_j \to \infty$ or $\tau_j^2 \to 0$.

Note that in a parameterization with respect to τ_j^2, a function $g_j(\cdot)$ can be estimated to lie exactly in the null space of the penalty (e.g., an exactly linear function in our running example) when $\widehat{\tau}_j^2 = 0$ is on the boundary of the parameter space $[0, \infty)$ for τ_j^2. This case corresponds to $\widehat{\rho}_j \to \infty$ in the $\rho_j \in (0, \infty)$ parameterization, which can never be exactly estimated, although the differences in $\widehat{\boldsymbol{\beta}}$ to a very large $\widehat{\rho}_j$ (very small $\widehat{\tau}_j^2$) are negligible. Considering the alternative τ_j^2 parameterization also makes explicit that there is a boundary issue for $\tau_j^2 = 0$ or $\rho_j \to \infty$ that is well-known to cause nonstandard asymptotic behavior, for example, when testing whether $g_j(\cdot)$ lies in the null-space of the penalty (Crainiceanu and Ruppert 2004; Greven and Crainiceanu 2013).

For estimation, the Laplace approximation (LAML) of the log marginal likelihood criterion (2)

$$\log \int f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\rho}) d\boldsymbol{\beta} = \log f(\boldsymbol{y}|\boldsymbol{\rho})$$

is maximized with respect to $\boldsymbol{\rho}$. This can be seen as approximately maximizing the posterior density

$$f(\boldsymbol{\rho}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\rho}) f(\boldsymbol{\rho}) \propto f(\boldsymbol{y}|\boldsymbol{\rho})$$

under independent improper uniform priors $\mathcal{U}(-\infty, \infty)$ for each ρ_j, $f(\boldsymbol{\rho}) \propto 1$. This prior corresponds to an improper prior on $(0, \infty)$ for either λ_j or τ_j^2 with density proportional to $1/\lambda_j$, respectively, $1/\tau_j^2$ and can lead to an improper posterior $f(\boldsymbol{\rho}|\boldsymbol{y})$ (Gelman 2006). Note that the posterior under proper $\mathcal{U}[a_j, b_j]$ priors has the same mode $\widehat{\boldsymbol{\rho}}$ as for improper $\mathcal{U}(-\infty, \infty)$ priors as long as $\widehat{\rho}_j \in [a_j, b_j]$ for all j. However, if the marginal likelihood is monotonically increasing in ρ_j, $\mathcal{U}[a_j, b_j]$ priors yield b_j as the posterior mode for ρ_j and not infinity.

Since the posterior distribution of $\boldsymbol{\rho}$ with $\mathcal{U}[a_j, b_j]$ priors, $a_j, b_j \in [-\infty, \infty]$, has density

$$\begin{aligned} f(\boldsymbol{\rho}|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\rho}) I(a_j \le \rho_j \le b_j \,\forall\, j) \\ &= \exp(\mathcal{V}_r(\boldsymbol{\rho})) I(a_j \le \rho_j \le b_j \,\forall\, j), \end{aligned} \tag{2}$$

where the marginal likelihood $\exp(\mathcal{V}_r(\boldsymbol{\rho}))$ is defined as a function of $\boldsymbol{\rho}$ analogous to equation (2) of the article, it is informative to look at the shape of the Laplace approximation $\exp(\mathcal{V}(\boldsymbol{\rho}))$. Figure 1 shows $\exp(\mathcal{V}(\rho))$ as a function of ρ for four simulations from model (1) with $n = 200$, $d = 0.1$, and $\sigma = 0.5$, estimating a smooth function for $m(\cdot)$ using the `gam` defaults in `mgcv` with `method = "REML"`. From left to right, estimates $\widehat{\rho}$ are increasing. In the rightmost panel, the LAML is maximized for $\rho \to \infty$. For comparison, the normal posterior densities for ρ as given by (6) in the article, using $\widehat{\rho}$ and $\boldsymbol{V}_\rho$ as returned by `mgcv` as mean and variance, are scaled to have the same maximum values and are overlaid in red. We can see that in all cases 1) the nondiminishing mass for $\rho \to \infty$ leads to an improper posterior that cannot be normalized, even though this is hardly visible in the leftmost panel with smallest $\widehat{\rho}$. In particular, the LAML stays practically constant for increasing ρ after a certain point that roughly results in a linear fit for $m(\cdot)$. 2) The normal approximation of equation (6) is reasonable locally near $\widehat{\rho}$ in the three leftmost panels, but does not hold for $\rho \to \infty$ or for the strongly regularized function estimate in the rightmost panel.

For the conditional posterior of $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}$, we know that it is asymptotically Gaussian $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ for each given $\boldsymbol{\rho}$, cf. equation (5) of the discussed article, where we have made the dependence of $\boldsymbol{V}_\beta$ on $\boldsymbol{\rho}$ explicit in our notation. The asymptotic marginal posterior of $\boldsymbol{\beta}|\boldsymbol{y}$ thus is generated similarly to a scale mixture of multivariate normal distributions with density

$$f(\boldsymbol{\beta}|\boldsymbol{y}) = \int f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}) f(\boldsymbol{\rho}|\boldsymbol{y}) d\boldsymbol{\rho}, \tag{3}$$

where $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ is the $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ density and the approximate shape of $f(\boldsymbol{\rho}|\boldsymbol{y})$ is shown in Figure 1.

To look at the marginal posterior of $\boldsymbol{\beta}$, we use the `jagam` function (Wood 2016) in `mgcv`, which allows to extract model specifications of a generalized additive model (GAM) and to fit the model using full Bayesian inference via MCMC in JAGS (Plummer 2016). We simulate again from running example (1) with $n = 100$ and $\sigma = 1$. The prior for the log-smoothing parameter ρ is set to $\rho \sim \mathcal{U}[-12, 12]$, since only proper priors are possible here, but estimates $\widehat{\rho}$ are always smaller than 12. Figure 2 shows a strongly nonlinear fit for $d = 2$ (top), and an

