Coverage Properties of Confidence Intervals for Generalized Additive Model Components*

Giampiero Marra
Statistical Science, University College London
Gower Street, London WC1E 6BT, U.K.

Simon N. Wood
Mathematical Sciences, University of Bath
Claverton Down, Bath BA2 7AY, U.K.

giampiero@stats.ucl.ac.uk

June 29, 2011

Abstract

We study the coverage properties of Bayesian confidence intervals for the smooth component functions of generalized additive models (GAMs) represented using any penalized regression spline approach. The intervals are the usual generalization of the intervals first proposed by Wahba and Silverman in 1983 and 1985, respectively, to the GAM component context. We present simulation evidence showing these intervals have close to nominal 'across-the-function' frequentist coverage probabilities, except when the truth is close to a straight line/plane function. We extend the argument introduced by Nychka in 1988 for univariate smoothing splines to explain these results. The theoretical argument suggests that close to nominal coverage probabilities can be achieved, provided that heavy oversmoothing is avoided, so that the bias is not too large a proportion of the sampling variability. Otherwise, because the Bayesian intervals account for bias and variance, the coverage probabilities are surprisingly insensitive to the exact choice of smoothing parameter. The theoretical results allow us to derive alternative intervals from a purely frequentist point of view, and to explain the impact that the neglect of smoothing parameter variability has on confidence interval performance. They also suggest switching the target of inference for component-wise intervals away from smooth components in the space of the GAM identifiability constraints. Instead intervals should be produced for each function as if only the other model terms were subject to identifiability constraints. If this is done then coverage probabilities are improved.

Key words: Bayesian confidence interval; Generalized additive model; Penalized regression spline.

1 Introduction

This paper is about the *coverage properties* of confidence intervals for the components of generalized additive models (GAMs), when the components are represented using penalized regression splines, and the smoothing parameters are estimated as part of fitting. Firstly we provide a novel adaptation of Nychka's (1988) analysis of the one dimensional smoothing spline case, to explain why the commonly

*Research Report No. 313, Department of Statistical Science, University College London. Date: June 2011.

used Bayesian confidence intervals for GAM components should have close to nominal 'across-the-function' coverage probabilities, in most cases. This also allows us to derive alternative intervals from a purely frequentist point of view, and to explain the impact that the neglect of smoothing parameter variability has on confidence interval performance. Secondly our analysis shows that we can expect component wise intervals to perform poorly in the case when the true component is 'close to' the null space of the component's smoothing penalty. Thirdly guided by the previous two results we provide a proposal to partially alleviate the cases that give poor interval performance. Finally we provide a simulation study which supports our findings, and illustrate the confidence intervals using a real dataset.

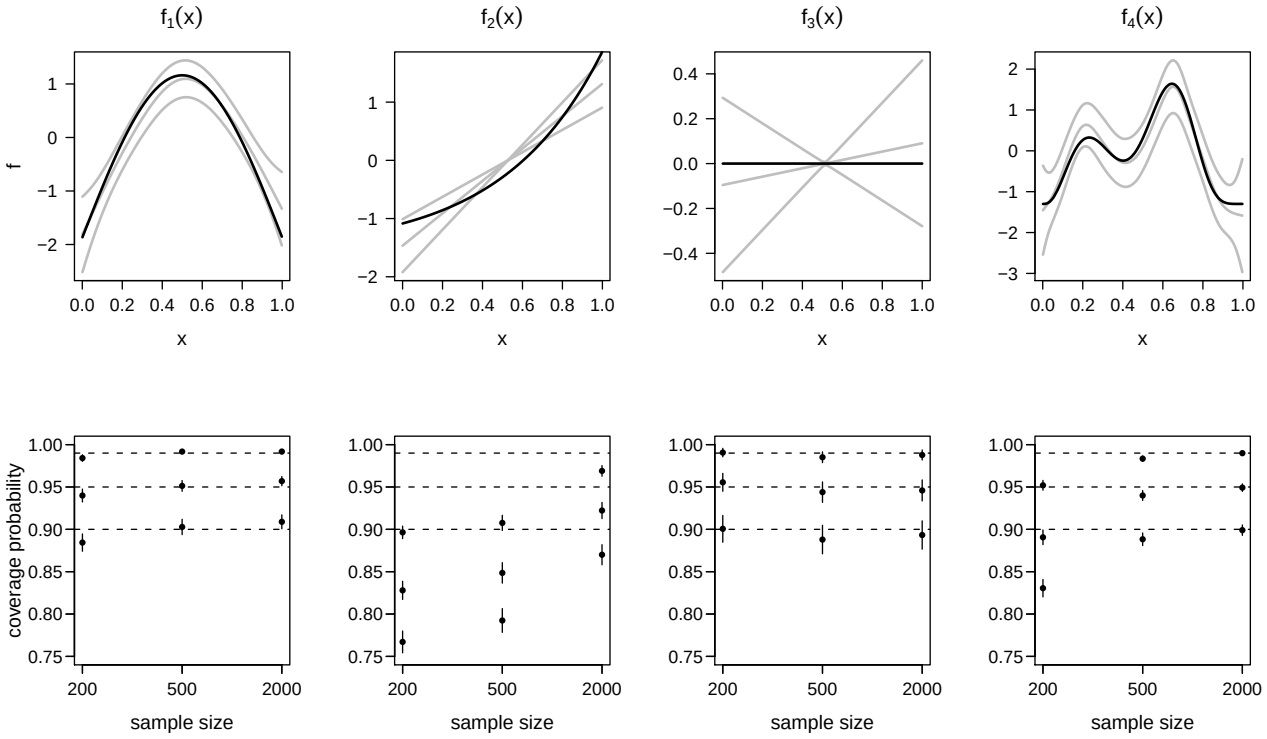

Figure 1: Results for component-wise Bayesian intervals for Bernoulli simulated data at three sample sizes. Observations were generated as $\text{logit}\{\mathbb{E}(Y_i)\} = \alpha + z_i + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i})$, where Y_i followed a bernoulli distribution and uniform covariates on the unit interval with correlations equal to 0.5 were employed (see section 4.1 for details). The function definitions are given in Table 1. The functions were scaled to have the same magnitude in the linear predictor and then the sum rescaled to produce probabilities in the range [0.02,0.98]. 1000 replicate datasets were then generated and GAMs fitted using penalized thin plate regression splines (Wood, 2003) with basis dimensions equal to 10, 10, 10 and 20, respectively, and penalties based on second-order derivatives. Multiple smoothing parameter selection was by generalized AIC (Wood, 2008). Displayed in the top row are the true functions, indicated by the black lines, as well as example estimates and 95% Bayesian confidence intervals (gray lines) for the smooths involved. The bottom row summarizes the interval coverage probability results. • represents the mean coverage probability from the 1000 across-the-function coverage proportions of the intervals, vertical lines show ±2 standard error bands for the mean coverage probabilities, and dashed horizontal lines show the nominal coverage probabilities considered.

