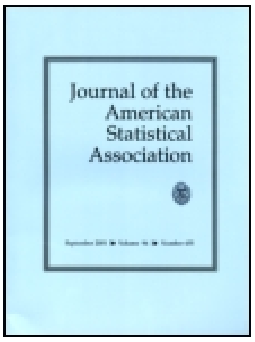
Wood, S. N., Pya, N., & Saefken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). Journal of the American Statistical Association, 111, 1548-1575. DOI: 10.1080/01621459.2016.1180986



Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Smoothing Parameter and Model Selection for General Smooth Models

Simon N. Wood, Natalya Pya & Benjamin Säfken

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

OPEN ACCESS

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
Published with license by Taylor and Francis.

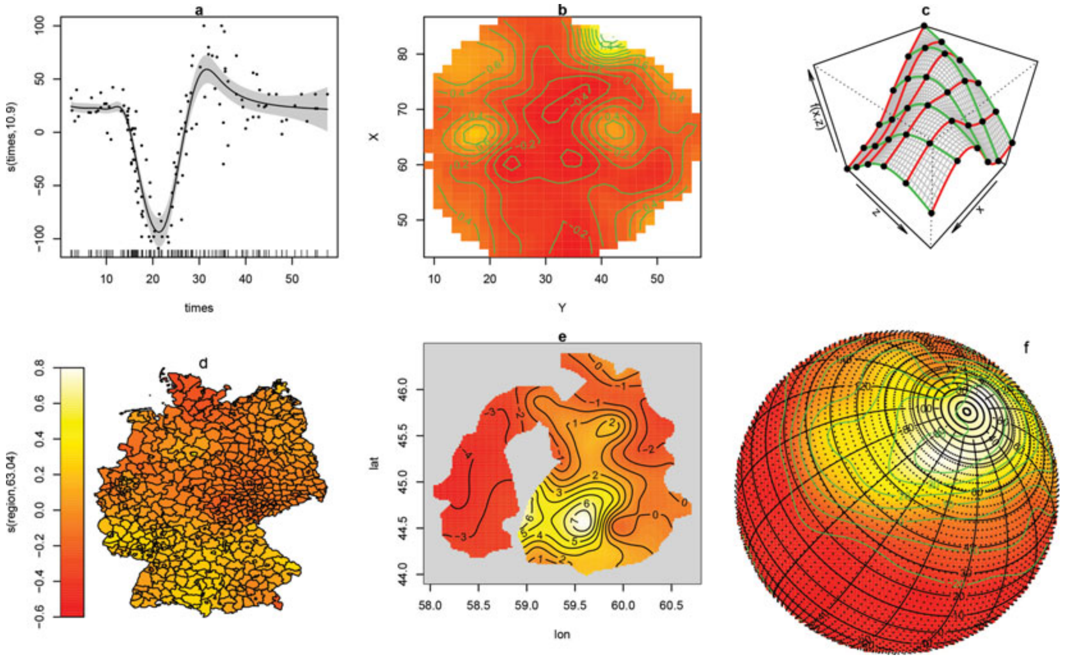


Figure 1. Examples of the rich variety of smooth model components that can be represented as reduced rank basis smoothers, with quadratic penalties and therefore can routinely be incorporated as components of a GAM. This article develops methods to allow their routine use in a much wider class of models. (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 $\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\boldsymbol{\mathcal{I}} + \mathbf{S}^\lambda)^{-1})$, where $\boldsymbol{\mathcal{I}}$ 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 $\text{diag}\{(\boldsymbol{\mathcal{I}} + \mathbf{S}^\lambda)^{-1}\boldsymbol{\mathcal{I}}\}$ 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