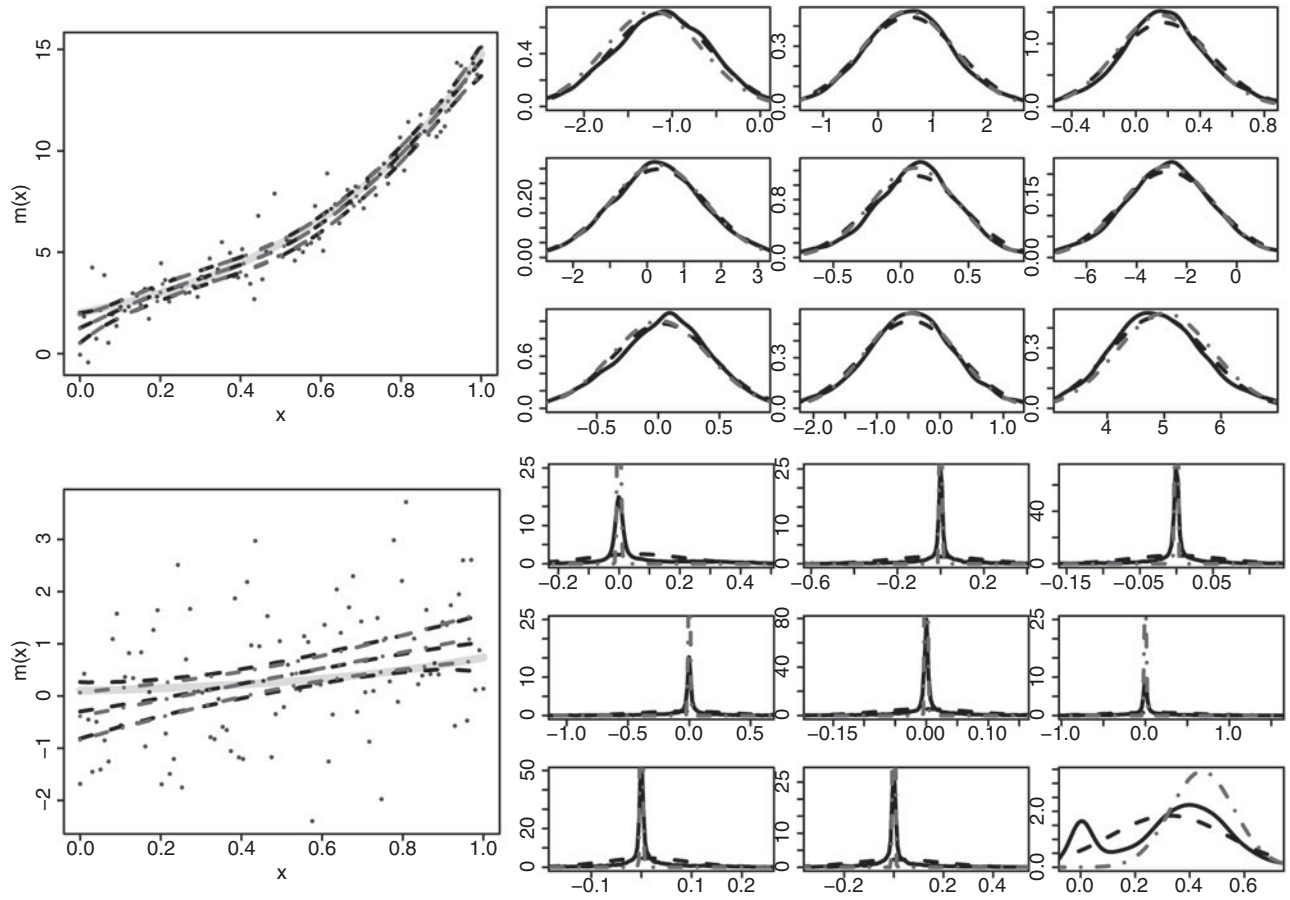

Figure 2. The marginal posterior densities of $\boldsymbol{\beta}|\boldsymbol{y}$ from `jagam`, separately for each component of $\boldsymbol{\beta}$ (nine small panels, in black), with normal distributions fitted to the posterior sample overlaid in dashed black, and the Gaussian approximation based on the smoothing parameter uncertainty corrected covariance $\boldsymbol{V}'_\beta$ from `mgcv` in dashed-dotted red for the estimate shown in the large panel (Bayesian estimate and credible intervals in dashed black, estimate and confidence interval from `mgcv` in dashed-dotted red, true function in grey, data as dots). Shown are two simulations from running example (1) with strong nonlinearity $d = 2$ (top) and small nonlinearity $d = 0.1$ (bottom). Vertical axis cropped in some panels for clarity.

almost linear fit for $d = 0.1$ (bottom). Note that the two estimates (left panels) from `mgcv` and `jagam` are not identical, as `mgcv` uses the posterior mode of $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ at the posterior mode $\widehat{\boldsymbol{\rho}}$ of $f(\boldsymbol{\rho}|\boldsymbol{y})$, while `jagam` uses the posterior mean of $f(\boldsymbol{\beta}|\boldsymbol{y})$. Confidence/credible bands and the posterior densities (right panels) are however based on the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ for both, with differences due to the normal approximation of the marginal posterior in `mgcv`. We can see that the posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ is close to normal in the strongly nonlinear setting, where $\widehat{\rho}$ is relatively small. For the almost linear estimate, however, where $\widehat{\rho}$ is large, most coefficients show a spiky posterior with most mass close to zero and heavier tails than a Gaussian, while one coefficient exhibits a bimodal distribution. Here, the normal approximation is not very close and the resulting confidence band (bottom left) of `mgcv` is noticeably narrower than the credible band from `jagam`.

3. Smoothing Parameter Uncertainty

Consider now the posterior covariance of $\boldsymbol{\beta}$, which is used for confidence band construction and in the conditional AIC. The usual uncorrected covariance $\boldsymbol{V}_\beta = \boldsymbol{V}_\beta(\boldsymbol{\rho})$ is based on the conditional posterior $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ with $\widehat{\boldsymbol{\rho}}$ plugged in. The authors propose a corrected covariance $\boldsymbol{V}'_\beta$ to account for smoothing parameter uncertainty, based on approximating the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ using a linear Taylor expansion.

Equation (6) of the article postulates an asymptotic posterior distribution $\boldsymbol{\rho}|\boldsymbol{y} \sim N(\widehat{\boldsymbol{\rho}}, \boldsymbol{V}_\rho)$ in the interior of the parameter space, where $\boldsymbol{V}_\rho$ is the inverse of the Hessian of the negative log marginal likelihood $-\mathcal{V}_r$ with respect to $\boldsymbol{\rho}$. The boundary case corresponds to $\widehat{\rho}_j \to \infty$, with $\widehat{\rho}_j$ values treated as "working infinity" during estimation when $\partial\mathcal{V}/\partial\rho_j \approx \partial^2\mathcal{V}/\partial\rho_j^2 \approx 0$. In this case, the authors propose to substitute a Moore–Penrose pseudoinverse of the Hessian. Consequences are most easily discussed for the case of a scalar ρ. When $\widehat{\rho} \to \infty$, the Hessian of the negative log marginal likelihood is $-\partial^2\mathcal{V}_r/\partial\rho^2|_{\rho=\widehat{\rho}} \approx -\partial^2\mathcal{V}/\partial\rho^2|_{\rho=\widehat{\rho}} \to 0$. The Moore-Penrose pseudoinverse of 0 is again 0, that is, $\boldsymbol{V}_\rho \to 0$. Thus, the posterior for ρ collapses to a point mass at $\widehat{\rho} \approx \infty$ (corresponding to a point mass at $\tau^2 \approx 0$). In this case, $\boldsymbol{J}$ and $\frac{\partial}{\partial\rho}\boldsymbol{R}_\rho$ also converge to $\boldsymbol{0}$, formula (7) gives $\boldsymbol{V}'_\beta \to \boldsymbol{V}_\beta$ and the corrected and uncorrected covariances coincide. This is also practically confirmed in simulations for our running example: When the estimated effective degrees of freedom approach 2, the differences in the two estimated covariance matrices approach zero.