Figure 1 illustrates our findings in one figure. Its 4 columns relate to the 4 functional components of a GAM for some simulated binary data, with the thick black curves in the upper panels showing the true component functions. Overlaid on each upper panel are example Bayesian confidence intervals for each component. The lower panels illustrate the realized across-the-function coverage probabilities

for such intervals over 1000 replicate datasets, at each of 3 sample sizes. Coverage properties are good for all components, except for f_2, where the truth is close to a function in the smoothing penalty null space for the term. In this case the function is often estimated to be in the null space of the penalty (as in the confidence interval shown for f_2 in the figure). When combined with the identifiability constraints necessary for GAM estimation, estimates in the penalty null space often have confidence intervals that are of zero width at some point. Zero or narrow width implies that the estimation bias exceeds its variance, and this in turn leads to a breakdown in the theoretical argument that underpins close to nominal interval coverage. (Interestingly it appears that this is the only situation in which the results show substantial sensitivity to smoothing parameter estimation.) A solution to this poor coverage problem is to modify the target of inference. Rather than basing intervals on estimates of the component functions in the null space of the identifiability constraints, we can base them on estimates of the function in the space in which the component of interest is unconstrained, while all identifiability constraints are carried by the other model components. Such intervals never have vanishing width, and consequently display closer to nominal coverage probabilities in simulations.

The remainder of this paper develops these results more fully. Section 2 specifies the model framework in more detail and establishes notation. Section 3 then presents the theoretical arguments for close to nominal component wise coverage properties in general, but poor coverage properties when 'close to' the penalty null space. It also suggests how to avoid poor interval coverage. Section 4 then presents the results of a simulation study supporting our findings, whereas Section 5 illustrate the confidence intervals using data on diabetes in Pima Indian women.

2 Preliminaries

This section covers some standard but essential background on GAM estimation. A GAM (Hastie and Tibshirani, 1990) is a generalized linear model (GLM; McCullagh and Nelder 1989) with a linear predictor involving smooth functions of covariates:

$$g\{\mathbb{E}(Y_i)\} = \mathbf{X}^*_i\boldsymbol{\theta}^* + \sum_j^J f_j(x_{ji}), \quad i = 1, \ldots, n \tag{1}$$

where $g(\cdot)$ is a smooth monotonic twice differentiable link function, Y_i is a univariate response that follows an exponential family distribution, $\mathbf{X}^*_i$ is the ith row of $\mathbf{X}^*$, which is the model matrix for any strictly parametric model components, with corresponding parameter vector $\boldsymbol{\theta}^*$, and the f_j are smooth functions of the covariates x_j, which may be vector covariates (so x_{ji} denotes the i^{th} observation of the j^{th} covariate). The f_j are subject to identifiability constraints, such as $\sum_i^n f_j(x_{ji}) = 0$ for all j. In this paper we concentrate on the case in which the f_j are represented using regression spline type bases, with associated measures of function roughness that can be expressed as quadratic forms in the basis coefficients (see Wood, 2006a for an overview). Given such bases the GAM is in principle just a parametric GLM and can be estimated as such, but to avoid overfitting it is necessary to estimate such models by penalized maximum likelihood estimation, in which the roughness measures are used to control overfit. In practice, the penalized likelihood is maximized by penalized iteratively reweighted

least squares (P-IRLS), so that the GAM is fitted by iterative minimization of the problem

$$\|\sqrt{\mathbf{W}^{[k]}}(\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta} \quad \text{w.r.t.} \quad \boldsymbol{\beta}.$$

k is the iteration index, $\mathbf{z}^{[k]} = \mathbf{X}\boldsymbol{\beta}^{[k]} + \mathbf{G}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$, $\mu_i^{[k]}$ is the current model estimate of $\mathbb{E}(Y_i)$, $\mathbf{G}^{[k]}$ is a diagonal matrix such that $G_{ii}^{[k]} = g'(\mu_i^{[k]})$, $\mathbf{W}^{[k]}$ is a diagonal matrix given by $W_{ii}^{[k]} = [G_{ii}^{[k]2} V(\mu_i^{[k]})]^{-1}$ where $V(\mu_i^{[k]})$ gives the variance of Y_i to within a response distribution scale parameter, ϕ, $\mathbf{X}$ includes the columns of $\mathbf{X}^*$ and columns representing the spline bases for the f_j, while $\boldsymbol{\beta}$ contains $\boldsymbol{\theta}^*$ and all the smooth coefficient vectors, $\boldsymbol{\beta}_j$. The $\mathbf{S}_j$ are matrices of known coefficients such that $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$ measures the roughness of f_j. The λ_j are smoothing parameters that control the trade-off between fit and smoothness. Smoothing parameter selection can be achieved, for example, by minimizing the Generalized Cross Validation score (GCV; Craven and Wahba, 1979) if a scale parameter has to be estimated, or the generalized Akaike's information criterion otherwise (AIC; Akaike, 1973). Fast stable computational methods for multiple smoothing parameter estimation are provided in Wood (2004, 2008).

Inference for *univariate* spline models can be effectively achieved using the Bayesian confidence intervals proposed by Wahba (1983) or Silverman (1985). As theoretically shown by Nychka (1988), for the case of univariate models whose Bayesian intervals have close to constant width, a very interesting feature of these intervals is that they work well when evaluated by a frequentist criterion, provided coverage is measured 'across-the-function' rather than pointwise. Specifically, consider the model

$$Y_i = f(x_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),$$

where f is a smooth function of x_i and the ϵ_i are mutually independent. According to Nychka's results, if f is estimated using a cubic smoothing spline for which the smoothing parameter is sufficiently reliably estimated (e.g. by GCV) that the bias in the estimates is a modest fraction of the mean squared error for $f(x)$, then the average coverage probability (ACP)

$$\text{ACP} = \frac{1}{n}\sum_{i=1}^{n} \Pr[f(x_i) \in BI_\alpha(x_i)]$$

is very close to the nominal level $1-\alpha$, where $BI_\alpha(x)$ indicates the $(1-\alpha)100\%$ Bayesian interval for $f(x)$ and α the significance level. This agreement occurs because the Wahba/Silverman type intervals include both a bias and variance component. Note that for convenience we define ACP only over the design points, rather than the whole function (but this restriction makes no practical difference for a smooth well sampled function).

Wahba/Silverman type intervals are straightforward to extend to non-Gaussian settings and GAM components (e.g. Gu, 1992; Gu, 2002; Gu and Wahba, 1993; Ruppert et al., 2003; Wood, 2006b). For example, by working in terms of the random vector $\mathbf{z}$ (the converged $\mathbf{z}^{[k]}$), it has been shown (Wood,

2006b) that the large sample posterior distribution for the regression spline coefficients of a GAM is

$$\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\lambda},\phi \sim N(\hat{\boldsymbol{\beta}},\mathbf{V}_{\boldsymbol{\beta}}), \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum penalized likelihood estimate of $\boldsymbol{\beta}$ which is of the form $(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{z}$, $\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\phi$, $\mathbf{W}$ is the diagonal weight matrix at convergence of the P-IRLS algorithm used for fitting, and $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. Notice that $\mathbf{W}$ and ϕ are equal to the identity matrix and σ^2, respectively, when a normal response is assumed and an identity link function selected; furthermore, in the normal errors case, result (2) holds independently of asymptotic arguments.

While the extension of the Wahba/Silverman results to the GAM component setting is well established, the coverage properties of the extension are unclear: Nychka's (1988) analysis has not previously been extended to the GAM component setting. The remainder of this paper therefore attempts to provide such an extension, and to examine component wise coverage properties by simulation.