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

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 $a_{\boldsymbol{i}j} b_{\boldsymbol{i}k} c^{\boldsymbol{i}l} + d_{jkl} = 0$ 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 $a_{i\boldsymbol{j}} b_{\boldsymbol{j}}$ 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, λ_i, $\rightarrow \infty$ 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 $\hat{\boldsymbol{\beta}}$, $\mathcal{V}^i_\rho$ and $\mathcal{V}^{ij}_{\rho\rho}$ by the inner algorithm.
   (b) Drop any $\mathcal{V}^i_\rho$, $\mathcal{V}^{ij}_{\rho\rho}$ and $\mathcal{V}^{ji}_{\rho\rho}$ for which $\mathcal{V}^i_\rho \simeq \mathcal{V}^{ii}_{\rho\rho} \simeq 0$. Let $\mathbb{I}$ denote the indices of the retained terms.
   (c) Test for convergence, that is, all $\mathcal{V}^i_\rho \simeq 0$ and the Hessian (elements $-\mathcal{V}^{ji}_{\rho\rho}$) is positive semidefinite.
   (d) If necessary perturb the Hessian (elements $-\mathcal{V}^{ji}_{\rho\rho}$) 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 $-\mathcal{V}^{\rho\rho}_{i\boldsymbol{j}}\mathcal{V}^{\boldsymbol{j}}_\rho$, 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 $\mathcal{L}^{\hat\beta\hat\beta}_{k\,j}$ denotes the inverse of $\mathcal{L}^{k\,j}_{\hat\beta\hat\beta}$.

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 $\boldsymbol{\mathcal{H}}$ 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 $\mathrm{d}\hat\beta_i/\mathrm{d}\rho_k = \mathcal{L}^{\hat\beta\hat\beta}_{i\,\boldsymbol{j}}\lambda_k S^k_{\boldsymbol{jl}}\hat\beta_{\boldsymbol{l}}$ and hence $l^{i\,j\,l}_{\hat\beta\hat\beta\rho} = l^{i\,j\,\boldsymbol{k}}_{\hat\beta\hat\beta\hat\beta}\mathrm{d}\hat\beta_{\boldsymbol{k}}/\mathrm{d}\rho_l$ (Section 3.1.3).
6. Compute $\mathrm{d}^2\hat\beta_i/\mathrm{d}\rho_k\mathrm{d}\rho_l = \mathcal{L}^{\hat\beta\hat\beta}_{i\,\boldsymbol{j}}\{(-l^{\boldsymbol{j}\,\boldsymbol{p}l}_{\hat\beta\hat\beta\rho} + \lambda_l S^l_{\boldsymbol{jp}})\mathrm{d}\hat\beta_{\boldsymbol{p}}/\mathrm{d}\rho_k + \lambda_k S^k_{\boldsymbol{jp}}\mathrm{d}\hat\beta_{\boldsymbol{p}}/\mathrm{d}\rho_l\} + \delta^l_k\mathrm{d}\hat\beta_i/\mathrm{d}\rho_k$, (Section 3.1.3).
7. Compute $\mathcal{L}^{\hat\beta\hat\beta}_{\boldsymbol{k}\,\boldsymbol{j}}l^{\boldsymbol{j}\,\boldsymbol{k}\,p\upsilon}_{\hat\beta\hat\beta\rho\rho}$ (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 $\mathcal{L}^{\hat\beta\hat\beta}_{k\,j}$. 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 $l^{i\,j}_{\beta\beta}$, given $\boldsymbol{\beta}$, along with $l^{i\,j\,k}_{\hat\beta\hat\beta\rho}$ given $\mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\rho_k$, and $\mathcal{L}^{\hat\beta\hat\beta}_{\boldsymbol{k}\,\boldsymbol{j}}l^{\boldsymbol{j}\,\boldsymbol{k}\,p\upsilon}_{\hat\beta\hat\beta\rho\rho}$ 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 $l^{j\,k\,p\upsilon}_{\hat\beta\hat\beta\rho\rho}$ 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 $\mathbf{S}^j$,

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

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,

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

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 Frobenius 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 $\hat{\beta}$