Thus, smoothing parameter uncertainty is not accounted for near the boundary of the parameter space, where coverage is most of an issue (Marra and Wood 2012). For very large smoothing parameters, confidence bands are effectively based on the

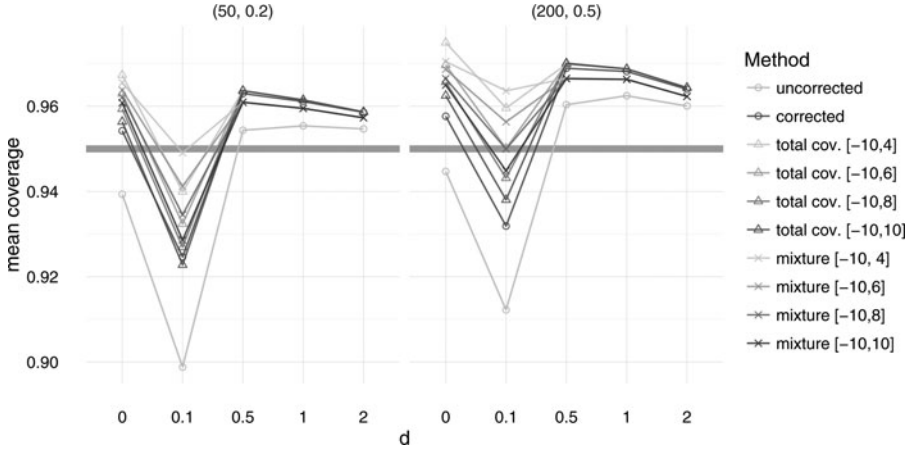

Figure 3. Average coverage across [0, 1] for nominal level $(1-\alpha) = 0.95$ in running example (1) with $(n, \sigma) = (50, 0.2)$ (left) and $(200, 0.5)$ (right). Coverages based on $\boldsymbol{V}_\beta$ and $\boldsymbol{V}'_\beta$ (circles) are denoted by uncorrected and corrected, respectively. The prior supports for the total covariance (triangles) and mixture approximation (crosses) methods are given in the legend.

submodel defined by the penalty nullspace. For example, in our running example, if $m(\cdot)$ is estimated to be linear, confidence bands are computed based on the linear submodel, ignoring the possibility of truly nonlinear functions. Confidence bands are too narrow and linear in this case. This leads to undercoverage for functions $m(\cdot)$ that are not linear, but close enough that there is a relevant probability of estimating a linear $\widehat{m}(\cdot)$, see Figure 3. For strongly nonlinear functions, linear estimates rarely occur and do not noticeably change coverage. For truly linear functions, computing confidence bands based on a linear submodel also does not lead to undercoverage. Thus, while coverage is not affected in most cases, for functions that are only slightly nonlinear relative to the noise level, there is a noticeable dip in coverage ("phase transition" between nonlinear and linear models).

Since the linear Taylor approximation does not work well on the boundary of the parameter space for $\widehat{\rho}_j \to \infty$, we sketch two possible alternatives. To work with a normal approximation for the marginal posterior of $\boldsymbol{\beta}$ that incorporates uncertainty in $\boldsymbol{\rho}$, we can use the law of total covariance

$$\begin{aligned}\operatorname{cov}(\boldsymbol{\beta}|\boldsymbol{y}) &= \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\operatorname{cov}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho})] + \operatorname{cov}_{\boldsymbol{\rho}|\boldsymbol{y}}[\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho})] \\ &= \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\boldsymbol{V}_\beta(\boldsymbol{\rho})] + \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho \widehat{\boldsymbol{\beta}}_\rho^\top] - \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho]\,\mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho]^\top,\end{aligned}$$

where expectations are with respect to the posterior distribution $f(\boldsymbol{\rho}|\boldsymbol{y})$ given in (2).

Alternatively, abandoning the normality assumption which might be questionable near the boundary, see Figure 2, we can directly use (3) to write $f(\boldsymbol{\beta}|\boldsymbol{y})$ as a continuous mixture of $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ densities. As $\boldsymbol{\beta}$ is multi-dimensional and we are usually interested in linear combinations of $\boldsymbol{\beta}$, it is easier to work with $f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y}) = \int f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho}) f(\boldsymbol{\rho}|\boldsymbol{y}) d\boldsymbol{\rho}$ for some relevant vector $\boldsymbol{x}$, which is a one-dimensional scale mixture of $\mathcal{N}(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{x}^\top\boldsymbol{V}_\beta(\boldsymbol{\rho})\boldsymbol{x})$ densities with mixing distribution given in (2).

To compute pointwise level $(1-\alpha)$ confidence bands for either approach, we first define a grid $\boldsymbol{\rho}_r$, $r = 1, \ldots, R$, of values covering the prior domain, for example, for a one-dimensional ρ a grid covering the interval $[a, b]$. We then compute weights $w(\boldsymbol{\rho}_r) = \exp(\mathcal{V}(\boldsymbol{\rho}_r))/\sum_{r=1}^R \exp(\mathcal{V}(\boldsymbol{\rho}_r))$ to use in numerical integration with respect to the posterior $f(\boldsymbol{\rho}|\boldsymbol{y})$. This requires R refits of the model with $\boldsymbol{\rho}$ fixed at $\boldsymbol{\rho}_r$, $r = 1, \ldots, R$, to obtain $\exp(\mathcal{V}(\boldsymbol{\rho}_r))$ as the LAML for this model fit. For the first approach, we can then approximate the total covariance as

$$\begin{aligned}\operatorname{cov}(\boldsymbol{\beta}|\boldsymbol{y}) \approx \sum_{r=1}^R w(\boldsymbol{\rho}_r)\left[\boldsymbol{V}_\beta(\boldsymbol{\rho}_r) + \widehat{\boldsymbol{\beta}}_{\rho_r}\widehat{\boldsymbol{\beta}}_{\rho_r}^\top\right] \\ - \left[\sum_{r=1}^R w(\boldsymbol{\rho}_r)\widehat{\boldsymbol{\beta}}_{\rho_r}\right]\left[\sum_{r=1}^R w(\boldsymbol{\rho}_r)\widehat{\boldsymbol{\beta}}_{\rho_r}\right]^\top,\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{\rho_r}$ and $\boldsymbol{V}_\beta(\boldsymbol{\rho}_r)$ are obtained from the model output as the estimate and uncorrected covariance when refitting the model with $\boldsymbol{\rho}$ fixed at $\boldsymbol{\rho}_r$. For the second approach, we can approximate $f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y})$ by a discrete mixture of $\mathcal{N}(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_{\rho_r}, \boldsymbol{x}^\top\boldsymbol{V}_\beta(\boldsymbol{\rho}_r)\boldsymbol{x})$ densities, with mixture weights again given by the $w(\boldsymbol{\rho}_r)$. The $\alpha/2$ and $(1-\alpha/2)$ quantiles for this mixture can then be obtained, for example, using the R package `nor1mix` (Mächler 2016).

Note that both approaches use the posterior density $f(\boldsymbol{\rho}|\boldsymbol{y})$, which is only proper if finite prior limits $[a_j, b_j]$ are used. We have seen in Section 2 that these prior limits, if chosen sufficiently large, do not have a strong influence on the model fit: if there is a maximum of the LAML within these bounds, it is not affected by them, and if the LAML is monotonically increasing for ρ_j, the estimate $\widehat{\boldsymbol{\beta}}$ is hardly affected by further increasing ρ_j once it approximately lies in the null space of the penalty. When using the posterior, however, these chosen boundaries do make a difference. Increasing b_j puts more prior weight on functions $g_j(\cdot)$ close to the penalty null space, while decreasing b_j puts more weight on functions deviating from the null space, for example, on linear and nonlinear $m(\cdot)$ in our running example, respectively. Thus, the two proposed alternatives also depend on the chosen prior limits.

