The definitive version is available at
onlinelibrary.wiley.com

University of Bath

Low rank scale invariant tensor product smooths for Generalized Additive Mixed Models

Simon N. Wood

Department of Statistics, University of Glasgow

Glasgow G12 8QQ U.K.

February 28, 2005

Abstract

A general method for constructing low rank tensor product smooths for use as components of GAMs or GAMMs is presented. A penalized regression approach is adopted in which tensor product smooths of several variables are constructed from smooths of each variable separately, these 'marginal' smooths being represented using a low rank basis with an associated quadratic wiggliness penalty. The smooths offer several advantages (i) they have one wiggliness penalty per covariate and are hence invariant to linear re-scaling of covariates, making them useful when there is no 'natural' way to scale covariates relative to each other; (ii) they have a useful tuneable range of smoothness, unlike single penalty tensor product smooths that are scale invariant; (iii) the relatively low rank of the smooths means that they are computationally efficient; (iv) the penalties on the smooths are easily interpretable in terms of function shape; (v) the smooths can be generated completely automatically from any marginal smoothing bases and associated quadratic penalties, giving the

modeler considerable flexibility to choose the basis penalty combination most appropriate to each modelling task; (vi) the smooths can easily be written as components of a standard linear or generalized linear mixed model, allowing them to be used as components of the the rich family of such models implemented in standard software, and to take advantage of the efficient and stable computational methods that have been developed for such models. Advantages (i) - (iii) are shared by methodology recently developed for the Smoothing Spline ANOVA approach to smooth modelling, but a small simulation study suggests that the new methods compare favourably with SS-ANOVA in terms of computational efficiency and mean square error performance.

1   Introduction

An Additive Mixed Model (special case of a GAMM, Lin and Zhang, 1999; Fahrmeir and Lang, 2001) has a structure something like

$$y_i = \mathbf{X}_i\boldsymbol{\theta} + w_{1i}f_1(x_{1i}) + w_{2i}f_2(x_{2i}, x_{3i}) + \ldots + \mathbf{Z}_i\mathbf{b} + \epsilon_i \qquad (1)$$

where y_i is a univariate response; $\boldsymbol{\theta}$ is a vector of fixed parameters; $\mathbf{X}_i$ is a row of a fixed effects model matrix; the w_{ji}s are covariates, dummy variables or often simply 1 (they are used in 'variable coefficient models': Hastie and Tibshirani, 1993); the f_js are smooth functions of covariates x_k; $\mathbf{Z}_i$ is a row of a random effects model matrix; $\mathbf{b} \sim N(0, \boldsymbol{\psi})$ is a vector of random effects coefficients with unknown positive definite covariance matrix $\boldsymbol{\psi}$; $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ is a residual error vector, with i^{th} element ϵ_i and covariance matrix $\boldsymbol{\Lambda}$, which is usually assumed to have some simple pattern. *Generalized* additive mixed models replace the normal residuals assumption with an assumption that $y_i|\mathbf{b}$ has some exponential family distribution and $E(y_i|\mathbf{b})$ is some monotonic function of the right-hand side of (1), excluding the ϵ_i term. These models are closely related to the geoadditive models of Kammann and Wand (2003), and Ruppert,

2

Wand and Carroll (2003) discuss a number of examples of models of this type. Additive mixed models also bear some relation to the models for designed experiments discussed, for example, by Verbyla *et al.* (1999) and implemented, for example, by Ball (2003). Generalized additive models (GAMs, Hastie and Tibshirani, 1990; see also Wahba, 1990) are a special case of GAMMs, which have no $\mathbf{Z}_i\mathbf{b}$ term.

GAMMs have an advantage over GAMs in that the more complex stochastic structure allows treatment of autocorrelation and repeated measures situations. The way in which smooths are actually incorporated into GAMMs varies. Lin and Zhang (1999) used cubic smoothing splines to represent the univariate smooths that they considered, while Wang (1998) represented a full smoothing spline ANOVA model (see e.g. Gu, 2002) as a normal linear mixed model. But other authors have tended to opt for the more computationally parsimonious penalized regression splines; either P-splines (Eilers and Marx, 1996) estimated using MCMC (Fahrmeir and Lang, 2001) or some variant on the thin plate spline basis or truncated power bases with estimation by REML (Kammann and Wand, 2003, Ruppert, Wand and Carroll, 2003).

Three approaches to representing smooths of more than one variable in GAMMs have been suggested. Either low rank approximations to thin plate splines have been employed (Kammann and Wand, 2003; Ruppert, Wand and Carroll, 2003) or tensor product P-splines have been suggested, with the single penalty given by the Kronecker product of the penalties associated with the marginal bases from which the smoothing basis is constructed (Fahrmeir and Lang, 2001). Finally, for smooths of 2 predictors in a fully Bayesian setting, and recognizing the undersmoothing that results from single Kronecker product penalties, Lang and Brezler (2004) suggested employing tensor products of equally spaced B-spline basis functions in conjunctions with spatially symmetric priors on the B-spline coefficients based on neighbouring coefficients. Lang and Brezler also generalized this to allow the degree of smoothing to vary over space:

3

these smooths perform well but are not invariant to re-scaling of the covariates.

By contrast, in non-GAMM settings the full tensor product smoothing splines of Wahba (1990) and Gu (2002) have a number of smoothing penalties associated with each component of an 'ANOVA - decomposition' of a smooth function, in such a way that the resulting smooths are invariant to covariate rescaling, but this approach is not easy to integrate with conventional approaches to (generalized) linear mixed modelling. Eilers and Marx (2002) have also used tensor products of B-splines to represent two dimensional surfaces, with separate difference penalties applied to the coefficients of the B-splines along the two covariate axes. When it is not appropriate to assume isotropy of a smooth of several variables then the invariance of such tensor product smooths is an important property.

The aims of this paper are: (i) to produce a general framework for constructing 'scale invariant' tensor product smooths from low rank penalized regression smoothers consisting of any set of basis functions and associated quadratic 'wiggliness' penalties, and (ii) to produce a general method for incorporating such smooths into standard (generalized) linear mixed models. This should substantially increase the flexibility both of smooth models, by allowing direct access to existing mixed model components and methods and of mixed modelling, by giving a straight-forward method of modelling smooth interactions, without the need for a complete change in modelling framework.

2   Low rank tensor product smooths

This section develops general methods for producing penalized regression smoothers which are functions of several variables. The smooths have the following properties: (i) they are invariant to linear rescaling of their covariates (while having a useful adjustable smoothness range); (ii) they are of relatively low rank, so that their computation is efficient and feasible for any size

of data set; (iii) they can be readily incorporated into existing penalized regression and mixed modelling frameworks, so that they add modelling functionality to these frameworks, rather than requiring the development of separate theory for fitting and inference; (iv) they allow the modeller considerable flexibility to easily construct smoothers designed for particular applied modelling tasks; (v) the penalties are fairly easy to interpret.

The basic approach of this section is to start from smooths of single covariates, represented using any basis with associated quadratic penalty measuring 'wiggliness' of the smooth. From these 'marginal smooths' a 'tensor product' construction is used to build up smooths of several variables. The basis construction is straightforward (and not novel): the key innovation is a method for obtaining simple wiggliness penalties for the tensor product smooth, which are interpretable in terms of function shape, and are induced naturally and automatically from the penalties of the marginal smooths.

The methods developed here can be used to construct smooth functions of *any* number of covariates, but the simplest introduction is via the construction of a smooth function of 3 covariates, x, z and v, the generalization then being trivial. The process starts by assuming that we have available low rank bases for representing smooth functions f_x, f_z and f_v of each of the covariates. That is we can write:

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x), \quad f_z(z) = \sum_{j=1}^{J} \delta_j d_j(z) \quad \text{and} \quad f_v(v) = \sum_{k=1}^{K} \beta_k b_k(v),$$

where the α_i, δ_j and β_k are parameters and the $a_i(x)$, $d_j(z)$ and $b_k(v)$ are known basis functions.