This section provides details for inner Steps 2–4. Newton iteration for $\hat{\boldsymbol{\beta}}$ requires the gradient vector, $\boldsymbol{\mathcal{G}}$, with elements $\mathcal{L}^{i}_{\beta} = l^{i}_{\beta} - \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{\boldsymbol{ij}} \beta_{\boldsymbol{j}}$ 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 $\mathbf{R}^{\mathsf{T}}\mathbf{R} = \boldsymbol{\mathcal{H}}' + \epsilon\mathbf{I}$ is repeated with increasing ϵ, starting from zero, until positive definiteness is obtained. The Newton step is then computed as $\boldsymbol{\Delta} = \mathbf{D}\mathbf{R}^{-1}\mathbf{R}^{-\mathsf{T}}\mathbf{D}\boldsymbol{\mathcal{G}}$. 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 $\mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{H}/\|\mathbf{H}\|_F + \mathbf{S}/\|\mathbf{S}\|_F$. 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 $\mathcal{L}^{i}_{\hat{\beta}} = l^{i}_{\hat{\beta}} - \lambda_{\boldsymbol{k}} S^{\boldsymbol{k}}_{\boldsymbol{ij}} \hat{\beta}_{\boldsymbol{j}} = 0$ and differentiating with respect to $\rho_k = \log\lambda_k$ yields

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

$$\frac{\mathrm{d}\hat{\beta}_i}{\mathrm{d}\rho_k} = \mathcal{L}^{\hat{\beta}\hat{\beta}}_{i\;\boldsymbol{j}}\lambda_k S^{k}_{\boldsymbol{jl}}\hat{\beta}_{\boldsymbol{l}},$$

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 $-l^{i\,j\,k\,l}_{\hat{\beta}\hat{\beta}\rho\rho} + \delta^{l}_{\boldsymbol{k}}\lambda_{\boldsymbol{k}}S^{\boldsymbol{k}}_{ij}$, and

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

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

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

and

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

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

$$\frac{\partial \log|\boldsymbol{\mathcal{H}}|}{\partial \rho_k} = \mathrm{tr}\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_k}\right) \text{ and}$$

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

The final term above is expensive if computed naively by explicitly computing each term $\partial^2\boldsymbol{\mathcal{H}}/\partial\rho_k\partial\rho_l$, but this is unnecessary and the computation of $\mathrm{tr}\left(\boldsymbol{\mathcal{H}}^{-1}\partial^2\boldsymbol{\mathcal{H}}/\partial\rho_k\partial\rho_l\right)$ 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 $l^{j}_{\beta^l} = l^{\boldsymbol{i}}_{\eta^l} X^l_{\boldsymbol{i}j}$ and $l^{j}_{\beta^l}{}^{k}_{\beta^m} = l^{\boldsymbol{i}}_{\eta^l}{}^{\boldsymbol{i}}_{\eta^m} X^l_{\boldsymbol{i}j} X^m_{\boldsymbol{ik}}$, which are also sufficient for first-order implicit differentiation.

LAML optimization also requires

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

$$= l^{\boldsymbol{i}}_{\hat{\eta}^l}{}^{\boldsymbol{i}}_{\hat{\eta}^m}{}^{\boldsymbol{i}}_{\hat{\eta}^{\boldsymbol{q}}} X^l_{\boldsymbol{i}j} X^m_{\boldsymbol{ik}}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p}.$$

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

$$l^{j}_{\hat{\beta}^l}{}^{k}_{\hat{\beta}^m}{}^{p}_{\rho}{}^{\upsilon}_{\rho} = l^{j}_{\hat{\beta}^l}{}^{k}_{\hat{\beta}^m}{}^{\boldsymbol{r}}_{\hat{\beta}^{\boldsymbol{q}}}{}^{\boldsymbol{t}}_{\hat{\beta}^{\boldsymbol{s}}}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p}\frac{\mathrm{d}\hat{\beta}^{\boldsymbol{s}}_{\boldsymbol{t}}}{\mathrm{d}\rho_\upsilon} + l^{j}_{\hat{\beta}^l}{}^{k}_{\hat{\beta}^m}{}^{\boldsymbol{r}}_{\hat{\beta}^{\boldsymbol{q}}}\frac{\mathrm{d}^2\hat{\beta}^{\boldsymbol{q}}_{\boldsymbol{r}}}{\mathrm{d}\rho_p\mathrm{d}\rho_\upsilon}$$