3 Confidence Intervals

The aim of this section is to develop a construction of variable width component-wise intervals. The primary purpose of this construction is to reveal the coverage properties of the usual component-wise extension of Wahba/Silverman type intervals as discussed, for instance, by Gu and Wahba (1993), Fahrmeir et al. (2004), Ruppert et al. (2003) and Wood (2006b), to name a few. The initial part of Section 3.1 is similar to the construction that can be found in Ruppert et al. (2003, Section 6.4), but thereafter we are forced to follow a line more similar to Nychka (1988) in order to establish coverage properties (which the Ruppert et al. (2003) derivation does not reveal). Our theoretical derivations explain why Wahba/Silverman type intervals work well in a frequentist setting, and show why intervals for smooth components that are in the penalty null space are problematic. The good coverage probabilities obtained using result (2) are hence explained and a remedy to the near-straight-line/plane case is introduced. The theoretical arguments also allow us to derive alternative intervals when a purely frequentist approach is adopted.

For clarity the normal error, identity link, case is covered first, with the generalization to GAMs discussed subsequently. We need to start by establishing some preliminary results.

3.1 Estimation of $\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2/n_b$, and σ^2

The subsequent arguments will require that we can estimate the expected mean squared error of linear transformations of the coefficient estimates $\hat{\boldsymbol{\beta}}$, so this needs to be addressed first.

Consider, then, an arbitrary linear transformation defined by a matrix of fixed coefficients $\mathbf{B}$, with n_b rows. We seek an estimate for the expected mean squared error, $\mathbb{E}(M_B)$, of $\mathbf{B}\hat{\boldsymbol{\beta}}$. From the Bayesian approach we have $\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\lambda},\sigma^2 \sim N(\hat{\boldsymbol{\beta}},\mathbf{V}_{\boldsymbol{\beta}})$. Taking expectations with respect to this distribution implies

$$\mathbb{E}(M_B) = \frac{1}{n_b}\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2 = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^{\mathsf{T}}\mathbf{B}\mathbf{V}_{\boldsymbol{\beta}}),$$

but this relies on adopting a fully Bayesian approach in which the random variability is ascribed to $\boldsymbol{\beta}$ rather than $\hat{\boldsymbol{\beta}}$, which will make it difficult to establish interval coverage properties. This section shows how we can instead estimate $\mathbb{E}(M_B)$ directly, or by making more limited use of some of the prior assumptions underpinning the Bayesian smoothing model.

First let $\tilde{\boldsymbol{\beta}}$ denote the unpenalized, and hence unbiased, estimate of $\boldsymbol{\beta}$. Let $\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ be the matrix such that $\hat{\boldsymbol{\beta}} = \mathbf{F}\tilde{\boldsymbol{\beta}}$ ($\mathbf{X}$ is the full model matrix here). It follows immediately that $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbf{F}\boldsymbol{\beta}$. It is then routine to show that the mean square error can be partitioned into a variance term and a mean squared bias term

$$\mathbb{E}(M_B) = \frac{1}{n_b}\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{V}_{\hat{\boldsymbol{\beta}}}) + \frac{1}{n_b}\|\mathbf{B}(\mathbf{F}-\mathbf{I})\boldsymbol{\beta}\|^2, \tag{3}$$

where $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S})^{-1}\sigma^2$. Two approaches to estimation are now possible.

1. Plugging $\hat{\boldsymbol{\beta}}$ into the right hand side of (3) yields the obvious estimator

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{V}_{\hat{\boldsymbol{\beta}}}) + \frac{1}{n_b}\|\mathbf{B}(\mathbf{F}-\mathbf{I})\hat{\boldsymbol{\beta}}\|^2. \tag{4}$$

2. Alternatively, we could seek to estimate (3) by assuming that the mean and variance assumptions of the Bayesian model hold and using the corresponding expected value of the right hand side of (3) as the estimator of $\mathbb{E}(M_B)$.

   Progress in this direction is most easily made by re-parameterizing using a 'natural' Demmler-Reinsch type parameterization (e.g. Wood, 2006a, 4.10.4). Forming the QR decomposition $\mathbf{X} = \mathbf{QR}$ and the eigen decomposition $\mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$ (where $\mathbf{U}$ is the orthogonal matrix of eigenvectors and $\mathbf{D}$ the diagonal matrix of eigenvalues), then the reparameterization leads to the new smooth coefficient vector $\mathbf{U}^\mathsf{T}\mathbf{R}\boldsymbol{\beta}$ and matrix of fixed coefficients $\mathbf{B}\mathbf{R}^{-1}\mathbf{U}$. In this parameterization the most important matrices are diagonal, for example $\mathbf{V}_{\boldsymbol{\beta}}/\sigma^2 = \mathbf{F} = (\mathbf{I}+\mathbf{D})^{-1}$ and $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{I}+\mathbf{D})^{-2}\sigma^2$. The prior assumptions then become $\mathbb{E}(\boldsymbol{\beta}) = \mathbf{0}$, $\mathrm{var}(\beta_k) = D_{kk}^{-1}\sigma^2$, unless $D_{kk} = 0$ (in which case the variance turns out to be immaterial), and the covariances are zero. It is then routine to show that

$$\frac{1}{n_b}\mathbb{E}\|\mathbf{B}(\mathbf{F}-\mathbf{I})\boldsymbol{\beta}\|^2 = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{H})\sigma^2,$$

   where $\mathbf{H}$ is a diagonal matrix with elements $H_{kk} = D_{kk}/(1+D_{kk})^2$. Recognizing that $\mathbf{H}\sigma^2 = \mathbf{V}_{\boldsymbol{\beta}} - \mathbf{V}_{\hat{\boldsymbol{\beta}}}$, and substituting into (3) we have the estimate

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n_b}\mathbb{E}\|\mathbf{B}\widehat{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\|^2 = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{V}_{\boldsymbol{\beta}}).$$

   Reversing the re-parameterization confirms that this is simply

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n_b}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}(\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S})^{-1})\sigma^2 \tag{5}$$

in the original parameterization. The point here is that if we are prepared to accept the Bayesian prior mean and variance assumptions as a reasonable model, then (5) follows as an estimate of $\mathbb{E}(M_B)$.

If $\mathbf{B} = \mathbf{X}$ then (5) leads easily to the usual estimator

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - \mathrm{tr}(\mathbf{F})}. \tag{6}$$

Alternatively, if we use (4), then we obtain

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 - \|\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2}{n - 2\mathrm{tr}(\mathbf{F}) + \mathrm{tr}(\mathbf{F}\mathbf{F}^{\mathsf{T}})}. \tag{7}$$

3.2 Intervals

Having completed the necessary preliminaries, we now consider the construction of intervals for some component function, $f(x)$, of a model. $f(x)$ is one of the functions $f_j(x_j)$, but to avoid clutter the index j is dropped here. Suppose that the vector of values of f evaluated at the observed covariate values can be written as $[f(x_1), f(x_2), \ldots, f(x_n)]^{\mathsf{T}} \equiv \mathbf{f} = \mathbf{X}\boldsymbol{\beta}$. Here $\mathbf{X}$ is the model matrix for just one model component: the matrix mapping the vector of all model coefficients to the evaluated values of just one smooth component (so, many of the columns of $\mathbf{X}$ may be zero). Our approach will modify Nychka's (1988) construction in order to obtain intervals of variable width, which are also applicable in the case where the function is only a component of a larger model.