Now consider how the smooth function of x, f_x, could be converted into a smooth function of x and z. What is required is for f_x to vary smoothly with z, and this can be achieved by allowing its parameters, α_i, to vary smoothly with z. Using the basis already available for

representing smooth functions of z we could write:

$$\alpha_i(z) = \sum_{j=1}^{J} \delta_{ij} d_j(z)$$

which immediately gives:

$$f_{xz}(x, z) = \sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} d_j(z) a_i(x).$$

Continuing in the same way, we could now create a smooth function of x, z and v by allowing f_{xz} to vary smoothly with v. Again, the obvious way to do this is to let the parameters of f_{xz} vary smoothly with v, and following the same reasoning as before we get:

$$f_{xzv}(x, z, v) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{ijk} b_k(v) d_j(z) a_i(x).$$

For any particular set of observations of x, z and v a model matrix, $\mathbf{X}$, can be produced which maps the parameters β_{ijk} (suitably arranged into a vector $\boldsymbol{\beta}$) to the evaluated function at these values. If $\mathbf{X}_x$, $\mathbf{X}_z$ and $\mathbf{X}_v$ are model matrices for the individual marginal smooths, and $\otimes$ is the usual Kronecker product, then it is easy to show that, given appropriate ordering of $\boldsymbol{\beta}$, the ith row of $\mathbf{X}$ is simply:

$$\mathbf{X}_i = \mathbf{X}_{xi} \otimes \mathbf{X}_{zi} \otimes \mathbf{X}_{vi}.$$

Clearly (i) this construction can be continued for as many covariates as are required; (ii) the result is independent of the order in which we treat the covariates and (iii) the covariates can themselves be vector covariates.

Having derived a 'tensor product' basis for representing smooth functions, it is also necessary to have some way of measuring function 'wiggliness', if the basis is to be useful for representing smooth functions in a penalized regression or mixed modelling context. Again it is possible to start from wiggliness measures associated with the marginal smooth functions, and again the three covariate case provides sufficient illustration. Suppose then, that each marginal

smooth has an associated functional that measures function wiggliness and can be expressed as a quadratic form in the marginal parameters. That is

$$J_x(f_x) = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S}_x \boldsymbol{\alpha}, \quad J_z(f_z) = \boldsymbol{\delta}^{\mathrm{T}} \mathbf{S}_z \boldsymbol{\delta}, \quad J_v(f_v) = \text{ß}^{\mathrm{T}} \mathbf{S}_v \text{ß}$$

The $\mathbf{S}_{\bullet}$ matrices contain known coefficients, and $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ and ß are vectors of coefficients of the marginal smooths. An example of a penalty functional is the cubic spline penalty, $J_x(f_x) = \int (\partial^2 f_x / \partial x^2)^2 dx$. Now let $f_{x|zv}(x)$ be $f_{xvz}(x, z, v)$ considered as a function of x only, with z and v held constant, and define $f_{z|xv}(z)$ and $f_{v|xz}(v)$ similarly. A natural way of measuring wiggliness of f_{xzv} is to use:

$$J(f_{xzv}) = \lambda_x \int_{z,v} J_x(f_{x|zv}) dz dv + \lambda_z \int_{x,v} J_z(f_{z|xv}) dx dv + \lambda_v \int_{x,z} J_v(f_{v|xz}) dx dz$$

where the $\lambda_{\bullet}$ are smoothing parameters controlling the tradeoff between wiggliness in different directions, and allowing the penalty to be invariant to the relative scaling of the covariates. As an example, if cubic spline penalties were used as the marginal penalties, then

$$J(f) = \int_{x,z,v} \lambda_x \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \lambda_z \left(\frac{\partial^2 f}{\partial z^2}\right)^2 + \lambda_v \left(\frac{\partial^2 f}{\partial v^2}\right)^2 dx dz dv.$$

Hence, if the marginal penalties are easily interpretable in terms of function shape, then so is the induced penalty. Numerical evaluation of the integrals in J is straightforward. As an example consider the penalty in the x direction. The function $f_{x|zv}(x)$ can be written as

$$f_{x|zv}(x) = \sum_{i=1}^{I} \alpha_i(z, v) a_i(x)$$

and it is always possible to find the matrix of coefficients $\mathbf{M}_{z,v}$ such that $\boldsymbol{\alpha}(z, v) = \mathbf{M}_{zv} \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is the vector of β_{ijk} arranged in some appropriate order. Hence

$$J_x(f_{x|zv}) = \boldsymbol{\alpha}(z, v)^{\mathrm{T}} \mathbf{S}_x \boldsymbol{\alpha}(z, v) = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{M}_{zv}^{\mathrm{T}} \mathbf{S}_x \mathbf{M}_{zv} \boldsymbol{\beta}$$

and so

$$\int_{z,v} J_x(f_{x|zv})dzdv = \boldsymbol{\beta}^{\mathrm{T}} \int_{z,v} \mathbf{M}_{zv}^{\mathrm{T}} \mathbf{S}_x \mathbf{M}_{zv} dzdv \boldsymbol{\beta}.$$

The last integral can be performed numerically, and it is clear that the same approach can be applied to all components of the penalty. However, a simple reparameterization can be used to provide an approximation to the terms in the penalty, which performs well in practice and avoids the need for explicit numerical integration.

To see how the approach works, consider the marginal smooth f_x. Let $\{x_i^* : i = 1, \ldots, I\}$ be a set of values of x spread evenly through the range of the observed x values. In this case we can always re-parameterize f_x in terms of new parameters

$$\alpha_i' = f_x(x_i^*).$$

Clearly under this re-parameterization $\boldsymbol{\alpha}' = \mathbf{A}\boldsymbol{\alpha}$ where $A_{ij} = a_i(x_j^*)$. Hence the marginal model matrix becomes $\mathbf{X}_x' = \mathbf{X}_x \mathbf{A}^{-1}$ and the penalty coefficient matrix becomes $\mathbf{S}_x' = \mathbf{A}^{-\mathrm{T}} \mathbf{S}_x \mathbf{A}^{-1}$.

Now suppose that the same sort of re-parameterization is applied to the marginal smooths f_v and f_z. In this case we have that

$$\int_{z,v} J_x(f_{x|zv})dzdv \approx h \sum_{jk} J_x(f_{x|z_j^* v_k^*}),$$

where h is some constant of proportionality related to the spacing of the z_j^*'s and v_k^*'s. Similar expressions hold for the other integrals making up J. It is straightforward to show that the summation in the above approximation is:

$$J_x^*(f_{xzv}) = \boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{S}}_x \boldsymbol{\beta} \ \text{ where } \ \tilde{\mathbf{S}}_x = \mathbf{S}_x' \otimes \mathbf{I}_J \otimes \mathbf{I}_K$$

where $\mathbf{I}_J$ is the rank J identity matrix. Exactly similar definitions hold for the other components of the penalty so that

$$J_z^*(f_{xzv}) = \boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{S}}_z \boldsymbol{\beta} \ \text{ where } \ \tilde{\mathbf{S}}_z = \mathbf{I}_I \otimes \mathbf{S}_z' \otimes \mathbf{I}_K$$

8

and

$$J_v^*(f_{xzv}) = \boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{S}}_v \boldsymbol{\beta} \text{ where } \tilde{\mathbf{S}}_v = \mathbf{I}_I. \otimes \mathbf{I}_J \otimes \mathbf{S}_v'$$

Hence

$$J(f_{xzv}) \approx J^*(f_{xzv}) = \lambda_x J_x^*(f_{xzv}) + \lambda_z J_z^*(f_{xzv}) + \lambda_v J_v^*(f_{xzv}).$$

where any constants, h, have been absorbed into the λ_j. Again, this penalty construction clearly generalizes to any number of covariates.