We compared the different methods in a small simulation study for our running example with $n \in \{50, 200\}$, $\sigma \in \{0.2, 0.5\}$, and $d \in \{0, 0.1, 0.5, 1, 2\}$ and worked with the prior limits $a = -10$ and $b \in \{4, 6, 8, 10\}$. (Results were not found to be sensitive to the lower limit here, as $\exp(\mathcal{V}(\rho))$ quickly decreases toward zero for small ρ values.) For the grid ρ_r we used $\hat{\rho} + \ln(10)k$ with $k = (-10, -9.9, \ldots, 9.9, 10)$, that is, an equidistant grid centered on $\hat{\rho}$, truncated to $[a, b]$.

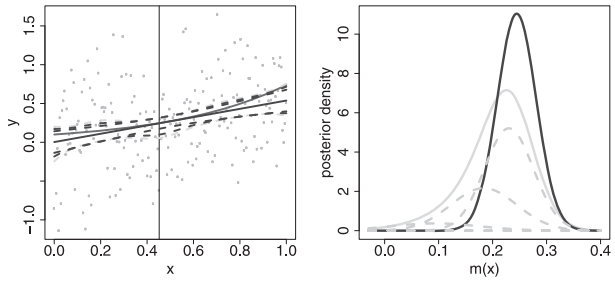

Figure 4. Left: data (gray dots) simulated from the running example with $n = 200$, $d = 0.1$, and $\sigma = 0.2$, true function $m(\cdot)$ (solid red), estimate and pointwise confidence band based on the corrected covariance (green) and confidence bands based on the mixture approximation for $a = -10$ and $b = 10$ (dark blue) or $b = 4$ (light blue). Right: for the x marked by a vertical line on the left, posterior density as given by the normal approximation with corrected covariance (green) and the approximation by a normal mixture distribution with $b = 4$ (light blue). Dashed gray lines show aggregated mixture components with average means and standard deviations as well as total weights within each fifth of the grid for ρ.

Figure 3 shows results for average coverage across [0, 1] for 12,000 replications for $(n, \sigma) = (50, 0.2)$ and $(200, 0.5)$. There is a clear pattern of undercoverage for the uncorrected covariance $\boldsymbol{V}_\beta$ for $d = 0.1$, which is only partly corrected by using the corrected covariance $\boldsymbol{V}'_\beta$. For these two settings, $d = 0.1$ is only slightly nonlinear compared to the noise level, resulting in a high percentage (29% respectively 41%) of (near-)linear estimates with estimated degrees of freedom below 2.1. For $d = 0$, computing confidence bands in the linear subspace leads to nominal or overcoverage, as does the normal approximation for large d values. Using either the total covariance or the mixture approximation to $f(\boldsymbol{\beta}|\boldsymbol{y})$ gives very similar coverage to the corrected covariance independent of b in the case of large d values, where estimates are almost never linear and the normal approximation works well. For small $d = 0.1$, undercoverage is improved compared to the corrected covariance, at the cost of some overcoverage for $d = 0$ (on the order of the overcoverage all methods show for larger d values). Coverage is dependent on upper limit b of the uniform prior for ρ, with smaller b placing more weight on nonlinear functions and thus leading to higher coverage values. We interpret the undercoverage we observe for some large b, that is, poorly calibrated credible/confidence intervals for $m(\cdot)$, as a result of prior-data-conflict in these settings: large values of b put an inordinate amount of prior weight on (approximately) linear functions, while the observations come from a nonlinear data-generating process. For $(n, \sigma) = (200, 0.2)$, near-linear estimates occur in only 2% of simulations for $d = 0.1$ due to the higher information content of the data, leading to no undercoverage using $\boldsymbol{V}'_\beta$ and no advantages of the alternative methods. The dip in coverage for $d = 0.1$ is also smaller for $(n, \sigma) = (50, 0.5)$, possibly due to the wide confidence bands in this setting with low signal-to-noise ratio. Based on our limited four simulations, the mixture approximation with a low b value seems to give coverage closest to constant across different values of d. Computing times for the confidence intervals based on the mixture approximation were 2.5–3.5 seconds on a laptop computer for $n = 200$ depending on b and without grid optimization or parallelization.

Figure 4 illustrates the corrected covariance and mixture approximation confidence intervals for our running example. The left panel shows the linear estimate for $m(\cdot)$ and the pointwise confidence band based on the corrected covariance in green. Confidence bands based on the mixture approximation with $a = -10$ and $b = 10$ or $b = 4$ are shown in dark and light blue, respectively. It is clear that these acknowledge the possibility of the true function being nonlinear and are wider in particular in the middle and toward the ends of the [0, 1] interval. This leads to increased coverage of the truly nonlinear function. The right panel shows the mixture approximation to the posterior of $m(x_i)$ for an example x_i in the middle of the interval. Uncertainty in ρ here leads to mixture components (for smaller ρ values) with smaller mean and larger variance. This results in a skewed posterior with long lower tail and thus in a smaller lower bound of the confidence/credible interval. It is also clear from the different locations of the maxima that there is a difference between the maximum of the marginal posterior density $f(\boldsymbol{\beta}|\boldsymbol{y})$ (light blue) and the maximum of the conditional posterior density $f(\boldsymbol{\beta}|\boldsymbol{y}, \widehat{\rho})$ (green). Thus, in addition to not being symmetrical, the interval based on the mixture approximation is also not centered on the same value as that based on the corrected covariance.

4. Summary

The proposal by Wood, Pya, and Säfken is an important step in the direction of accounting for smoothing parameter uncertainty in inference for $\boldsymbol{\beta}$. While it works well in most settings, for the "phase transition" between functions in the null space of the penalty and those far from the null space it does not lead to well calibrated inference. Although the problem does not seem to be extremely large, we did see some undercoverage in confidence bands due to the neglect of smoothing parameter uncertainty when the function is effectively estimated to lie in the null space of the penalty ($\widehat{\rho}_j \to \infty$ case).

We have discussed two alternative approaches to the problem, which seem to improve undercoverage occurring in some of the cases we looked at. To more fully develop either approach would require a more efficient way to choose a suitable grid with the smallest possible number of grid points and a much wider simulation study including models with more than one smoothing parameter and other response distributions. Preliminary results for a simple logistic GAM indicate that the two discussed alternatives seem to be transferable in principle. We also saw that results remain sensitive to the chosen prior limits of the uniform priors for the log-smoothing parameters $\boldsymbol{\rho}$. The best combination of the discussed alternatives and the Wood, Pya, and Säfken approach might be to only compute the mixture approximation when the estimate is close to the penalty null-space, as the corrected covariance seems to work well in all other cases. We think it remains worthwhile to think about alternative approaches for smoothing parameter uncertainty near the boundary as well as the relationship to model selection via the newly proposed conditional AIC, for which the corrected covariance is also used.