Given some convenient constants, C_i, we seek a constant, A, such that

$$\mathrm{ACP} = \frac{1}{n}\mathbb{E}\left\{\sum_i \mathbb{I}(|\hat{f}(x_i) - f(x_i)| \leq z_{\alpha/2}A/\sqrt{C_i})\right\} = 1 - \alpha, \tag{8}$$

where $\mathbb{I}$ is an indicator function, α is a constant between 0 and 1 and $z_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. To this end, define $b(x) = \mathbb{E}\{\hat{f}(x)\} - f(x)$ and $v(x) = \hat{f}(x) - \mathbb{E}\{\hat{f}(x)\}$, so that $\hat{f} - f = b + v$. Defining I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$ we have

$$\begin{aligned} \mathrm{ACP} &= \Pr\left(|b(x_I) + v(x_I)| \leq z_{\alpha/2}A/\sqrt{C_I}\right) \\ &= \Pr\left(|B + V| \leq z_{\alpha/2}A\right) \end{aligned}$$

by definition of the randomised scaled bias and scaled variance, $B = \sqrt{C_I}b(x_I)$ and $V = \sqrt{C_I}v(x_I)$, respectively. Approximation of the distribution of $B + V$ is central to the following argument.

Let $[b(x_1), b(x_2), \ldots, b(x_n)]^{\mathsf{T}} \equiv \mathbf{b} = \mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f}$. Hence, defining $\mathbf{c} = (\sqrt{C_1}, \sqrt{C_2}, \ldots, \sqrt{C_n})^{\mathsf{T}}$, we have

$$\mathbb{E}(B) = \sum_i \frac{1}{n} b(x_i)\sqrt{C_i} = \mathbf{c}^{\mathsf{T}}(\mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f})/n.$$

This quantity is the average bias across the component function and in practice tends to be very

small, unless heavy oversmoothing is employed. In any case it can be estimated as

$$\widehat{\mathbb{E}(B)} = \mathbf{c}^\mathsf{T}(\hat{\hat{\mathbf{f}}} - \hat{\mathbf{f}})/n = \mathbf{c}^\mathsf{T}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/n. \tag{9}$$

Now consider V. Defining $[v(x_1), v(x_2), \ldots, v(x_n)]^\mathsf{T} \equiv \mathbf{v} = \hat{\mathbf{f}} - \mathbb{E}(\hat{\mathbf{f}})$, we have $\mathbb{E}(\mathbf{v}) = \mathbf{0}$, and hence

$$\mathbb{E}(V) = \sum_i \frac{1}{n} v(x_i)\sqrt{C_i} = 0.$$

The covariance matrix of $\mathbf{v}$ is, $\mathbf{V}_{\hat{\mathbf{f}}} = \mathbf{X}\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\mathsf{T}$, the same as that of $\hat{\mathbf{f}}$. Hence

$$\mathrm{var}(V) = \sum_i \frac{1}{n}\mathbb{E}\{v(x_i)^2 C_i\} = \mathrm{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})/n,$$

where $\mathbf{C}$ is the diagonal matrix with leading diagonal elements C_i. Now since $\mathbf{v} \sim N(\mathbf{0}, \mathbf{V}_{\hat{\mathbf{f}}})$ (multivariate normality of $\mathbf{v}$ following from multivariate normality of $\hat{f}$), V is a mixture of normals, which is inconvenient unless $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ is independent of i. If this constant variance assumption holds then $V \sim N(0, \mathrm{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})\sigma^2/n)$ (and the lack of dependence on i implies a lack of dependence on $b(x_i)$, implying independence of B and V).

It is the distribution of $B+V$ that is needed. $\mathbb{E}(B+V) = \mathbb{E}(B)$ and by construction $\mathrm{var}(B+V) = \mathbb{E}(M) - \mathbb{E}(B)^2$ where

$$M = \frac{1}{n}\sum_i C_i\{\hat{f}(x_i) - f(x_i)\}^2 = \|\sqrt{\mathbf{C}}(\hat{\mathbf{f}} - \mathbf{f})\|^2/n = \|\sqrt{\mathbf{C}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n.$$

Now we can exactly re-use Nychka's (1988) argument: provided B is small relative to V then $B + V$ will be approximately normally distributed, i.e. approximately

$$B + V \sim N(\mathbb{E}(B), \mathbb{E}(M) - \mathbb{E}(B)^2).$$

The assumption that B is small relative to V is examined in more detail in the Appendix. We can estimate $\mathbb{E}(B)$ (either as 0 or using (9)) and $\mathbb{E}(M)$ (using (4) or (5)). So, defining $\hat{\sigma}^2_{bv} = \widehat{\mathbb{E}(M)} - \widehat{\mathbb{E}(B)^2}$, we have the approximate result

$$B + V \sim N(\widehat{\mathbb{E}(B)}, \hat{\sigma}^2_{bv}).$$

Routine manipulation then results in

$$\hat{f}(x_i) - \widehat{\mathbb{E}(B)}/\sqrt{C_i} \pm z_{\alpha/2}\hat{\sigma}_{bv}/\sqrt{C_i}, \tag{10}$$

as the definition of intervals achieving close to $1-\alpha$ ACP (i.e. $A = \sigma_{bv}$). Intuitively, it is the fact that the convolution of B and V is close to a normal that leads the intervals to have good across-the-function coverage.

So far the choice of C_i has not been discussed, but the constant variance requirement, for $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ to be independent of i, places strong restrictions on what is possible here. Two choices are interesting:

1. $C_i^{-1} = [\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ ensures that the constant variance assumption is met exactly. Note that in this case, if we use (5) as the expected mean squared error estimate,

$$\widehat{\mathbb{E}(M)} = \frac{\hat{\sigma}^2}{n} \sum_i \frac{[\mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathsf{T}}]_{ii}}{[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}}.$$

   In effect the resulting intervals are using the frequentist covariance matrix $\mathbf{V}_{\hat{\mathbf{f}}}$, but 'scaled up' to the 'size' of the Bayesian covariance matrix $\mathbf{V}_{\mathbf{f}} = \mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathsf{T}}$. Alternatively using (4) we have

$$\widehat{\mathbb{E}(M)} = 1 + \|\sqrt{\mathbf{C}}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2/n, \tag{11}$$

   and combining this with estimator (5) yields intervals

$$\hat{f}(x_i) - \widehat{\mathbb{E}(B)}\sqrt{[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}} \pm z_{\alpha/2}\sqrt{\left\{\widehat{\mathbb{E}(M)} - \widehat{\mathbb{E}(B)^2}\right\}[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}}. \tag{12}$$

2. $C_i^{-1} = [\mathbf{V}_{\mathbf{f}}]_{ii}$. If $[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii} \approx \gamma[\mathbf{V}_{\mathbf{f}}]_{ii}$ for some constant γ, then this choice approximately meets the constant variance assumption. If we use the (typically accurate) approximation $\widehat{\mathbb{E}(B)} \approx 0$, along with the mean squared error estimate (5), then the resulting intervals are exactly Bayesian intervals of the Wahba/Silverman kind, i.e.

$$\hat{f}(x_i) \pm z_{\alpha/2}\hat{\sigma}\sqrt{[\mathbf{V}_{\mathbf{f}}]_{ii}}. \tag{13}$$

So, we arrive at the key result. Given the derivation of the intervals started from (8), we expect component-wise Wahba/Silverman type intervals to have close to nominal coverage properties across the function.

Notice the limited role of smoothing parameter selection in the above:

1. The requirement for B to be smaller than V, requires that we choose smoothing parameters so as not to oversmooth too heavily, but otherwise the choice of smoothing parameter is rather unimportant.

2. Accuracy of $\widehat{\mathbb{E}(B)}$ depends on $\hat{f}$ being close to f, but this is exactly what smoothing parameter selection methods try to achieve. In any case this term is typically small enough to be practically negligible.