Given its model matrix and penalties, the coefficients and smoothing parameters of a tensor product smooth can be estimated using the multiply penalized regression methods given in Wood (2004), or incorporated into a generalized additive model and estimated in the same way. Alternatively the smooth can be re-parameterized for representation as a component of a mixed model, as section 3 shows. Whatever the estimation method, the smooths are invariant to rescaling of the covariates, provided only that the marginal smooths are similarly invariant (which is always the case in practice).

2.1   Nesting and ANOVA decompositions

Consider a set of marginal smoothing bases used to construct interaction smooths of various orders. The tensor product construction immediately implies that any interaction of some set of covariates, is nested within all higher order interactions including that set of covariates (subject only to the mild technical restriction that all marginal bases include the constant function in their span). So, for example, the model

$$y_i = f_x(x_i) + f_z(z_i) + \epsilon_i$$

is strictly nested within the model

$$y_i = f_{zx}(x_i, z_i) + \epsilon_i$$

9

provided that the bases used for f_x and f_z are the marginal bases used for f_{xz}. This immediately opens up the possibility of working with 'ANOVA decomposition' type models:

$$y_i = f_x(x_i) + f_z(z_i) + f_{zx}(x_i, z_i) + \epsilon_i,$$

$$y_i = f_x(x_i) + f_z(z_i) + f_v(v_i) + f_{xz}(x_i, z_i) + f_{xv}(x_i, v_i) + f_{zv}(z_i, v_i) + f_{xzv}(x_i, z_i, v_i) + \epsilon_i,$$

etc. since all that is required to do this is to impose appropriate identifiability constraints on the model terms. When using such decompositions, each component smooth will have associated penalties and smoothing parameters (e.g. the three covariate decomposition given above would have 12 smoothing parameters).

This notion of an ANOVA decomposition of functions has been pioneered in the smoothing spline literature, and is the subject of a monograph by Gu (2002). Fuller comparisons with existing SS-ANOVA are provided in section 5.2.

3 Smooths as mixed model components

The aim of this section is to show how single penalty smooths represented by any basis and quadratic penalty and the tensor product smooths of section 2 can be included as components of a linear mixed model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \text{ where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda}) \tag{2}$$

where $\mathbf{y}$ is a response vector, $\mathbf{X}$ and $\mathbf{Z}$ are model matrices, $\boldsymbol{\beta}$ and $\mathbf{b}$ are fixed parameters and random effects, respectively, and $\boldsymbol{\epsilon}$ is a residual error vector. The covariance matrix $\boldsymbol{\psi}$ is usually parameterized in terms of some unknown parameters, while $\boldsymbol{\Lambda}$ is usually either diagonal, or has some other simple form induced by a simple correlation model. The idea is to add to the already extensive range of models readily available in this framework by providing straightforward means to include a wide range of penalized smooths and tensor product smooths in such models.

First consider a smooth with a single smoothing parameter. For example,

$$f(\mathbf{x}) = \sum_{j=1}^{J} b_j(\mathbf{x})\beta_j$$

with associated wiggliness measure $J(f) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}$, where $\mathbf{S}$ is a positive semi-definite matrix of coefficients (only semi-definite because most penalties treat some space of functions as having zero wiggliness). Given $(y_i, \mathbf{x}_i)$ data, it is straightforward to produce a model matrix $\mathbf{X}^f$, where $X_{ij}^f = b_j(\mathbf{x}_i)$, so that $\mathbf{X}^f\boldsymbol{\beta}$ is a vector of $f(\mathbf{x}_i)$ values.

The mixed model approach to estimating f starts from the premise that, by stating that f is smooth, we really believe that it is more probable that f is smooth than that f is wiggly. This can be formalized by specifying a prior for the wiggliness of the model which is $\propto \exp(-\lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}/2)$, say. Such a prior implies an improper Gaussian prior for $\boldsymbol{\beta}$ itself (Silverman, 1985). This improper distribution for $\boldsymbol{\beta}$ does not fit easily into standard linear mixed modelling approaches (e.g. Pinheiro and Bates, 2000). Some re-parameterization is therefore needed, so that the new parameters divide into a set with a proper distribution, to be treated as random effects, and a set (of size M) with an improper uniform distribution which can be treated as fixed effects. To achieve this, consider the eigen-decomposition, $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ is an orthogonal matrix, the columns of which are the eigenvectors of $\mathbf{S}$, and $\mathbf{D}$ is a diagonal matrix with the corresponding eigenvalues arranged in descending order on the leading diagonal. Let $\mathbf{D}_+$ denote the smallest sub-matrix of $\mathbf{D}$ containing all the strictly positive eigenvalues. Now re-parameterize, so that the new coefficient vector can be written $(\mathbf{b}_R^{\mathrm{T}}, \boldsymbol{\beta}_F^{\mathrm{T}})^{\mathrm{T}} \equiv \mathbf{U}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\beta}_F$ is of dimension M. It is clear that $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta} = \mathbf{b}_R^{\mathrm{T}}\mathbf{D}_+\mathbf{b}_R$ and that the coefficients $\boldsymbol{\beta}_F$ are unpenalized. Partitioning the eigenvector matrix so that $\mathbf{U} \equiv [\mathbf{U}_R : \mathbf{U}_F]$, where $\mathbf{U}_F$ has M columns and defining $\mathbf{X}_F \equiv \mathbf{X}^f\mathbf{U}_F$ while $\mathbf{X}_R = \mathbf{X}^f\mathbf{U}_R$, the mixed model representation of the smooth in terms of a linear predictor and random effects distribution is now

$$\mathbf{X}_F\boldsymbol{\beta}_F + \mathbf{X}_R\mathbf{b}_R, \quad \mathbf{b}_R \sim N(\mathbf{0}, \mathbf{D}_+^{-1}/\lambda)$$

11

where λ and $\boldsymbol{\beta}_F$ are fixed parameters to be estimated. For convenient estimation with standard software a further re-parameterization is useful. Defining $\mathbf{b} = \sqrt{\mathbf{D}_+^{-1}}\mathbf{b}_R$ and $\mathbf{Z} = \mathbf{X}_R\sqrt{\mathbf{D}_+}$ then the mixed model representation of the term, evaluated at its covariate values is

$$\mathbf{X}_F\boldsymbol{\beta}_F + \mathbf{Z}\mathbf{b}, \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{I}/\lambda)$$

Including such a term in a mixed model of the form (2) is simply a matter of appending the columns of $\mathbf{X}_F$ to the fixed effect model matrix, appending the columns of $\mathbf{Z}$ to the random effects model matrix and specifying the given random effects covariance matrix. Obviously, the multiple smooth terms of an additive model are easily combined (although some simple identifiability constraints are then required).

When representing tensor product smooths, which have multiple smoothing parameters, the only change is that the positive semi-definite pseudoinverse of the covariance matrix for $\boldsymbol{\beta}$ is now of the form $\sum_{i=1}^{d} \lambda_i \tilde{\mathbf{S}}_i$, where $\tilde{\mathbf{S}}_i$ is defined in section 2. The degree of rank deficiency of this matrix, M_T, is readily shown to be given by the product of the dimensions of the null spaces of the marginal penalty matrices $\mathbf{S}_i$ (provided that $\lambda_i > 0 \ \forall \ i$). Again re-parameterization is needed, this time by forming,

$$\sum_{i=1}^{d} \tilde{\mathbf{S}}_i = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$$

where $\mathbf{U}$ is an orthogonal matrix of eigenvectors and $\mathbf{D}$ is a diagonal matrix of eigenvalues, with M_T zero elements at the end of the leading diagonal. Notice that there are no λ_i parameters in the sum that is decomposed: this is reasonable since the null space of the penalty does not depend on these parameters (however given finite precision arithmetic it might be necessary to scale the $\tilde{\mathbf{S}}_i$ matrices in some cases).