$$= l^{\boldsymbol{i}}_{\hat{\eta}^l}{}^{\boldsymbol{i}}_{\hat{\eta}^m}{}^{\boldsymbol{i}}_{\hat{\eta}^{\boldsymbol{q}}}{}^{\boldsymbol{i}}_{\hat{\eta}^{\boldsymbol{s}}} X^l_{\boldsymbol{i}j} X^m_{\boldsymbol{ik}}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p}\frac{\mathrm{d}\hat{\eta}^{\boldsymbol{s}}_{\boldsymbol{i}}}{\mathrm{d}\rho_\upsilon} + l^{\boldsymbol{i}}_{\hat{\eta}^l}{}^{\boldsymbol{i}}_{\hat{\eta}^m}{}^{\boldsymbol{i}}_{\hat{\eta}^{\boldsymbol{q}}} X^l_{\boldsymbol{i}j} X^m_{\boldsymbol{ik}}\frac{\mathrm{d}^2\hat{\eta}^{\boldsymbol{q}}_{\boldsymbol{i}}}{\mathrm{d}\rho_p\mathrm{d}\rho_\upsilon}.$$

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, $l^{i}_{\hat{\mu}^l}{}^{i}_{\hat{\mu}^m}{}^{i}_{\hat{\mu}^q}{}^{i}_{\hat{\mu}^s}$ rather than $l^{i}_{\hat{\eta}^l}{}^{i}_{\hat{\eta}^m}{}^{i}_{\hat{\eta}^q}{}^{i}_{\hat{\eta}^s}$, 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 $l^{i}_{\hat{\mu}^j}{}^{i}_{\hat{\mu}^j}{}^{i}_{\hat{\mu}^k}{}^{i}_{\hat{\mu}^k}$ as at all awkward, but we have

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

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

The general method requires $\mathcal{L}^{\hat{\beta}}_{\boldsymbol{k}}{}^{\hat{\beta}}_{\boldsymbol{j}} l^{\boldsymbol{j}}_{\hat{\beta}}{}^{\boldsymbol{k}}_{\hat{\beta}}{}^{p}_{\rho}{}^{\upsilon}_{\rho}$ to be computed, which would have $O\{M(M+1)n\mathcal{P}^2/2\}$ cost if the terms $l^{\boldsymbol{j}}_{\hat{\beta}}{}^{\boldsymbol{k}}_{\hat{\beta}}{}^{p}_{\rho}{}^{\upsilon}_{\rho}$ 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 $B_{ij} = \mathcal{L}^{\hat{\beta}}_{i}{}^{\hat{\beta}}_{j}$. Defining