3. $\mathbb{E}(M)$ will be more reliably estimated by (4) if $\hat{f}$ is close to f: again this is what smoothing parameter selection tries to achieve. Similarly $\mathbb{E}(M)$ will be more reliably estimated by (5) if the prior variance assumption for $\boldsymbol{\beta}$ is plausible. This variance is controlled by the smoothing parameter for f, which the smoothing parameter selection method is designed to approximate well.

The forgoing immediately suggests the circumstances under which the intervals will behave poorly. If smoothing parameters are substantially over-estimated, so that we substantially oversmooth some

component, then the requirement for B to be smaller than V will be violated. Two situations are likely to promote oversmoothing. Firstly, if a true effect is almost in the penalty null space then the estimated smoothing parameter may tend to infinity, forcing the estimate into the penalty null space, and as we will see below, this can be problematic. A second possibility is that highly correlated covariates are likely to mean that it is difficult to identify which corresponding smoothing parameters should be high and which low. For example if one covariate has a very smooth effect, and another a very wiggly effect, but they are highly correlated, it is quite possible that their estimated effects will have the degrees of smoothness reversed. This means that one of the covariate effects is substantially oversmoothed. There is some evidence for this effect in the simulation results shown in figure 7, but it seems to be a less serious issue than the first problem.

In summary: we have shown that Bayesian component-wise variable width intervals, or our proposed alternative, for the smooth components of an additive model should achieve close to nominal across-the-function coverage probability, provided only that we do not oversmooth so heavily that average bias dominates the sampling variability for a term estimate. Beyond this requirement not to oversmooth too heavily, the results appear to have rather weak dependence on smoothing parameter values, suggesting that the neglect of smoothing parameter variability should not significantly degrade interval performance. Note however, that our results rely on several heuristic assumptions: i) that $E(B)$ is negligible, or can be reasonably well estimated by (11), (ii) that $E(M)$ can be reasonably well estimated by (4) or (5), (iii) that B is generally substantially smaller than V and, for Bayesian intervals, iv) that $[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii} \approx \gamma[\mathbf{V}_{\mathbf{f}}]_{ii}$. All are highly plausible, but are difficult to treat more rigorously.

3.3 Generalized additive model case

The results of sections 3.1 and 3.2 can be routinely extended to the *generalized* additive model case. In the case of section 3.1 the results follow as large sample approximations with the substitutions

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\phi, \tag{14}$$

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\phi, \tag{15}$$

$$\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}, \tag{16}$$

and $\sqrt{\mathbf{W}}\mathbf{X} = \mathbf{Q}\mathbf{R}$. Then (5) becomes

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1})\sigma^2.$$

Similarly, (6) becomes

$$\hat{\phi} = \frac{\|\sqrt{\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\|^2}{n-\mathrm{tr}(\mathbf{F})},$$

while the generalization of (7) is

$$\hat{\phi} = \frac{\|\sqrt{\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\|^2 - \|\sqrt{\mathbf{W}}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})\|^2}{n-2\mathrm{tr}(\mathbf{F})+\mathrm{tr}(\mathbf{F}\mathbf{F}^\mathsf{T})}.$$

Section 3.2 also follows as before, but again with the substitutions (14) – (16). The key requirement that $\mathbf{v} \sim N(\mathbf{0}, \mathbf{V}_{\hat{\mathbf{f}}})$ is now a large sample approximation, which follows from the large sample normality of $\hat{\boldsymbol{\beta}}$, which can readily be established (e.g. Wood, 2006b).

3.4 What the results explain

Our results explain the success of Bayesian, component-wise, variable width intervals, which is evidenced in the simulation study of the next section. More interestingly, they explain the cases where the Bayesian intervals fail. The major failure, evident from simulations (see Figure 1), occurs when a smooth component is close to a function in the null space of the component's penalty (i.e. to a straight line or plane, for the examples in this paper) and may therefore be estimated as exactly such a function. The component will have been estimated subject to an identifiability constraint, but when intervals are constructed subject to such a constraint, the observed coverage probabilities are poor. The preceding theory explains why. For example, when a term is estimated as a straight line but *subject to an identifiability constraint* then the associated confidence interval necessarily has width 0 where the line passes through the zero line. In this case the sampling variability must be smaller than the bias over some interval surrounding the point, and the assumption that B is less than V will fail, given also that the C_i associated with this interval will be very large.

The theory also suggests a remedy to this problem: compute each term's interval as if it alone were unconstrained, and identifiability was obtained by constraints on the other model terms (see section below).

Notice that the interval failure in the constrained straight line/plane case is not just the result of failing to meet the original Nychka (1988) constant width assumption: variable width intervals are quite acceptable under our extension of Nychka's argument: the problem occurs when the interval width shrinks to zero somewhere.

The poor component-wise coverages reported in Wood (2006b) are also explained by our results. The reported simulations were performed (in 2001/2) with the original smoothness selection method proposed in Wood (2000), and using uncorrelated covariates. This method alternated Newton updates of the smoothing parameters with a computationally cheap global search for an 'overall' smoothing parameter. While superficially appealing, this global search can miss a shallow minimum altogether, and place one or more smoothing parameters in a part of the smoothing parameter space in which the smoothness selection criteria is completely flat. Once smoothing parameters are at such a point, then they can only return by accident, in another global search, and this rarely happens. The upshot is that too many straight line/plane are estimated, and as we have seen this degrades the performance of the associated intervals. Step guarding procedures were eventually introduced in the code implementing Wood (2000) in the `R` package `mgcv`, but the importance of then repeating the simulations in Wood (2006b) was not appreciated by the second author, until now. The bootstrapping method proposed by Wood (2006b) to improve interval coverage appears to have actually been fixing an artefact of the smoothing parameter selection method, rather than a real deficiency in the confidence interval methods.

3.5 Component interval computation

As already mentioned we can improve interval performance by changing the target of inference, a little. For each component smooth term, intervals can be constructed by applying identifiability constraints to all other model components, but allowing the component of interest to 'carry the intercept'. This yields improved coverage probabilities, especially for the near straight line/plane case (see section below), and in the authors' experience also produces intervals that correspond more closely to the way in which users tend to interpret component-wise intervals. Specifically, assume that the linear predictor is

$$\eta_i = \mathbf{X}^*_i \boldsymbol{\beta}^* + \sum_j f_j(x_{ji})$$

where $\sum_i f_j(x_{ji}) = 0$ for all j. Suppose that $\boldsymbol{\beta}_j$ is the coefficient vector for f_j, so that the complete coefficient vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}^{*\mathsf{T}}, \boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \ldots)^\mathsf{T}$, with Bayesian covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$. Let $\alpha = \mathbf{1}^\mathsf{T}\mathbf{X}^*\boldsymbol{\beta}^*/n$ define the intercept. We are interested in inference about $\tilde{f}_j(x_j) = \alpha + f_j(x_j)$.

Now let $\mathbf{X}$ denote the full $n \times p$ model matrix, and assume that its first p^* columns are given by $\mathbf{X}^*$. Further, let $\mathbf{X}^j$ denote the matrix $\mathbf{X}$ with all columns zeroed, except for those corresponding to the coefficients of f_j, and the first p^*: these are modified to $X^j_{ik} = \sum_i X_{ik}/n$ for $1 \leq k \leq p^*$. Then $\mathbf{X}^j$ is the model matrix for $\tilde{f}_j$, i.e. $[\tilde{f}_j(x_{j1}), \tilde{f}(x_{j2}), \ldots, \tilde{f}(x_{jn})]^\mathsf{T} \equiv \tilde{\mathbf{f}}_j = \mathbf{X}^j\boldsymbol{\beta}$. Then the Bayesian covariance matrix for $\tilde{\mathbf{f}}_j$ is just $\mathbf{X}^j\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{j\mathsf{T}}$ and interval computation is straightforward.