It is not now possible to achieve the sort of simple representation of a term that was obtained with a single penalty, so the re-parameterization is simpler. Partitioning the eigenvector matrix so that $\mathbf{U} \equiv [\mathbf{U}_R : \mathbf{U}_F]$ where $\mathbf{U}_F$ has M_T columns, it is necessary to define $\mathbf{X}_F \equiv \mathbf{X}^f\mathbf{U}_F$,

$\mathbf{Z} \equiv \mathbf{X}^f \mathbf{U}_R$ and $\boldsymbol{\mathcal{S}}_i = \mathbf{U}_R^T \tilde{\mathbf{S}}_i \mathbf{U}_R$. A mixed model representation of the tensor product term (i.e. the linear predictor and random effects distribution) is

$$\mathbf{X}_F \boldsymbol{\beta}_F + \mathbf{Z}\mathbf{b}, \quad \mathbf{b} \sim N\left(\mathbf{0}, \left(\sum \lambda_i \boldsymbol{\mathcal{S}}_i\right)^{-1}\right)$$

where the λ_i and $\boldsymbol{\beta}_F$ parameters have to be estimated. Clearly the covariance matrix structure is not completely standard, but neither is it difficult to implement. For example, a new `pdMat` class implementing this covariance matrix structure can readily be written for use with the `nlme` software of Pinheiro and Bates (2000). Given such a class, incorporation of one or more tensor product terms into a linear mixed model is straightforward.

4 Generalization and Confidence Intervals

The discussion so far has focused on additive mixed models, and general methods for setting these up in a manner allowing estimation using standard software such as the `nlme` library of Pinheiro and Bates (2000). Estimation in the generalized case can proceed in a completely straightforward manner using the approximate PQL methods of e.g. Breslow and Clayton (1993). Venables and Ripley (2002) provide a suitable function `glmmPQL` based on iterative calls to the mixed modelling function `lme` from the `nlme` library.

The remaining issue is the calculation of confidence intervals. In most applications of GAMMs these would be required primarily for the smooth terms and the fixed effects. If this is the case then, following Silverman (1985), a Bayesian posterior covariance matrix for the coefficients of these terms can be obtained. Conditioning on the parameter estimates for the random effects, it is first necessary to calculate the covariance matrix for the response data (or pseudodata in the PQL case) implied by the estimated random effects structure *excluding* the smooth terms; suppose this is $\mathbf{V}$. Then if $\boldsymbol{\theta}$ is the vector of all the fixed parameters plus the

13

coefficient of the smooths, $\mathbf{X}$ is the model matrix corresponding to these terms and $\mathbf{S}_i$ is the ith penalty matrix (padded with zeros if necessary so that $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_i\boldsymbol{\theta}$ is the correct penalty) then

$$\boldsymbol{\theta}|\mathbf{y} \sim N(\hat{\boldsymbol{\theta}}, (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X} + \sum \lambda_i\mathbf{S}_i)^{-1})$$

where $\hat{\boldsymbol{\theta}}$ is the vector of estimates or predictions of the elements of $\boldsymbol{\theta}$. This is essentially the approach taken in Lin and Zhang (1999), and allows the required intervals to be obtained. The only quantity not readily available from standard software is the estimate $\mathbf{V}$, but with some effort it is possible to extract it, at least from `lme` fits. As usual the degrees of freedom per element of $\boldsymbol{\theta}$ can be estimated from the leading diagonal of $(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X} + \sum \lambda_i\mathbf{S}_i)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$.

5   Comparison with existing techniques

There is a substantial practical difference between using the approach suggested in section 2 and a simpler tensor product approach employing a single penalty matrix such as $\mathbf{S}_\pi = \mathbf{S}_w \otimes \mathbf{S}_x \otimes \mathbf{S}_z$ (e.g. Fahrmeier and Lang, 2001). The problem with such single penalties is their degree of rank deficiency. For example, a smooth of three variables constructed from three cubic spline bases, each of rank five, would have 125 parameters and a penalty of rank 27. Hence the effective degrees of freedom of the term would have to lie between 98 and 125, rendering the penalization effectively useless. In contrast, using the same marginal bases and the approach advocated here, the degrees of freedom of the smooth would lie between 8 and 125: a much more useful range for practical work. Alternatively one could employ a higher rank single penalty in association with a tensor product basis, but in that case the resulting smooth is no longer invariant to linear re-scaling of the arguments of the smooth. All proposals for tensor product smoothing with a single penalty are unsatisfactory for this reason: it is possible to have a useful tuneable smoothness range or scale invariance, but not both, and for practical purposes

both are needed.

5.1   Tensor products of P-splines

Eilers and Marx (2003) proposed constructing smooths of two variables from tensor products of P-splines, with separate difference penalties applied to each covariate direction, and hence two smoothing parameters to estimate. This suggestion is a special case of the method of section 2, but without the re-parameterization (or integration) step in the construction of the penalty. A small simulation comparing the Eilers and Marx method and the section 2 method was performed. 400 data were simulated from f_1 from section 5.2 with additive i.i.d. gaussian noise ($\sigma = 1$). 200 replicate data sets were produced and from each reconstructions of f_1 were attempted using both alternative methods. Each employed P-splines with a cubic basis and second order penalties as marginal penalties. Smoothing parameters were estimated by GCV. Mean square error performance was on average 8.5% better using the section 2 smooths paper, and these outperformed the Eilers and Marx smooths in 83% of replicates. Similar modest improvements were obtained at other noise levels, and are presumably attributable to the care that has been taken to preserve the penalties' relationship to function shape when moving from the marginal smooths to the tensor product smooth in section 2.

5.2   Smoothing Spline ANOVA

Smoothing Spline ANOVA, for which Gu (2002) is an excellent guide, provides a self contained framework for modelling with smooth functions, including an approach to mixed modelling. The tensor product smooths used in this framework have the scale invariance properties of the smooths introduced in section 2, although a different construction is used to achieve this. Until recently these methods were rather computationally costly, but Kim and Gu (2004) have

15

gone a long way towards improving this situation using low rank approximations, so a small simulation study was undertaken to compare SS-ANOVA to the methods proposed in section 2, in a penalized regression context. Data were simulated from 4 models of the general form:

$$y_i = f_j(x_i, z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where the ϵ_i were independent, the covariates x and z were independent uniform random deviates on $(0, 1)$ and the functions f_j were:

$$
\begin{aligned}
f_1(x, z) &= 10\pi\sigma_x\sigma_z \left(1.2e^{-(x-.2)^2/\sigma_x^2-(z-.3)^2/\sigma_z^2} + .8e^{-(x-.7)^2/\sigma_x^2-(z-.8)^2/\sigma_z^2}\right), \\
f_2(x, z) &= 2\sin(\pi * x) + \exp(2z), \\
f_3(x, z) &= 1.9(1.45 + \exp(x)\sin(13(x-.6)^2))\exp(-z)\sin(7z)
\end{aligned}
$$

and $f_4 = (f_1 + f_2)/2$ ($\sigma_x = .3$ and $\sigma_z = .4$). 200 replicate simulations were made using each model for each combination of $\sigma = 2$ and $\sigma = 0.5$ with sample sizes 200 and 1000. For each replicate, four reconstructions of f_j were produced and their mean square prediction error at the covariate values was assessed. The four approaches were (i) fit an SS-ANOVA model, by Kim and Gu's method as implemented in routine `ssanova1` of Gu's R package `gss`, (ii) fit a single section 2 smooth, (iii) fit an ANOVA decomposition model based on section 2 smooths, as described in section 2.1, (iv) chose whichever of (ii) and (iii) has the lowest GCV score. All alternatives used GCV to estimate smoothing parameters, and models (ii) and (iii) were estimated by the method of Wood (2004) as implemented in R package `mgcv`. All models had 49 coefficients.

The results are shown in figure 4. Method (iv) gave better average mean square error performance than method (i) in all cases, and required between 1/4 (n=200) and 1/12 (n=1000) of the computation time. In fact method (ii) was only worse than method (i) in terms of mean square error for the strictly additive f_2, but required between 1/10 and 1/30 of the computation time.