References

Crainiceanu, C. M., and Ruppert, D. (2004), "Likelihood Ratio Tests in Linear Mixed Models With One Variance Component," *Journal of the Royal Statistical Society*, Series B, 66, 165–185. [1569]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models" (comment on article by Browne and Draper), *Bayesian Analysis*, 1, 515–534. [1569]

Greven, S., and Crainiceanu, C. M. (2013), "On Likelihood Ratio Testing for Penalized Splines," *AStA Advances in Statistical Analysis*, 97, 387–402. [1569]
Greven, S., and Scheipl, F. (2016), "A General Framework for Functional Regression Modelling," *Statistical Modelling*, to appear. [1568]
Mächler, M. (2016), *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*, available at *http://CRAN.R-project.org/package=nor1mix*. R package version 1.2-2. [1571]
Marra, G., and Wood, S. N. (2012), "Coverage Properties of Confidence Intervals for Generalized Additive Model Components," *Scandinavian Journal of Statistics*, 39, 53–74. [1570]
Plummer, M. (2016), *rjags: Bayesian Graphical Models using MCMC*, available at *https://CRAN.R-project.org/package=rjags*. R package version 4-6. [1569]
Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, UK: Cambridge University Press. [1569]
Scheipl, F., Staicu, A.-M., and Greven, S. (2015), "Functional Additive Mixed Models," *Journal of Computational and Graphical Statistics*, 24, 477–501. [1568]
Scheipl, F., Gertheiss, J., and Greven, S. (2016), "Generalized Functional Additive Mixed Models," *Electronic Journal of Statistics*, 10, 1455–1492. [1568]
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, 73, 3–36. [1568]
Wood, S. N. (2016), "Just Another Gibbs Additive Modeller: Interfacing Jags and mgcv," *arXiv:1602.02539*. [1569]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1250583

Rejoinder

Simon N. Wood[a], Natalya Pya[b], and Benjamin Säfken[c]

[a]School of Mathematics, University of Bristol, Bristol, UK; [b]KIMEP University, Almaty, Kazakhstan; [c]Georg-August-Universität Göttingen, Göttingen, Germany

The Boundary of the Smoothing Parameter Space

Thomas Kneib, Sonja Greven, and Fabian Scheipl make insightful comments on the problems associated with smoothing parameters on the parameter space boundary, where our smoothing parameter uncertainty correction does not hold, with the approximate posterior for the log smoothing parameters then being improper. For the AIC correction proposed in the article, we are effectively correcting the AIC for the larger model only when it does not coincide with the smaller model, so that its smoothing parameters are not estimated on the boundary. When both model estimates coincide then being able to correct for smoothing parameter uncertainty in the parameters on the boundary is anyway irrelevant.

However, as Sonja Greven and Fabian Scheipl point out in their impressive discussion contribution, interval estimates can be very suboptimal when smoothing parameters are estimated to be on the boundary, if no correction for smoothing parameter uncertainty is then made. Their proposed fixes offer a substantial improvement at the "phase transition" from completely smooth estimates to estimates with nonzero smoothing penalty, but are also relatively complex to implement in general. Picking up their basic approach, a very simple alternative is to base the smoothing parameter uncertainty correction on a model that is statistically indistinguishable from the estimated model, but for which the approximate posterior for the log smoothing parameters is proper.

This is very easy to implement with Newton-based optimization. We can identify smoothing parameters on the edge of the feasible parameter space as those for which $|\partial \mathcal{V}/\partial \rho_i| \approx |\partial^2 \mathcal{V}/\partial \rho_i^2|$ (and ρ_i is large) at Newton method convergence. Those so identified can then be reduced until the log RE/ML has changed by some small target amount that can be treated as statistically insignificant: for example, 0.1 or 0.01. Clearly, statistically we cannot distinguish the resulting model from the best fit, but the Hessian of $\mathcal{V}$ will now have full rank (otherwise we would not have actually changed the RE/ML). The corrected covariance matrix for the model coefficients can now be computed from this statistically indistinguishable model. Obviously, it is a bad idea to decide between models when one of them has been shifted in a way that increases the difference between the models, so we do not recommend this approach for computing a corrected AIC.

Figure 1 shows the results of replicating the small study that leads to Figure 3 of Greven and Scheipl, using our simple proposal (which adds less than 5% to the computing time). The over coverage of the corrected intervals seen in both figures is to be expected, given the Nychka (1988) result suggesting close to nominal coverage *without* accounting for smoothing parameter uncertainty: clearly increasing the width of the intervals can only lead to higher coverage. Like Greven and Scheipl's proposals, our simple proposal looks promising, and we agree that the whole topic merits further study.

Parameterization, Propriety, and Invariance

Thomas Kneib's query about whether our uncertainty corrections are parameterization invariant and Sonja Greven and

CONTACT Simon Wood *simon.wood@bristol.ac.uk* School of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom.

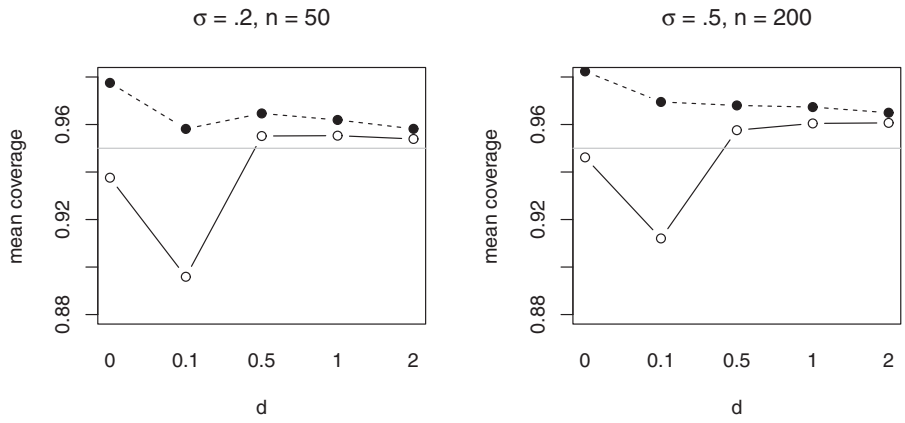

Figure 1. Replication of Greven and Scheipl Figure 3, with the simple alternative boundary correction proposed in the text (dashed and solid symbols). The uncorrected alternative is shown as open symbols and continuous lines.

Fabian Scheipl's comments about improper posteriors are also somewhat related. An appealing parameterization would be in terms of $\alpha_j = \text{tr}\{(\mathcal{I} + \mathbf{S}^\lambda)^{-1}\lambda_j \mathbf{S}^j\}$, which have an interpretation as the "suppressed degrees of freedom" per penalty, and for which a uniform prior is proper since the α_j are bounded between 0 and the rank of $\mathbf{S}^j$ (for smooths with multiple penalties there are also restrictions on the sum of the α_j). For a singly penalized smooth with basis dimension k_j the effective degrees of freedom is just $k_j - \alpha_j$. So, this parameterization does away with the possibility of a formally improper posterior, while also operating on the degrees of freedom scale on which there is some hope of specifying meaningful priors, should something other than uniform be desired (and it is easy to incorporate such priors into estimation and posterior inference). Of course a Taylor approximation at the parameter space boundary is still not helpful.