This fix is quite important since the near straight-line/plane case is the one that is most important to get right for model selection purposes, both for deciding on which terms should be removed from a model, and which should be treated purely parametrically. In other words, the problem fixed by the preceding theory is of some importance in practice.

Notice that this proposal is not the same as basing intervals on the standard error of overall model predictions, with all but the covariate of interest held constant (e.g. Ruppert et al., 2003 Chapter 8). Such intervals are typically much wider than our proposal. Also note that trying to reduce the near straight-line problem by using alternative constraints will unnecessarily widen confidence intervals, since only the given constraint results in $\hat{\mathbf{f}}_j \perp \mathbf{1}$.

4 Simulation study

A Monte Carlo simulation study was conducted to compare the practical performance of several component-wise variable width confidence intervals. Specifically, based on equation (10), three kinds of intervals were considered:

1. Standard Bayesian intervals: the Wahba/Silverman type Bayesian intervals for smooth functions subject to identifiability constraints derived employing. i.e. (13) using (6).

2. Bayesian intervals with intercept: the same intervals as the previous ones but re-defining the model matrix $\mathbf{X}$ as discussed in section 3.5.

3. Alternative intervals with intercept: these intervals are derived using the model matrix of section 3.5, and are given by (12) using (7).

For the cases in which non-Gaussian data were considered, the substitutions of section 3.3 were carried out. Under a wide variety of settings, and employing the test functions displayed in Figures 1 and 2, the confidence intervals were compared in terms of coverage probabilities.

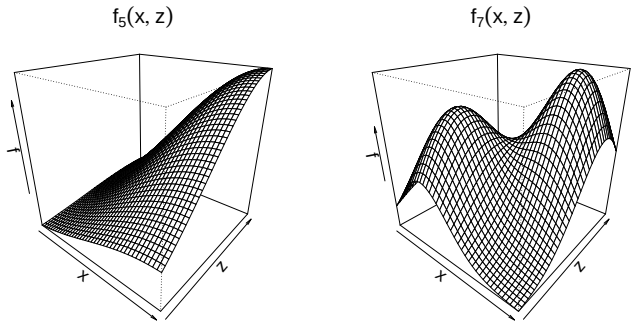

Figure 2: Two of the three two-dimensional test functions used in the linear predictor $\eta_{2,i}$. $f_6(x, z)$ is omitted since it is not informative.

$$
\begin{aligned}
&f_1(x) = 2\sin(\pi x)\\
&f_2(x) = e^{2x}\\
&f_3(x) = 0\\
&f_4(x) = x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10}\\
&f_5(x,z) = 0.7e^{-\{(-3x+3)^2+0.7(3z-3)^2\}/5}\\
&f_6(x,z) = 0\\
&f_7(x,z) = 0.075e^{-\frac{(x-0.3)^2}{0.3^2}-(z-0.3)^2} + 0.094e^{-\frac{(x-0.8)^2}{0.3^2}-\frac{(z-0.8)^2}{0.4^2}}
\end{aligned}
$$

Table 1: Test function definitions. f_1 - f_4 are plotted in Figure 1, and f_5 and f_7 in Figure 2.

4.1 Design

Two different linear predictors have been used for the simulation study. The first one was made up of a parametric component, z, plus four one-dimensional test functions (see Figure 1 and Table 1)

$$\eta_{1i} = \alpha + z_i + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}).$$

The second one was made up of a parametric component plus three two-dimensional test functions (see Figure 2 and Table 1)

$$\eta_{2i} = \alpha + z_i + f_5(x_{5i}, x_{6i}) + f_6(x_{7i}, x_{8i}) + f_7(x_{9i}, x_{10i}).$$

Uniform covariates on (0, 1) with equal correlations were obtained using the algorithm from Gentle (2003). For instance, using `R`, three uniform variables with correlations approximately equal to ρ could be obtained as follows:

```
library(mvtnorm)
cor <- array(c(1,rho,rho,rho,1,rho,rho,rho,1),dim=c(3,3))
var <- pnorm(rmvnorm(n,sigma=cor))
```

```
x1 <- var[,1]; x2 <- var[,2]; x3 <- var[,3]
```

This procedure was employed to obtain correlation among all covariates involved in a linear predictor. The cases in which ρ was set to 0, 0.5 and 0.9 were considered. The functions were scaled to have the same range and then summed. Data were simulated under four error models at each of three signal to noise ratio levels, at each of three sample sizes, $n = 200, 500, 2000$ (further details are given in Table 2). One-thousand replicate data sets were then generated at each sample size, distribution and error level combination, and generalized additive models fitted using penalized thin regression splines (Wood, 2003) based on second-order derivatives and with basis dimensions equal to 10, 10, 10 and 20 respectively for the first linear predictor, and with basis dimensions equal to 20, 20 and 50 for the second linear predictor. The smoothing parameters were chosen by GCV in the normal and gamma cases and by generalized AIC in the Poisson and binomial cases (Wood, 2004, 2008). For each replicate, 90%, 95% and 99% confidence intervals for the linear predictor as well as for each smooth function, evaluated at the simulated covariate values, were obtained. Then for each value of ρ, test function, sample size, signal to noise ratio, error model and $1 - \alpha$ level, an overall mean coverage probability from the resulting 1000 across-the-function coverage proportions was taken, and its standard deviation calculated.

	binomial	*gamma*	*Gaussian*	*Poisson*
$g(\mu)$	*logit*	*log*	*identity*	*log*
$l \leq \eta \leq u$	$[0.02, 0.98]$	$[0.2, 3]$	$[0, 1]$	$[0.2, pmax]$
s/n	n_{bin}= 1, 3, 5	$\phi = 0.6, 0.4, 0.2$	σ= 0.4, 0.2, 0.1	$pmax = 3, 6, 9$

Table 2: Observations were generated from the appropriate distribution with true response means, laying in the specified range, obtained by transforming the linear predictors by the inverse of the chosen link function. l, u and s/n stand for lower bound, upper bound and signal to noise ratio parameter, respectively. The linear predictor for the binomial case was scaled to produce probabilities in the range $[0.02, 0.98]$; observations were then simulated from binomial distributions with denominator n_{bin}. In the gamma case the linear predictor was scaled to have range $[0.2, 3]$ and three levels of ϕ used. For the Gaussian case normal random deviates with mean 0 and standard deviation σ were added to the true expected values, which were then scaled to lay in $[0, 1]$. The linear predictor of the Poisson case was scaled in order to yield true means in the interval $[0.2, pmax]$.

4.2 Coverage probability results

To save space, not all simulation results are shown. Instead the most important examples chosen. We have been careful to choose plots that are representative of the results in general, so that intuition gained from the plots shown fairly reflects the intuition that would be gained from looking at all the plots from the study.

Figures 3, 4 and 6 show coverage probability results when covariates are moderately correlated ($\rho = 0.5$) and data are generated using linear predictor $\eta_{1,i}$. Figure 8 refers to the case when employing $\eta_{2,i}$. Figures 5 and 7 serve to show the impact that covariate correlation has on confidence interval performance.