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

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

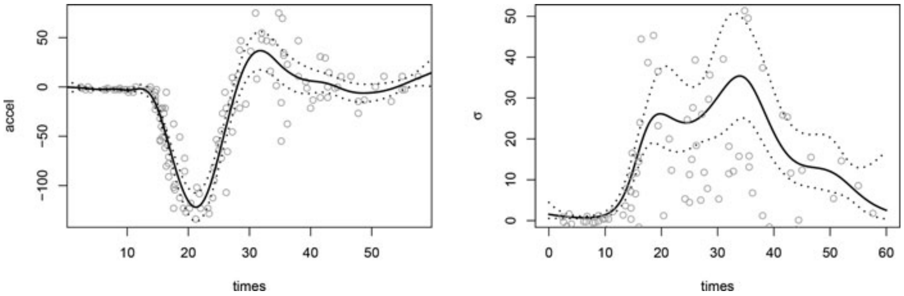


Figure 2. A smooth Gaussian location scale model fit to the motorcycle data from Silverman (1985), using the methods developed in Section 3.2. The left plot shows the raw data as open circles and an adaptive p-spline smoother for the mean overlaid. The right plot shows the simultaneous estimate of the 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 $\mathbf{B}\left(\begin{smallmatrix}\mathbf{X}^1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^2\end{smallmatrix}\right)^{\mathsf{T}}$ (which need only have $O(n\mathcal{P}^2)$ cost), each trace computation has $O(Mn\mathcal{P})$ cost (since $\mathrm{tr}(\mathbf{C}^{\mathsf{T}}\mathbf{D}) = D_{\boldsymbol{ij}}C_{\boldsymbol{ij}}$).

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 $\boldsymbol{\theta}, \phi$ 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^{i}_{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)

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

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 $\boldsymbol{\beta}|\mathbf{y} \overset{d}{=} \hat{\boldsymbol{\beta}}_{\rho^*} + \mathbf{R}^{\mathsf{T}}_{\rho^*}\boldsymbol{z}$ 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

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

where r is a lower order remainder term and $\mathbf{J} = \mathrm{d}\hat{\boldsymbol{\beta}}/\mathrm{d}\boldsymbol{\rho}|_{\hat{\boldsymbol{\rho}}}$. 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

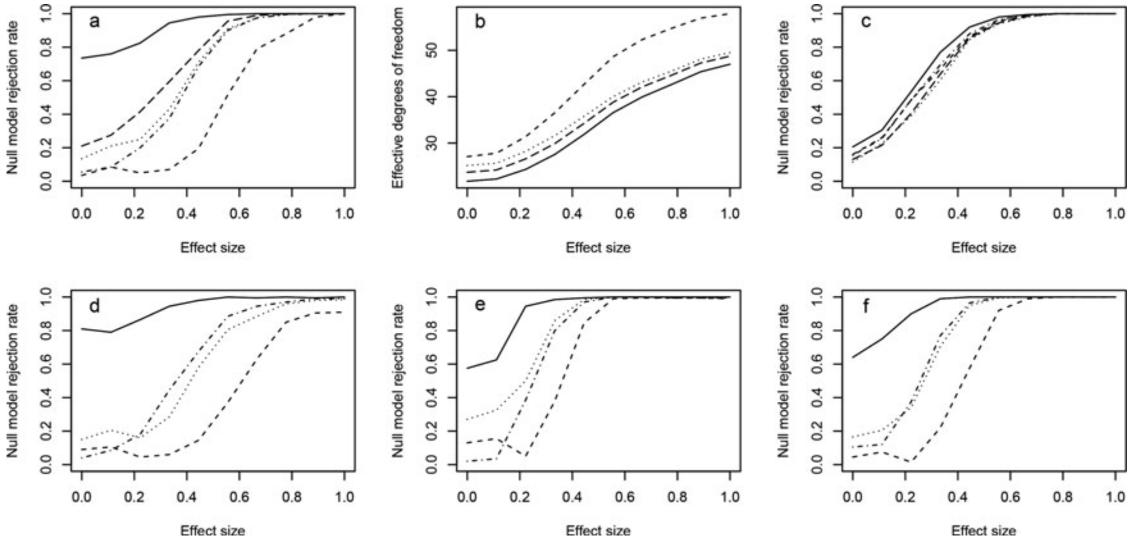


Figure 3. Simulation based illustration of the problems with previous AIC type model selection criteria and the relatively good performance of the Section 5 version. In all panels: (i) the solid curves are for conventional conditional AIC, (ii) the dotted curves are for the Section 5 version, (iii) the 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.