None the less, the derivatives $\partial\alpha_j/\partial\rho_k$ are readily computed given our article's methods, so that a matrix inversion yields the Jacobian terms $\partial\rho_k/\partial\alpha_j$. Transforming the Hessian of the LAML, we obtain

$$\mathcal{V}^{ij}_{\alpha\alpha} = \frac{\partial\rho_l}{\partial\alpha_j}\mathcal{V}^{kl}_{\rho\rho}\frac{\partial\rho_k}{\partial\alpha_i} + \mathcal{V}^k_\rho\frac{\partial^2\rho_k}{\partial\alpha_i\partial\alpha_j}.$$

So, our correction (7) will only match in the α and ρ parameterizations if we neglect the $\mathcal{V}^k_\rho\partial^2\rho_k/\partial\alpha_i\partial\alpha_j$ term in the Hessian transform: the correction is not generally invariant. Interestingly if we do use α parameterization with a (proper) uniform prior, then to ensure that the Jacobian can be inverted numerically requires moving the α_j an "insignificant distance" back from the boundary if it lies on it, as in the simple boundary correction above. If we also neglect the $\mathcal{V}^k_\rho\partial^2\rho_k/\partial\alpha_i\partial\alpha_j$ term in the Hessian transformation then we end up with the same correction as in the simple method discussed earlier.

Bayes, Calibration, Better Models, and Implementational Cost

Thomas Kneib's comments on frequentist versus Bayesian inference also deserve comment. Our view is that the distinction is somewhat blurred in smoothing, given the well-established equivalence of quadratically penalized smoothers and (intrinsic) Gaussian random fields, especially when using marginal likelihood for smoothing parameter estimation. We view our approach as "empirical Bayes" and think that the real distinction is rather between approximate-direct and "exact"-simulation based computational strategies. Where we are firmly frequentist in outlook is in wanting well calibrated inference in the frequentist sense. It is then often the case that the conditions required for the fully Bayesian approach to be well calibrated are the same as those required for our approximate posterior to be a good approximation. If it is, then it is straightforward to use the approximation for direct and rapid posterior simulation, and then to generate samples from the posterior of any quantity predicted by the model. That said there are many cases in which the Bayesian simulation approach is superior, both in speed of implementation and in the flexibility of the random effects structures that it readily handles. (Thomas Kneib's final question under this heading is much too difficult and we do not have a general answer.)

Rather neatly, Thomas Yee's Figure 1 and the associated discussion nicely shows that Thomas Kneib's comments on the limitations of our simple example multivariate Gaussian model are spot on, and a more flexible structure would offer some improvements. We agree however, that beyond two dimensions the challenges of coming up with an interpretable model structure are considerable.

Thomas Yee makes a worthwhile and interesting point about the costs and benefits of the method. In part, the complexity of the method is what gives access to the quantities that allow us to correct for smoothing parameter uncertainty and substantially improve AIC performance: this we think is worthwhile in practical data analysis. Second, the method structure is not driven by a desire for complexity, but from the numerous practical examples of convergence failures that the first author accumulated from users of the `mgcv` software when it was based exclusively on the type of working model approach described in Wood (2004). The problem with estimating smoothing parameters for iteratively recomputed working linear models is that there is generally no guarantee of convergence (a problem that seemed to get worse as one moves from the exponential family to the more complex models, such as the GAMLSS class, see Wood 2005). That said, the Reiss and Ogden (2009) demonstration of REML's decreased tendency to multiple minima relative to GCV does suggest that REML based estimation in such schemes should be less

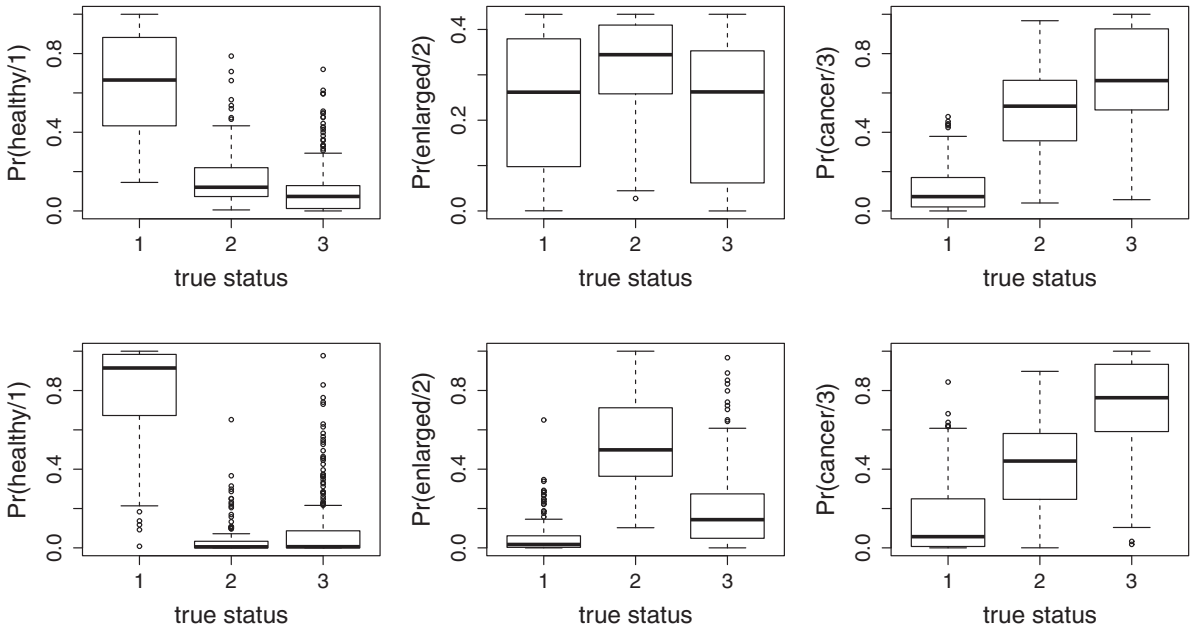

Figure 2. Classification results for the ordered categorical (upper) and multinomial logit (lower) models for the prostate screening example. Boxplots are of probabilities of each disease category according to the model, for each actual disease category. Clearly, the multinomial model (with two signal regression linear predictors) gives superior results.

problematic. It will be interesting to see how the VGAM implementation works out in this respect. It is also worth noting that substantial parts of the work of implementing new distributions for GAMLSS type models can be automated. For example, the 3rd and 4th order derivatives of the generalized extreme value distribution were so horrifying that a morning was spent producing an automatic method for generating them in R. This consisted of exporting the derivatives from Maxima as Maxima expressions, reading these into R and auto-translating into R code, and then auto-simplifying the code. The `gevlss` family in `mgcv` is the result. The auto-generated derivative code required only limited hand modification for stability and efficiency.

On the other points that Thomas Yee raises: `mgcv` offers rather limited single index model support at present, and it is true that the general framework could offer something rather more efficient. The point about the prostate screening model is also right. Figure 2 shows the classification rates for the ordered categorical model presented in the article (top row) and for a logistic multinomial model with separate signal regression linear predictors for the enlarged and cancerous classes. The latter is clearly superior.

The other points raised perhaps require no rejoinder, but are much appreciated and we thank the discussants for them.