4.2.1 Results for standard Bayesian intervals

The standard Bayesian intervals yield realized coverage probabilities that appear to be fairly close to their nominal values in most cases, even at small sample sizes when the signal-to-noise ratio is not too low. The exception is for function $f_2(x)$ where the smoothing parameter methods tend to select more straight lines than they should really be selecting. However, as explained in section 3.4, it is because of the identifiability constraint that the confidence intervals are too narrow, and the coverage probabilities are poor as a consequence of the violation of the assumption that B is less than V. It is worth observing that when not much information is present in the data, it is likely to come up with straight line estimates for smooth functions like $f_2(x)$. In combination with the identifiability constraint this has a detrimental effect on confidence interval performance, which is why confidence intervals should really be constructed including the intercept of the model.

4.2.2 Results for Bayesian intervals 'with intercept'

As suggested in section 3.4, interval performance can be improved by allowing the component of interest to 'carry the intercept'. When this is done, the component-wise Bayesian intervals produce overall better coverage probabilities, especially for the near straight line/plane case; see functions $f_2(x)$ and $f_5(x, z)$.

4.2.3 Results for alternative intervals 'with intercept'

The alternative intervals also exhibit good component-wise interval performance but with slower convergence rate: this is probably because of the higher number of component quantities that have to be estimated for these intervals.

4.2.4 Impact of covariate correlation

Covariate correlation has an impact on confidence interval performance. Figures 5, 6 and 7 show the coverage probabilities obtained for Poisson data when correlated covariates were generated using $\rho = 0$, 0.5 and 0.9 respectively. It can be seen that although mild correlation does not spoil the coverage probabilities, a heavier one degrades some of them. As explained in section 3.2, the confidence intervals exhibit good practical performance if the smoothing parameters are chosen such that the estimated smooth components are not too heavily oversmoothed, but this is less likely to happen when covariates are heavily correlated. In this case, when $\rho = 0.9$ the covariates are quite confounded and this may lead to heavy oversmoothing for some component function of a GAM (often with some other component correspondingly undersmoothed, but this is less detrimental). Looking at the coverage probability results from the standard Bayesian intervals and comparing the results across the different values of ρ reveal that the coverages for $f_1(x)$ and $f_2(x)$ worsen. This means that for these two functions a smoother estimate is often selected, and this is why interval performance degrades. Specifically, the major failure evident from our results occurs for $f_2(x)$ where a substantial number of straight line estimates are selected. On the other hand, it would be rather remarkable if a smoothing method could select the right curve for a function which is so close to a straight line,

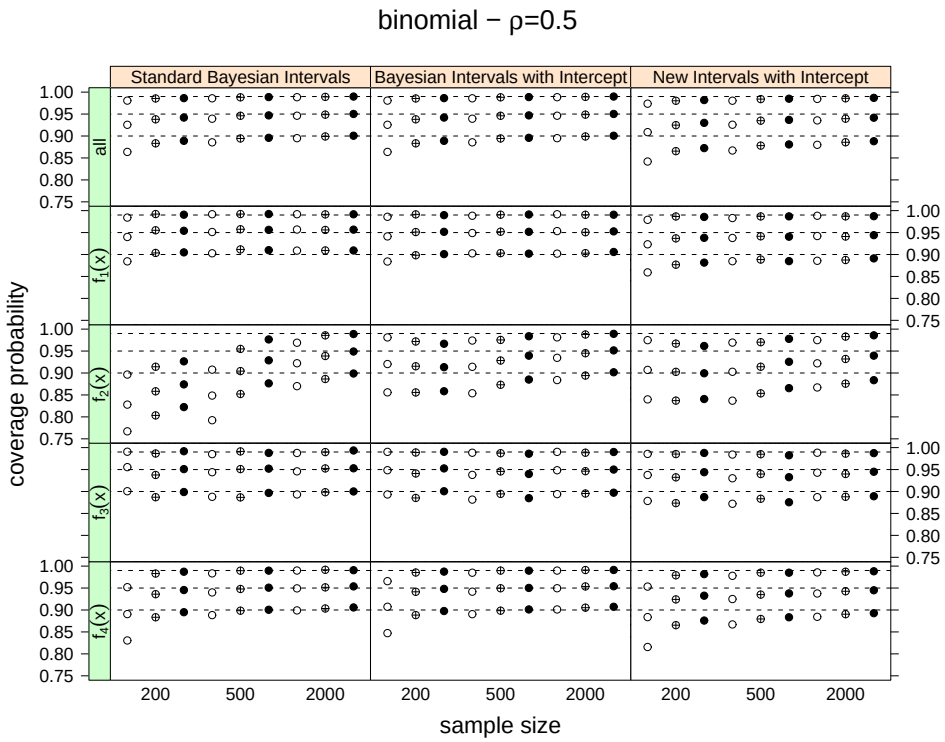

Figure 3: Coverage probability results for binomial data generated using $\eta_{1,i}$ as a linear predictor. Covariate correlation was equal to 0.5. Details are given in section 4. $\circ$, $\oplus$ and $\bullet$ stand for high, medium and low noise level respectively. Standard error bands are not reported since they are smaller than the plotting symbols. Notice the improvement in the performance of the component-wise intervals for $f_2(x)$, when the intercept is included in the calculations.

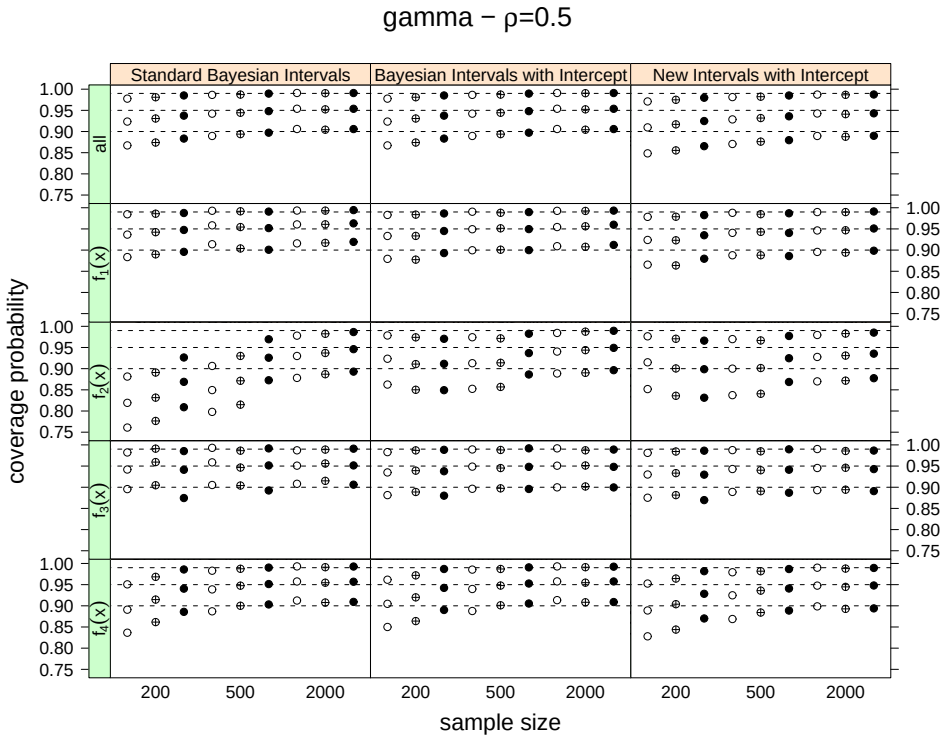

Figure 4: Coverage probability results for gamma data. Details are given in the caption of Figure 3.