16

A smaller three covariate simulation was also undertaken. 500 triples x_i, z_i, v_i were generated from independent uniform distributions on $(0, 1)$ and response data were generated from:

$$y_i = 1.5e^{-(x_i-.2)^2/5-(z_i-.5)^2/3-(v_i-.9)^2/4} \quad + \quad .5e^{-(x_i-.3)^2/4-(z_i-.7)^2/2-(v_i-.4)^2/6}$$

$$+ \quad e^{-(x_i-.1)^2/5-(z_i-.3)^2/5-(v_i-.7)^2/4} + \epsilon_i$$

where the ϵ_i are i.i.d. Normal random deviates with $\sigma = 0.02$. 100 replicates of this simulation were performed. Three types of model were fitted to each replicate: (a) a tensor product smooth of x_i, z_i and v_i; (b) an SS-ANOVA type model with all 7 terms represented using the tensor product smooths of section 2 and (iii) an SS-ANOVA model estimated using the method of Kim and Gu (2004). Kim and Gu recommend that, to avoid over-fitting, each model degree of freedom should be treated as if it were 1.4 degrees of freedom when calculating GCV scores, so each model was estimated both by conventional GCV and this modified version (for the 2 covariate simulations the modified GCV score provided no improvement, possibly because rather fewer smoothing parameters are estimated in that case). For this example all models had 64 coefficients plus smoothing parameters to be estimated.

The results are summarized in figure 5. In this case method (c) took 190 times as long as method (b) and 550 times as long as method (a). Method (b) had the best MSE performance followed by method (c). With so many smoothing parameters in the ANOVA type models (12 for (b) and 19 for (c)), the nature of the penalty is effectively being estimated, as well as the degree of smoothness, and it may be this degree of flexibility that allows better MSE performance.

Experimentation with a 4 covariate smooth produced similar results: the 65 smoothing parameter, full SS-ANOVA becomes even more costly in this case, and again, better MSE performance seems to be obtainable more cheaply using the smooths of section 2.

A theoretical objection to the section 2 methods is that the number of model coefficients

tends to increases as the power of the number of covariates of the smooth. This need not happen with existing SS-ANOVA techniques, although it is doubtful that good function reconstructions will be obtained in practice if the number of parameters does not depend on dimension in a somewhat similar manner. In any case, no computational advantage of the SS-ANOVA approach is apparent for the above examples, which have up to 4 way interactions.

To conclude: the methods presented in this paper provide a useful complementary approach to the smoothing spline ANOVA framework described in Gu (2002) and may offer the following advantages. (i) in a penalized regression context, the simulation study suggests that the new methods compare favourably with SS-ANOVA in both computational speed and MSE performance; (ii) because the bases of the smooths do not depend on the smoothing parameters, the new smooths are readily incorporated into existing methodological frameworks for linear and generalized linear mixed modelling, and indeed non-linear mixed effects modelling: this in turn gives ready access to a rich variety of modelling tools including all the inferential machinery of mixed models, and the reliable and extensive software libraries for mixed modelling; (iii) the new tensor product wiggliness penalties are relatively interpretable in terms of function shape, provided that the penalties associated with the marginal smooths are interpretable, whereas the penalties associated with full SS-ANOVA models require quite a high degree of sophistication to interpret; (iv) the interpretation of the distribution of the model coefficients treated as random effects is straightforward: a negative exponential distribution is assumed for function 'wiggliness'; (v) the new methods make it very easy to build smooths of several variables from smooths designed for particular modelling tasks: for example, for some practical applications smooths should be very 'flat' away from high densities of supporting data, and the new methods can easily accommodate such 'designer smooths'. Point (ii) is particularly significant from an applied perspective: the literature on mixed modelling is large and growing, and high quality software

is available to implement a rich variety of mixed models and associated inferential techniques. It is of therefor of some utility to be able to add good smoothing methods to this framework, rather than having to try and develop parallel mixed modelling methodologies specifically applicable to smooth modelling. On the other hand, the methods developed here do not offer the complete framework for smooth modelling that the SS-ANOVA approach provides.

6 Examples

This section illustrates the utility of the methods using two simulated examples and a short real example. Firstly, data were simulated from the model

$$y_i = f(x_i) + e_i, \quad i = 1, \ldots, 400$$

where $f(x) = x^{11}[10(1-x)]^6/5 + 10(10x)^3(1-x)^{10}$, $e_i = 0.6e_{i-1} + \epsilon_i$ for $i = 1 \ldots 400$, $e_0 = 0$, $\epsilon_i \sim N(0, 1.5^2)$ and the x_i were uniformally spaced on [0,1]. The function f was then treated as unknown and represented by a rank 20 P-spline basis (cubic B-splines, penalized by a 2nd order difference penalty: see Eilers and Marx, 1996), while the noise was modelled as an AR(1) process with unknown correlation parameter. Note that the P-spline basis does not have immediately identifiable fixed (i.e. unpenalized) and random (penalized) components, so the approach of section 3 is required. After representing the model as a linear mixed model it was estimated using REML (S routine `lme`, Pinheiro and Bates, 2000). For comparison, fits were also made assuming i.i.d. errors using REML and performing estimation by penalized likelihood using GCV for smoothness selection (`gam` from R package `mgcv`). Figure 1 shows typical results: the mixed model with AR(1) errors produces a reasonable reconstruction of the truth, with plausible 95% confidence bands, while the methods that neglect autocorrelation overfit, and produce overly narrow confidence bands.

The second example uses a 'repeated measures' additive mixed model with one tensor product term and Poisson errors. 400 Poisson data y_i were simulated from $y_i \sim \mathrm{Poi}(e^{\eta_i})$ where

$$\eta_i = f_1(x_i, z_i) + f_2(w_i) + b_j \text{ if observation } i \text{ is from group } j.$$

There were 10 b_j terms which were i.i.d. $N(0,1)$; each group contained 40 observations; the x_i and w_i were independent uniform random deviates on (0,1); the z_i were independent random deviates on (0,0.05); $f_1(x, z) = 2 \exp(-(x - 0.2)^2/\sigma_x^2 - (z - 0.015)^2/\sigma_z^2) + 1.3 \exp(-(x - 0.7)^2/\sigma_x^2 - (z - 0.04)^2/\sigma_z^2)$ where $\sigma_x = 0.3$ and $\sigma_z = 0.02$; $f_2(w) = \sin(2\pi w)$. The response data are plotted against the three covariates, and a spurious covariate v, in figure 2. Section 6.1 presents a typical study requiring such a model.

Three models were fitted to the data: all assumed Poisson errors and a log link; represented f_2 with a 10 knot cyclic penalized cubic regression spline and included a final nuisance term (f_3, not in the truth) represented by a 10 knot 'P-spline'. The first model was a GAMM including a random effect for group and representing f_1 with a tensor product of penalized cubic regression splines with 6 knots per direction (piecewise cubic Hermite polynomial bases). The second model was the same as the first except that f_1 was represented by a rank 36 isotropic smooth (a thin plate regression spline, Wood, 2003). The final model was as the first, but without the random effect and estimated by penalized likelihood maximization with smoothing parameters chosen by an unbiased risk estimator (see Wahba 1990), which is approximately AIC.

Results for a typical replicate simulation are shown in figure 3. Clearly an isotropic smooth is unsatisfactory here, while neglecting the correlation structure in the data leads to over-fitting. Hence, for this type of data, the work reported in this paper is a necessary addition to GAMM and GAM methods.

6.1 Mackerel example

Fish stock assessments are sometimes undertaken by surveying the eggs of a particular species in order to work out egg abundance, from which total mass of the spawning stock of fish can be inferred. One such survey was undertaken in 1992 off the west coasts of Britain, Eire and France targeting Mackerel eggs. Several fisheries research vessels sampled on an 'irregular grid' by hauling a fine meshed net vertically through the water column and counting the mackerel eggs found in the net (see left most panel figure 6). GAMs were used to model these data by Borchers et al. (1997). The best models in terms of explaining the egg abundances tend to depend almost exclusively on geographic predictors, such as longitude, latitude and distance from the 200m sea bed contour (a proxy for the distance from the edge of the continental shelf). These are fine for stock assessment, but less satisfactory in terms of biological interpretability, since they depend on quantities which the fish are unlikely to directly respond to.