References

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143. [1573]

Reiss, P. T., and Ogden, T. R. (2009), "Smoothing Parameter Selection for a Class of Semiparametric Linear Models," *Journal of the Royal Statistical Society*, Series B, 71, 505–523. [1574]

Wood, S. (2005), Discussion of "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society*, Series C, 54, 545–546. [1574]

——— (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686. [1574]

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

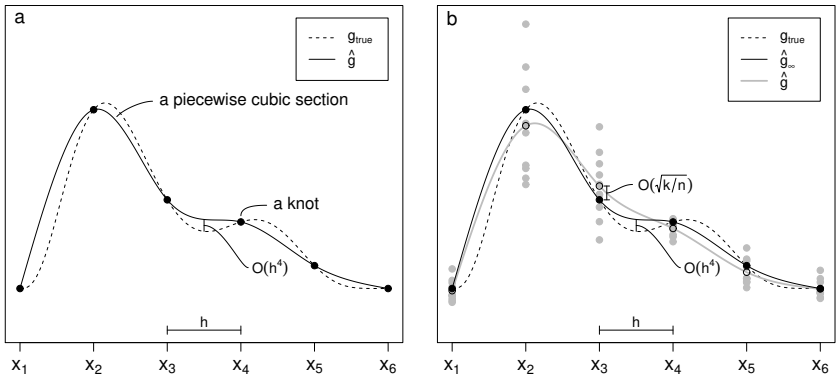

Figure 1: **a.** A cubic interpolating spline (continuous curve), interpolating 6 'knot' points (black dots) with evenly spaced x co-ordinates, from a true function (dashed curve). The spline is made up of piecewise cubic sections between each consecutive pair of knots. The approximation error is $O(h^4)$, 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 $\mathcal{K}^m \phi_j(x) = \Lambda_j \phi_j(x)$,

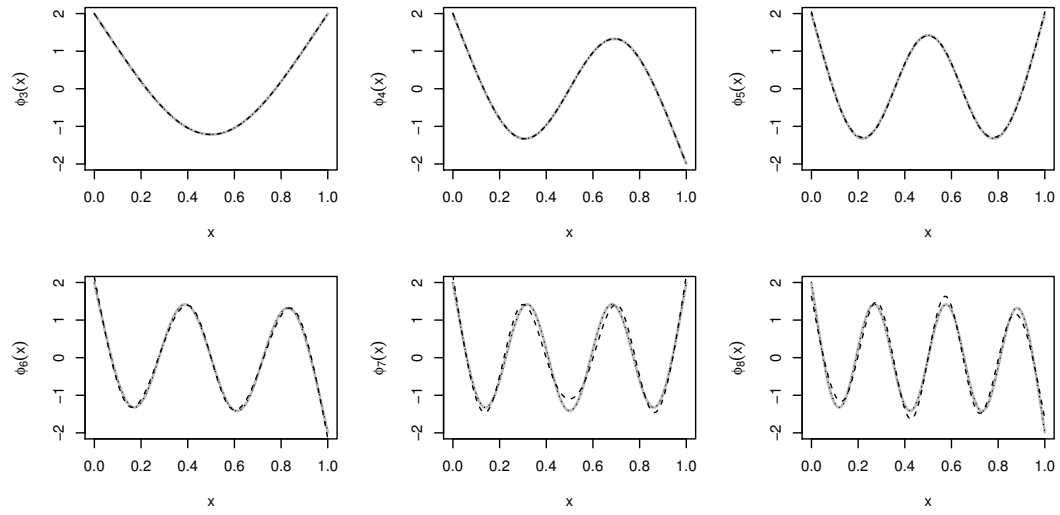

Figure 2: Eigenfunctions of $\mathcal{K}^2$ shown in grey, with Demmler-Reinch spline basis functions overlaid in black. The first two linear functions are not shown. The dashed curves are for a rank 8 cubic spline basis, while the dotted curve, exactly overlaying the grey curves, are for a rank 16 cubic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

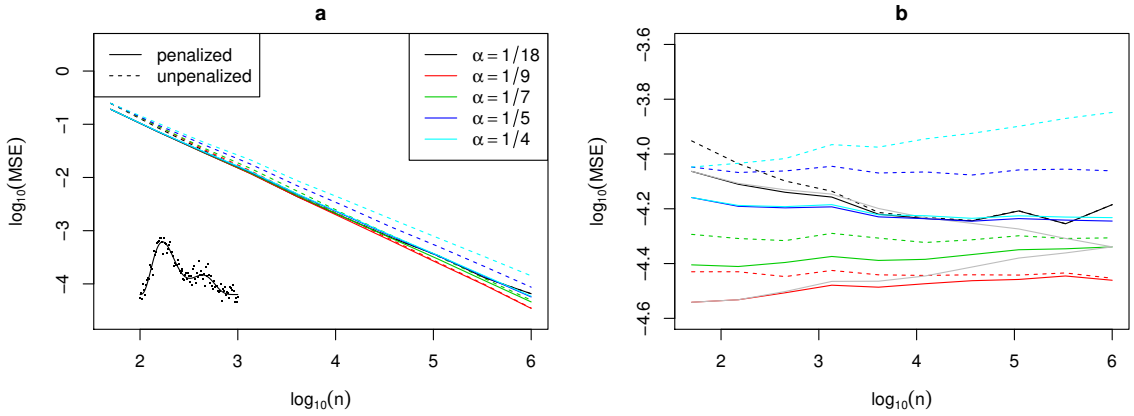

Figure 3: **a.** Example of MSE convergence for simple Gaussian smoothing. The true function is shown at lower left, with 100 noisy samples also shown. The coloured lines show log MSE averaged over 100 replicates against log sample size when the basis size $k \propto n^\alpha$ for various α values (all 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

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

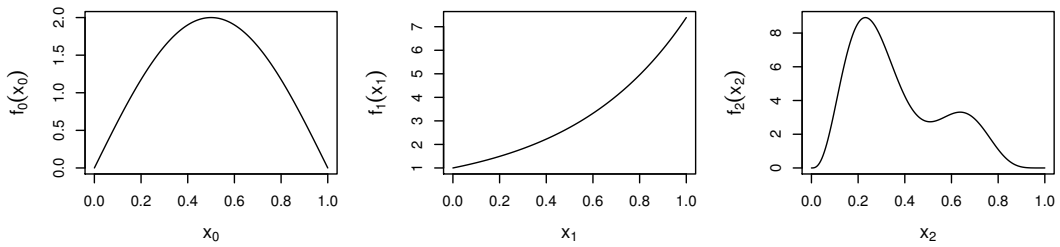

Figure 4: Shapes of the functions used for the simulation study (from Gu and Wahba, 1991). $f_3(x_3) = 0$.