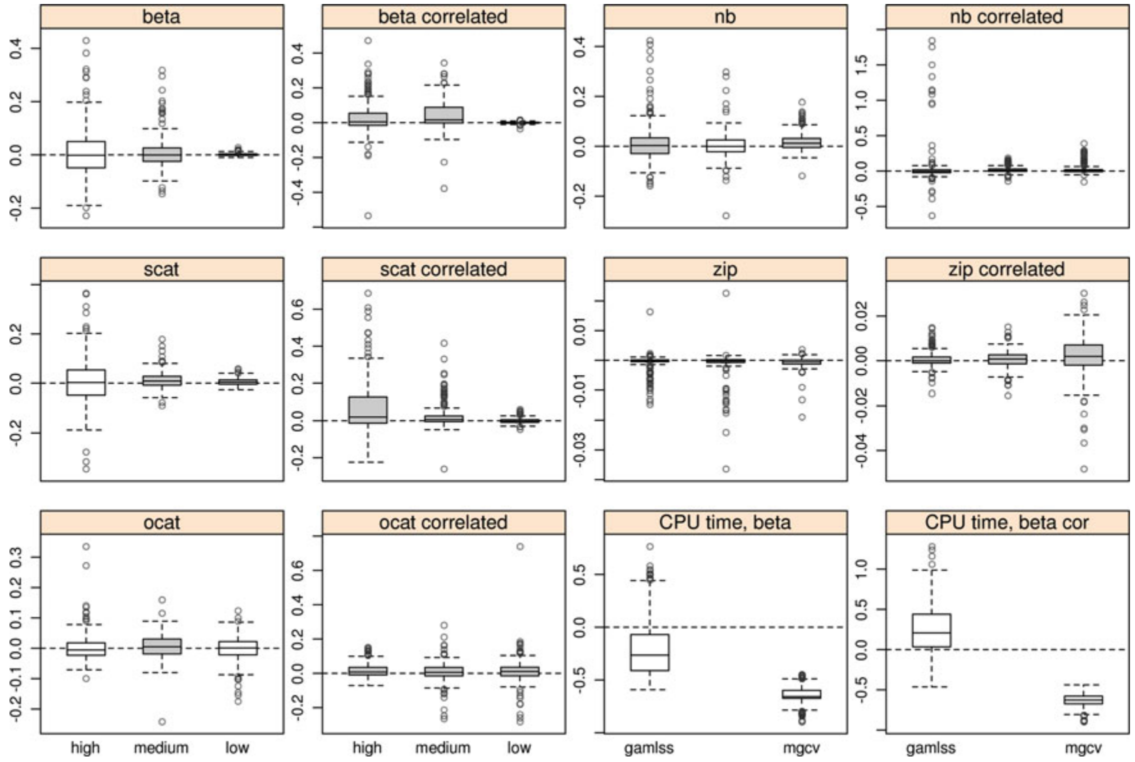


Figure 4. Results of simulation comparison with `gamlss` (beta, nb, scat, zip) and `BayesX` (ocat) packages for one dimensional P-spline models. The two plots at lower right show comparisons of $\log_{10}$ computing times for the case with the smallest time advantage for the new method — Beta regression. 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

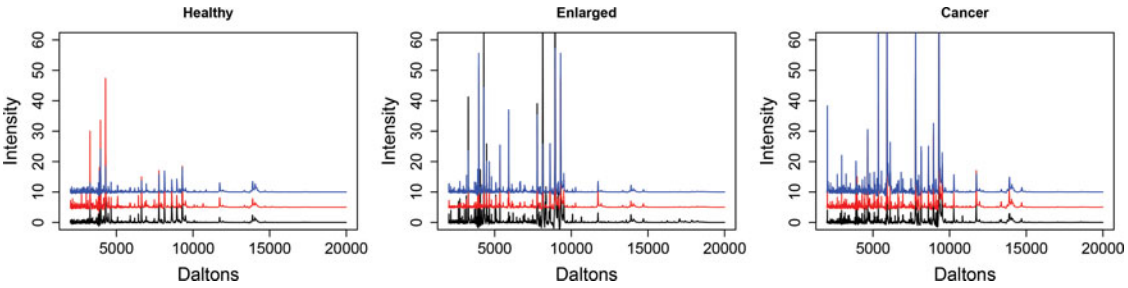


Figure 5. Three representative protein mass spectra (centered and normalized) from serum taken from patients with apparently healthy prostate, enlarged prostate, and prostate cancer. It would be useful to be able to predict disease status from the spectra. The red and blue spectra have been shifted upward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