Biologically, it would be interesting to try and base prediction entirely on variables that the fish might be able to sense, such as salinity, water temperature, sea bed depth and perhaps latitude (since day length varies with latitude over the survey area). For the purposes of this example, square root of observed egg density per square metre of sea surface, y, is used as the response, and this is modelled as having a normal distribution (modelling the counts directly and using a Poisson distribution is also possible, but in that case there is substantial overdispersion to be dealt with). The model used was then:

$$\sqrt{y_i} = f_1(\texttt{r.bd}_i, \texttt{lat}_i, \texttt{temp}_i) + f_2(\texttt{sal}_i) + b_j + \epsilon_i$$

assuming that observation i was obtained by boat j. The random effects b_j are assumed i.i.d. Normal, while the vector of residuals is $\epsilon \sim N(\mathbf{0}, \mathbf{\Lambda})$, $\mathbf{\Lambda}$ being given by the assumption that the residuals are correlated in a manner that decays exponentially with geographic distance between

observations nested within vessel (see Pinheiro and Bates, 2000). The vessel effect allows for differences in operating procedures etc. between the boats. The spatial correlation is to account for aggregation not explicable by the covariates. It is nested within vessel, since in practice different vessels tend to be separated in time when proximate in space. The smooth function f_1 was represented using a tensor product smooth, with marginal cubic regression spline bases of dimension 6: it is a function of the square root of sea bed depth, latitude and temperature at 20 metres depth. f_2 depends on salinity and was represented using a rank 10 thin plate regression spline. Salinity is unlikely to interact strongly with the other covariates.

The model was estimated by likelihood maximization (REML estimates are very similar). The salinity effect is estimated to be a straight line with slope very close to zero, and no sensible model selection criterion would leave it in the model, so it was dropped. The standard deviation of the vessel effect was estimated to be only 1% of the residual standard deviation, and the spatial auto-correlation was similarly close to zero, however these were not dropped, given their role as nuisance factors included purely to avoid being misled about the other effects.

Figure 7 shows some slices through the estimated f_1: note the apparent preference for relatively cool deep water, and the way that temperature preference does not seem to change greatly with latitude. Figure 6 also shows predicted square root of egg density and its standard deviation. Notice how the bulk of the distribution is off the shelf edge, and the survey area is failing to cover the whole distribution: in part this is because the fish were expected to be rather closer to the shelf edge (200 metre contour) than appears to actually be the case.

7 Conclusions

The main innovation reported in this paper is a general method for producing low rank, scale invariant tensor product smooths for inclusion into GAMMs and GAMs, which have a practi-

cally useful smoothness range when smoothness is to be estimated as part of model fitting, and which can be constructed from a wide variety of 'marginal' smooths. The importance of scale invariance is well illustrated in the first row of figure 2, where an example of the suggested smooths is compared with an isotropic smooth: if the covariates of a smooth are not on the same scale, then assuming isotropy can lead to very poor results, which the proposed method overcomes by using a separate penalty for each covariate direction. The ability to include these smooths directly into standard mixed models is also useful, both as an extension to what can be achieved with such models, and because of the ready access that it gives to mixed modelling theory for doing things like setting confidence intervals on smoothing parameters.

The methods of section 2 can be viewd as a generalization of the proposal of Eilers and Marx (2003) to any number of covariates, any marginal smoothing bases and quadratic penalties, and mixed model settings, while offering improvements in the performance of such smooths by virtue of the novel approach to penalty construction. As discussed in section 5.2, the methods complement existing SS-ANOVA methods by providing a means of representing tensor product smooths and ANOVA decompositions of functions which is easily extended (e.g. to variable coefficient models and non-linear mixed models) and easily incorporated into existing mixed modelling methodology, while allowing considerable freedom to design smooths for particular purposes. The simulation study suggests that when compared to existing SS-ANOVA models, the new methods perform favourably in terms of computation time and MSE performance.

The methods described here are implemented in package `mgcv` for R (R Core Development Team, 2003).

References

Ball, R.D. (2003) lmeSplines. *R News* **3**(3), 24-28

Borchers, D.L., S.T. Buckland, I.G. Priede and S. Ahmadi (1997) Improving the precision of the daily egg production method using generalized additive models. *Can. J. Fish. Aquat. Sci.* **54**:2727-2742.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9- 25.

Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* **11**:89-121

Eilers, P.H.C. and Marx, B.D. (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems* **66**:159-174

Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors *Applied Statistics* **50**, 201-220

Gu, C. (2002) *Smoothing spline ANOVA models* Springer-Verlag, New York

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized additive models.* Chapman & Hall, London

Hastie, T.J. and Tibshirani, R.J. (1993) *Varying coefficient models. J.R. Statist. Soc. B* **55**, 757-796.

Kim, Y.J. and Gu C. (2004) *Smoothing spline Gaussian regression: more scalable computation via efficient approximation J. Roy. Statist. Soc. B* **66**: 337-356

Kammann, E.E. and Wand M.P. (2003) Geoadditive models. *Applied Statistics* **52**:1-18

Lang, S. and Bresger, D. (2004) Bayesian P-splines.*J. Comp. Graph. Statist.* **13**:183-212

Lin X. and Zhang, D. (1999) Inference in generalized additive mixed models using smoothing splines. *J.R. Statist. Soc. B* **61**, 381-400.

Pinheiro, J.C. and Bates, D.M. (2000) *Mixed- Effects Models in S and S-PLUS* Springer-Verlag, New York.

R Development Core Team (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) *Semiparametric Regression.* Cambridge

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J.R. Statist. Soc. B* **47**,1-52.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics in S 4th ed.* , Springer-Verlag, New York.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines *Applied Statistics* **48**: 269-313

Wahba (1990) *Spline models for observational data. CBMS-NSF Reg. Conf. Ser. Appl. Math.*: **59**.

Wood, S.N. (2003) Thin plate regression splines. *J. R. Statist. Soc. B* **65**, 95-114.