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

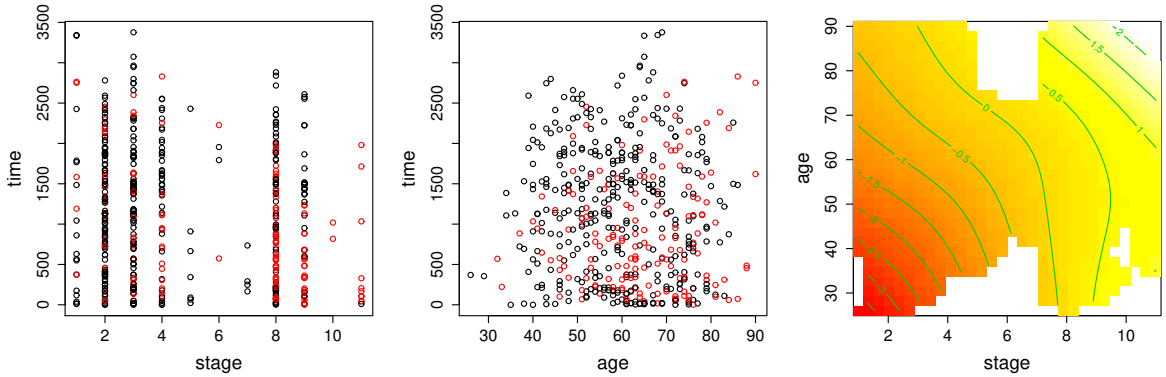

Figure 5: Left: Survival times (red) and censoring times (black) against disease stage for patients with kidney renal clear cell carcinoma. Middle: times against patient age. Right: the combined smooth effect of age and stage on the linear predictor scale from a Cox Proportional hazards survival model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

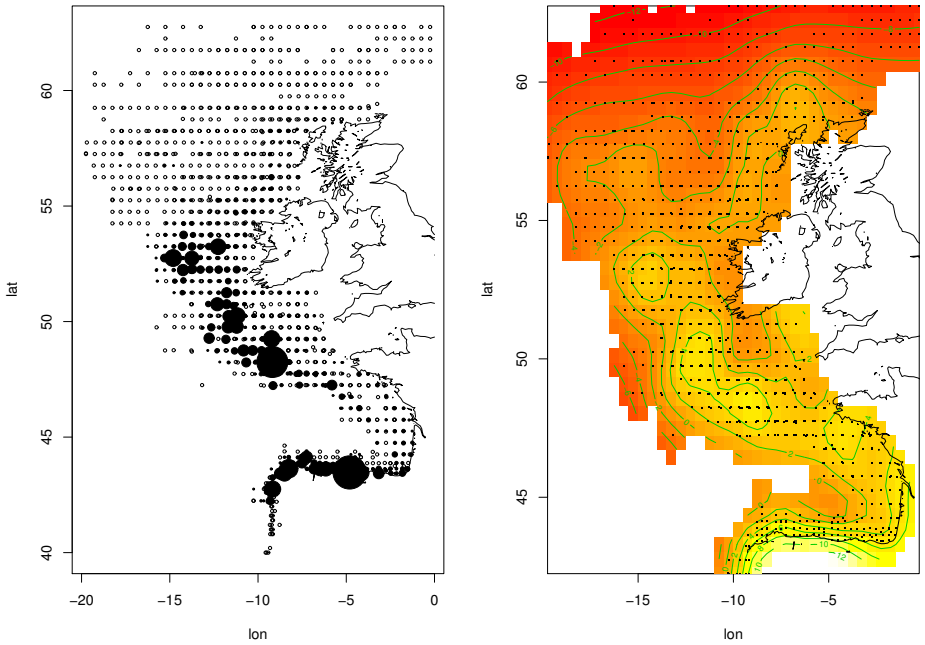

Figure 6: Left: 2010 Horse Mackerel egg survey data. Open circles are survey stations with no eggs, while solid symbols have area proportional to number of eggs sampled. Right: Fitted spatial effect from the best fit negative binomial based model.

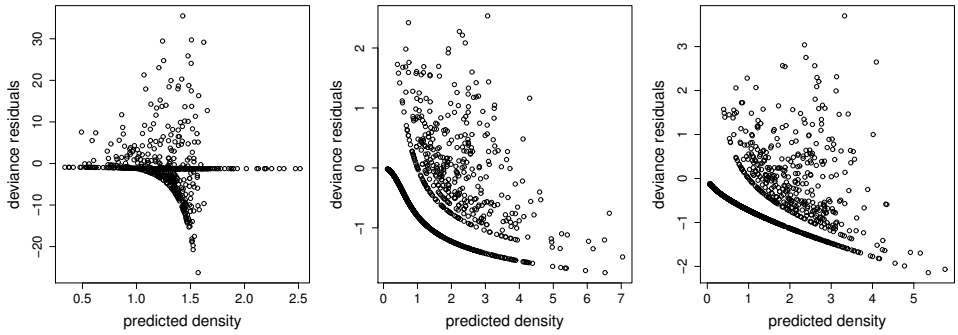

Figure 7: Residual plots for three Horse Mackerel egg models. Deviance residuals have been scaled by the scale parameter so that they should have unit variance for a good model. The fourth root of fitted values is used to best show the structure of the residuals. Left is for a zero inflated Poisson model: 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.

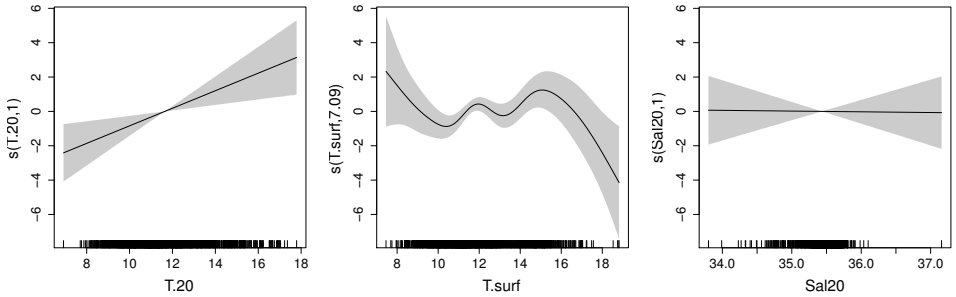

Figure 8: Horse Mackerel model univariate effect estimates.

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 $\boldsymbol{\theta}$.

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

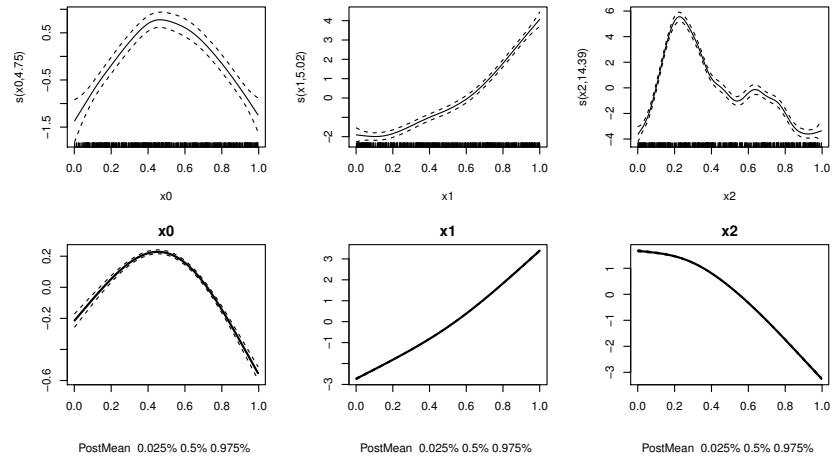

Figure 9: `mgcv` (top row) and INLA (lower row) fits (with 95% credible intervals) to the simple 3 term additive model simulated in Appendix L. Each row is supposed to be reconstructing the same true function, which in reality looks like the estimate in the upper row. On this occasion INLA encounters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