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

where $f(D)$ is an unknown smooth function of mass D (in Daltons) and $\nu_i(D)$ 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

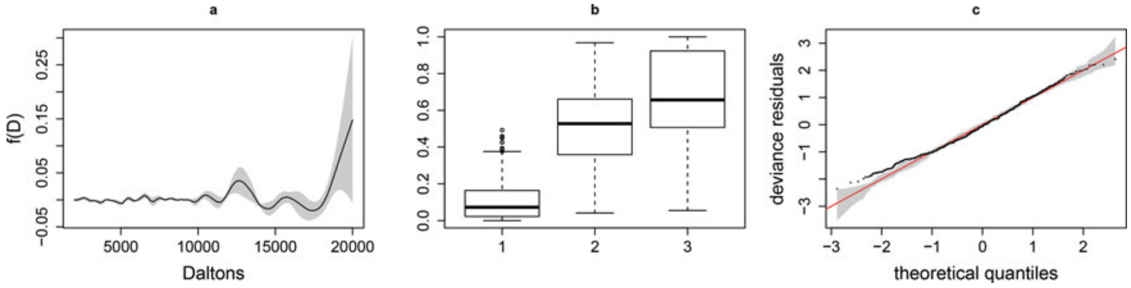


Figure 6. Results from the ordered categorical prostate model fit. (a) The estimated coefficient function $f(D)$ with 95% confidence interval. (b) Boxplots of the model probability of cancer, for the 3 observed states (1, healthy, 2, enlarged and 3, cancer). (c) QQ-plot of ordered deviance residuals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

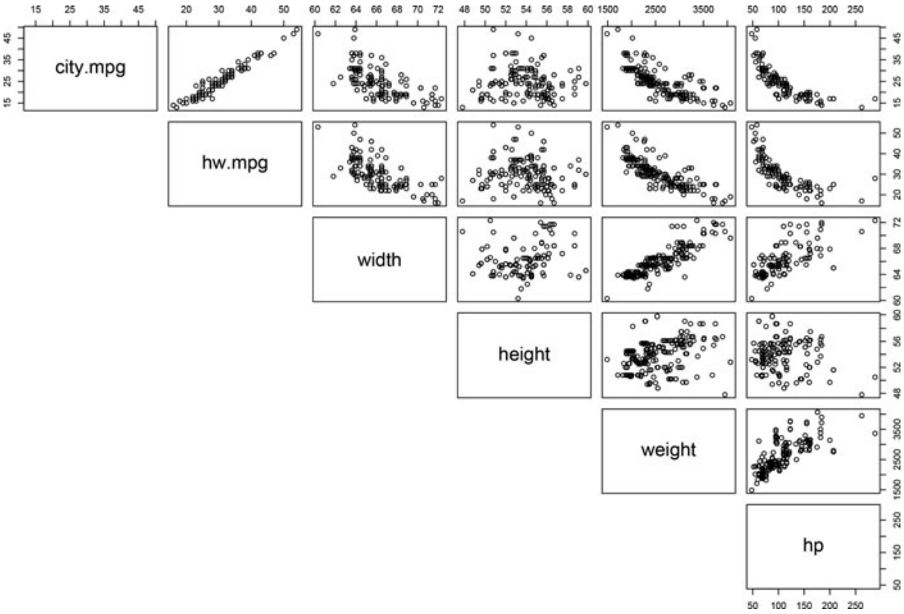


Figure 7. Part of a dataset from the USA on fuel efficiency of cars.