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Ass.* **99**:637-686

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance *J.R. Statist. Soc. B* **60**, 159-174.
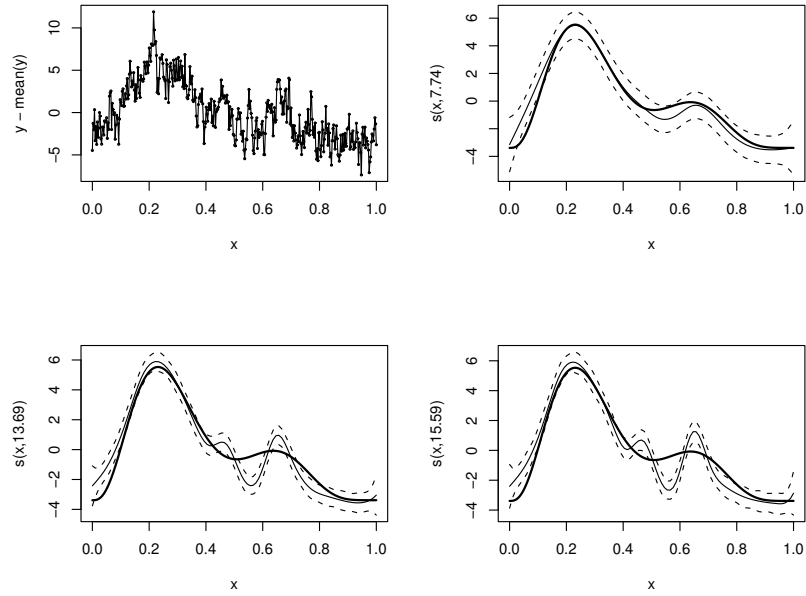
Figure 1: Reconstructing a smooth function sampled with auto-regressive error. The upper left plot shows the data. The upper right plot shows the reconstruction using a mixed model representation of a P-spline model for the smooth function with an AR(1) error model, estimated using REML; the bold line is the true function, the thin continuous line the reconstruction and the dashed lines are 95% confidence limits. The lower left panel is similar but assuming i.i.d. errors. The lower right panel is as the lower left, but estimated using penalized likelihood maximization with smoothness selected by GCV. In all panels the plots are centered to have zero mean over the covariate values. The figures in the y-axis labels give the estimated degrees of freedom for the smooths.
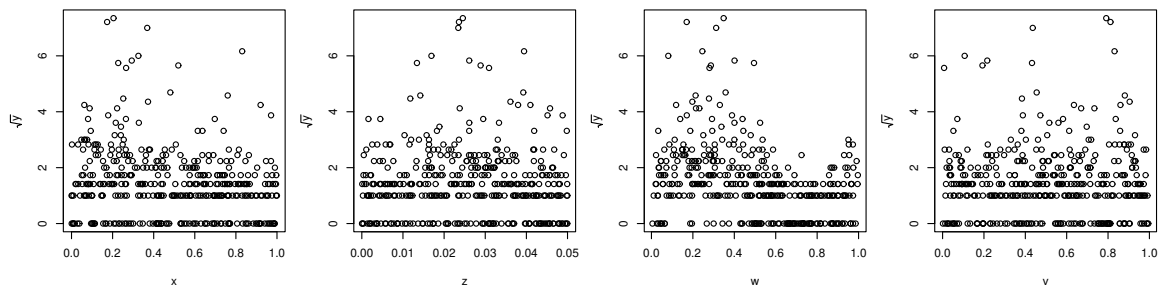
Figure 2: Scatter plots of the square root of the response data against each candidate covariate for the GAMM repeated measures example. Note how difficult it would be to judge what the appropriate scaling of x and z ought to be by straightforward inspection of the data.
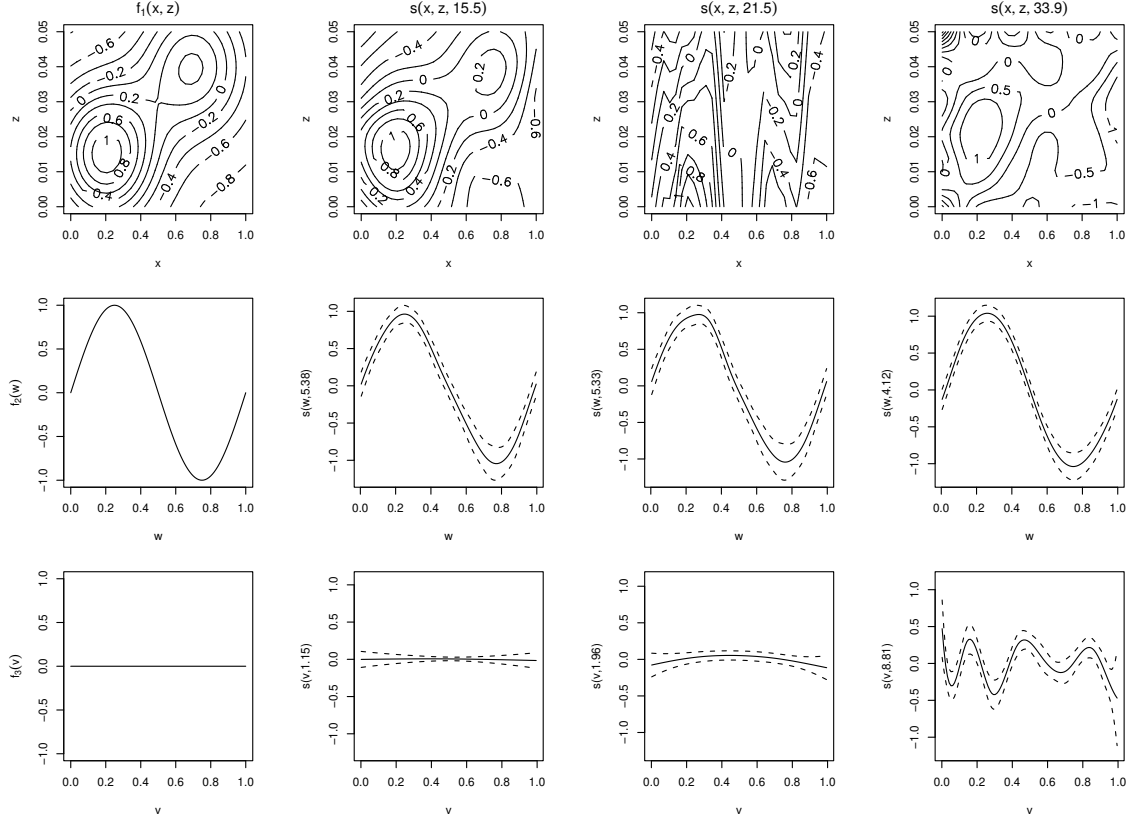
Figure 3: The true and estimated component functions of the simulated repeated measures GAMM example. The rows show, from top to bottom f_1, f_2 and the spurious function (f_3, say). The columns, from left to right, show: the true functions used in simulation; the component functions of a GAMM estimated by PQL with f_1 represented as a tensor product term and with a random effect for group; the same as the previous column, but with an isotropic smooth term for f_1; finally a GAM assuming i.i.d. errors, but with a tensor product smooth for f_1, estimated by penalized likelihood maximization, with smoothing parameters chosen by unbiased risk estimation (approximate AIC). In all cases the figure in the response axis label gives the effective degrees of freedom of the plotted smooth term estimate.
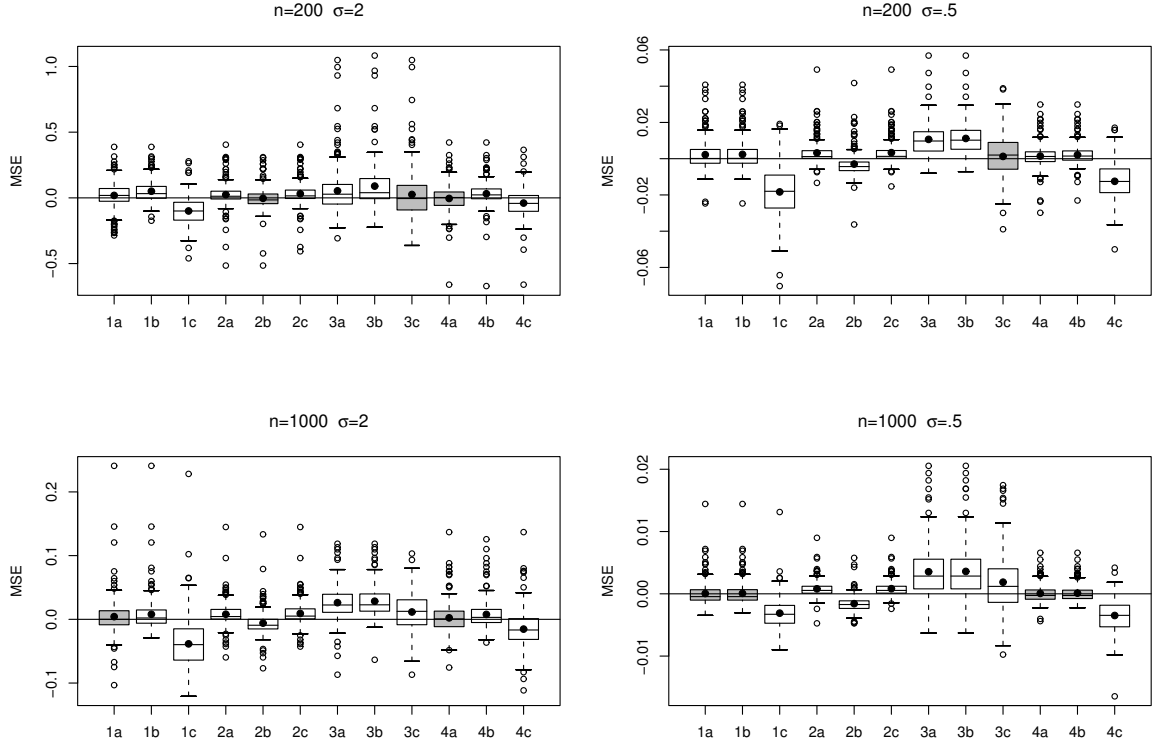
28