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

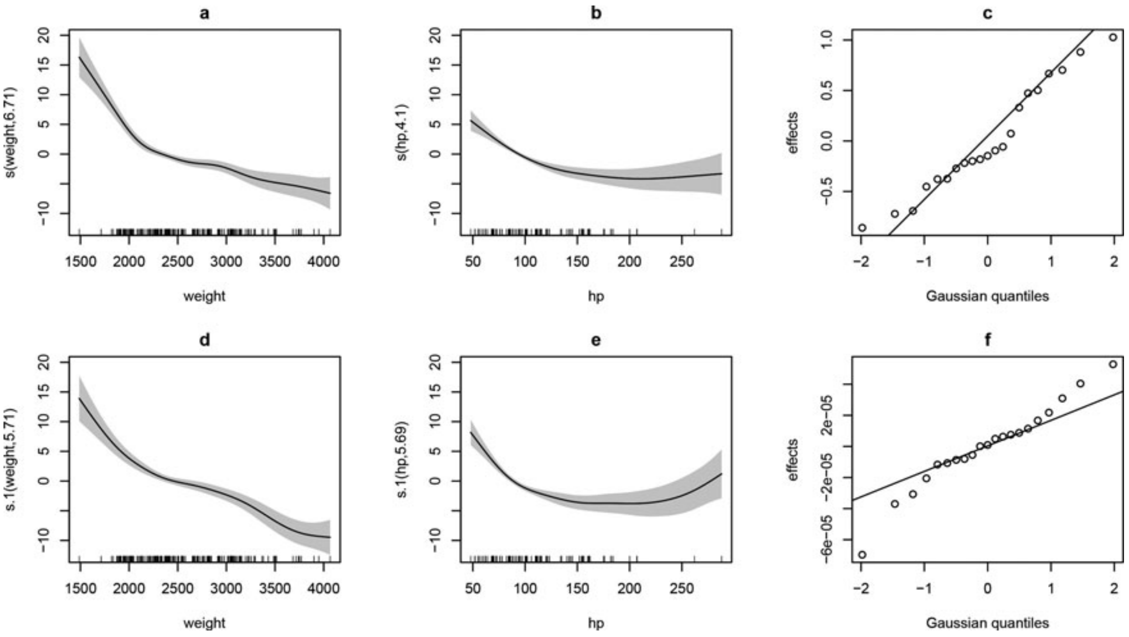


Figure 8. Fitted smooth and random effects for final car fuel efficiency model. Panels (a)–(c) relate to the city fuel consumption, while (d)–(f) are for the highway. (c) and (f) are normal QQ-plots of the predicted random effects for manufacturer, which in the case of highway MPG are effectively zero.

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 $\mathcal{D}^{\hat\beta\hat\beta}_{i\,j}$ denote elements of the inverse of the Hessian matrix $(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}^\lambda)$ with elements $\mathcal{D}^{i\ \,j}_{\hat\beta\hat\beta}$, 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

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

Differentiating once more yields

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

$$\left.+\mathcal{D}^{\boldsymbol{l}\ \boldsymbol{p}\ k}_{\hat\beta\hat\beta\,\theta}\frac{\mathrm{d}\hat\beta_{\boldsymbol{p}}}{\mathrm{d}\theta_j}+\mathcal{D}^{\boldsymbol{l}\ \ j\,k}_{\hat\beta\,\theta\,\theta}\right).$$

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

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

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

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

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

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,

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

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

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

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

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: $D^i_\beta=D^{\boldsymbol{k}}_\eta X_{\boldsymbol{k}i}$, $D^{i\ j}_{\beta\beta}=D^{\boldsymbol{kk}}_{\eta\eta}X_{\boldsymbol{k}i}X_{\boldsymbol{k}j}$, $D^{i\ j\,k}_{\beta\beta\beta}=D^{\boldsymbol{l}\,\boldsymbol{l}\,\boldsymbol{l}}_{\eta\eta\eta}X_{\boldsymbol{l}i}X_{\boldsymbol{l}j}X_{\boldsymbol{l}k}$. 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

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

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

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

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

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.