Figure 4: Comparisons of Mean Square Error (MSE) performance for the tensor product smooths of section 2 and the low rank SS-ANOVA method of Kim and Gu (2004) for smooth functions of two covariates. Each box plot shows the SS-ANOVA MSE minus the MSE of the alternative method based on the smooths in section 2, over 200 replicates. The number in the x axes labels refers to the index of the true function from which the data were simulated. The letters in the labels are as follows: (a) is where whichever of a single smooth and an SS-ANOVA style decomposition of a smooth had the lower GCV score was selected as the best model; (b) is for a single section 2 style smooth; (c) is for a smooth represented using an SS-ANOVA type decomposition, but using the section 2 smooths to represent this. The four panels relate to different combinations of sample size and noise level. All values above the zero line indicate that the section 2 based method had better MSE than the Kim and Gu SS-ANOVA method. Grey shaded boxplots indicate comparisons where the mean difference in MSE was not significantly different from zero at the 5% level. The type (a) smooths, based on the methods of section 2, consistently outperform the Kim and Gu (2004) SS-ANOVA method, while using less computer time to estimate.
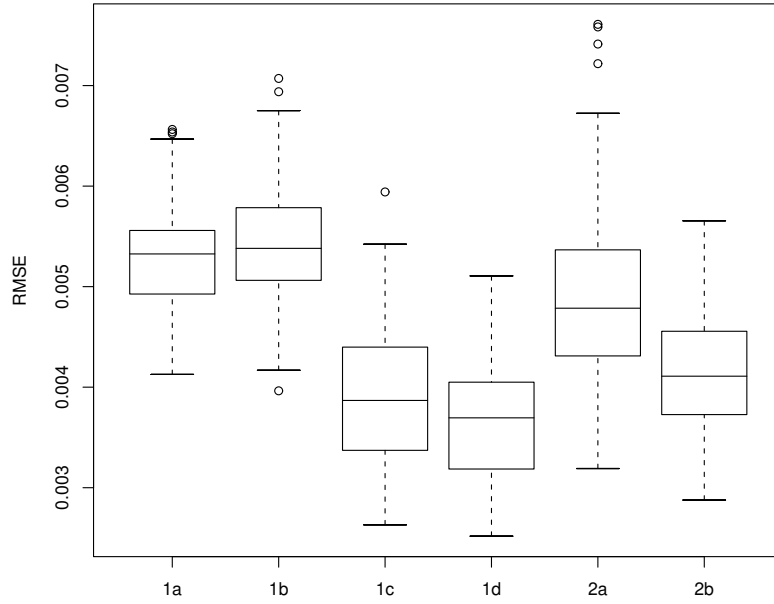
Figure 5: Comparisons of Root Mean Square Error (RMSE) performance for the tensor product smooths of section 2 and the low rank SS-ANOVA method of Kim and Gu (2004) for a smooth function of three covariates. Each box plot shows the RMSE for one alternative model. Smoothing parameters were estimated by GCV or vGCV in which each model degree of freedom counts as 1.4 degrees of freedom. 1a and 1b are for single section 2 smooths estimated by GCV and vGCV respectively; 1c and 1d are for SS-ANOVA style decompositions represented using section 2 smooths and estimated by GCV and vGCV respectively; 2a and 2b are for Kim and Gu SS-ANOVA models estimated by GCV anf vGCV respectively. 1c consistently gave lower GCV scores than 1a, while 1d consistently gave lower vGCV scores than 1b.
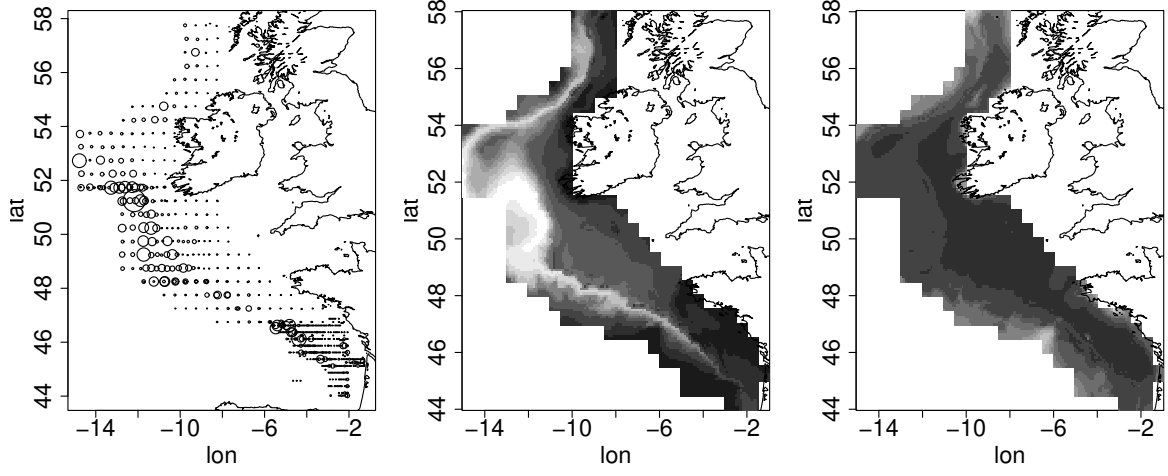
Figure 6: Left panel: locations of mackerel egg samples with symbol areas proportional to egg density per square metre of sea surface. Middle panel: model predicted square root egg density over the survey area. Right panel: $5\times$ the standard error of the estimates in the middle panel, on the same scale as the middle panel.
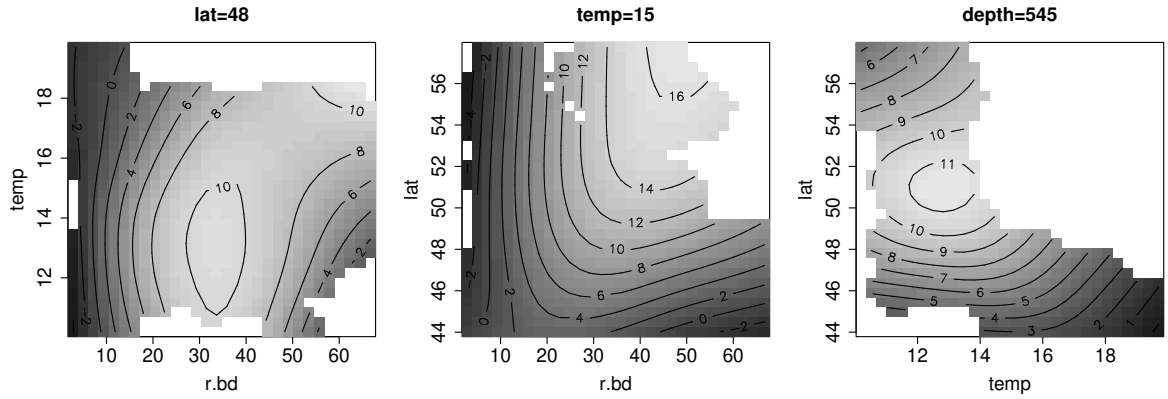

Figure 7: Each figure shows f_1 against two covariates with the other covariate held at its mean value in the data set. The function is only plotted for values of the covariates sufficiently close to values observed in the data. Note that this figure serves to emphasize the importance of a useful smoothness range: a tensor product smooth with a single penalty, with coefficient matrix constructed from a Kronecker product of marginal penalty matrices, could not have represented a function as smooth as this.