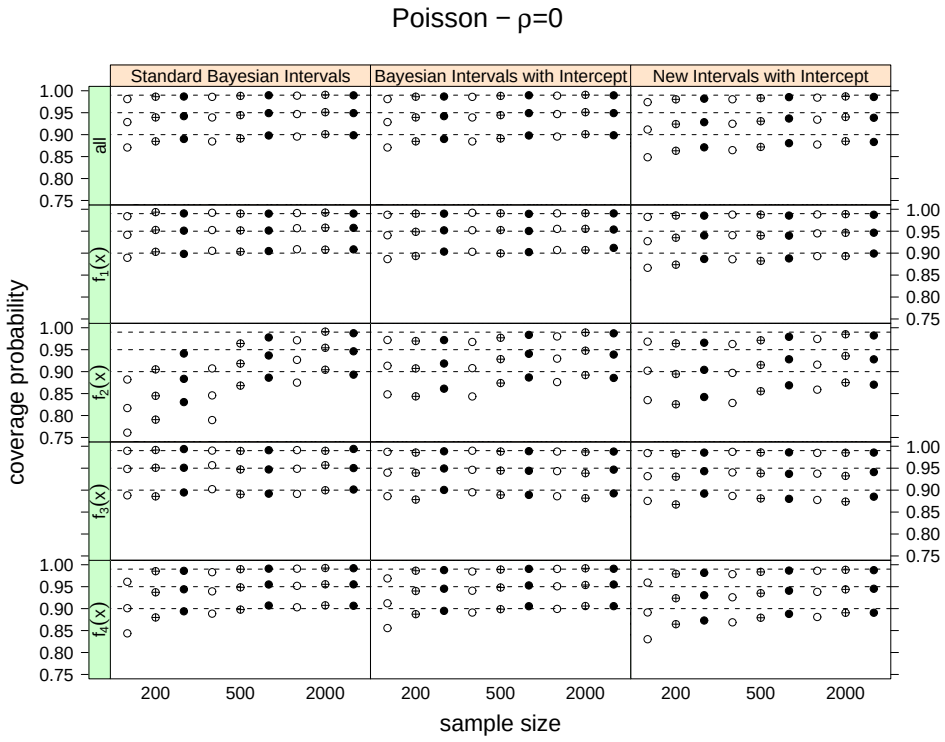

Figure 5: Coverage probability results for Poisson data for the case in which correlated uniform covariates were obtained setting $\rho = 0$. Details are given in the caption of Figure 3.

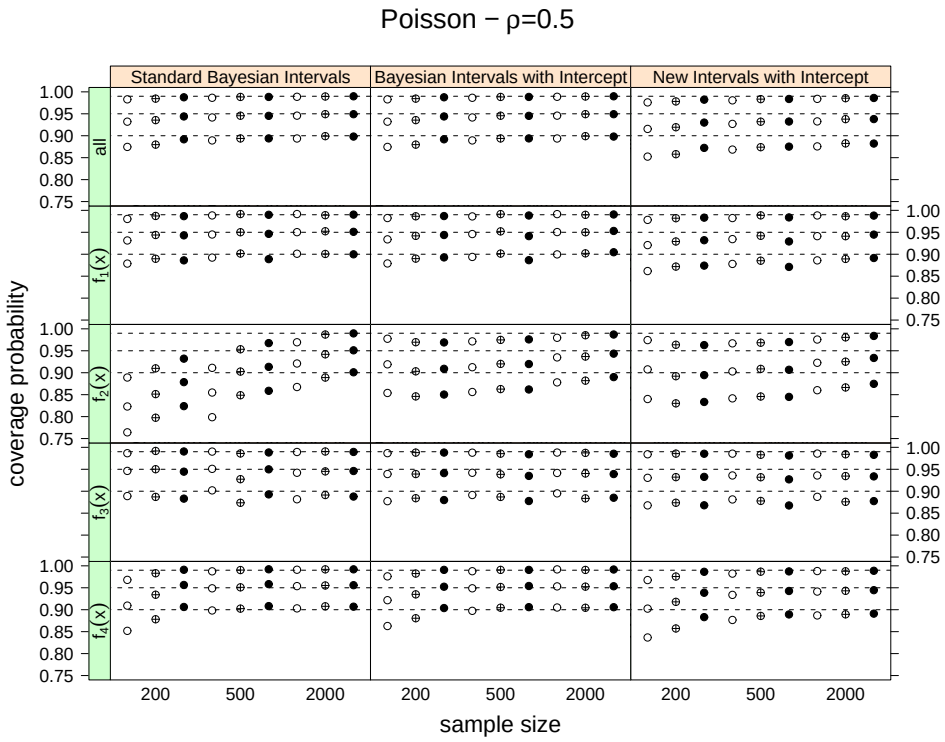

Figure 6: Coverage probability results for Poisson data for the case in which ρ was set to 0.5. Details are given in the caption of Figure 3.

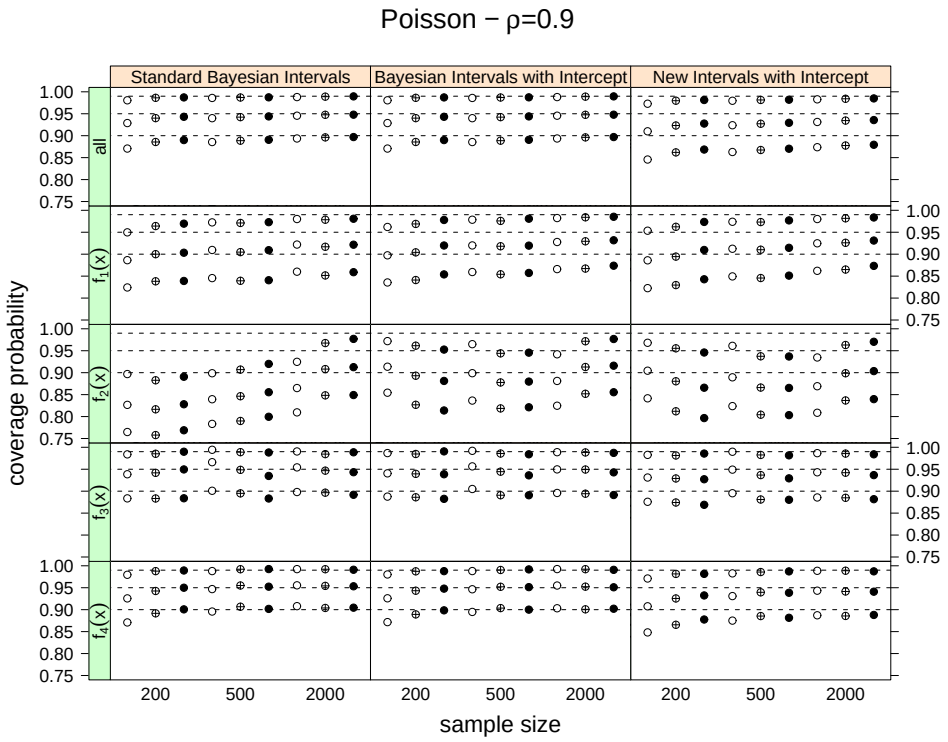

Figure 7: Coverage probability results for Poisson data for the case in which ρ was set to 0.9. Details are given in the caption of Figure 3. Notice how the confidence interval performance for $f_1(x)$ and $f_2(x)$ degrades when oversmoothing, due to high covariate correlation, occurs.

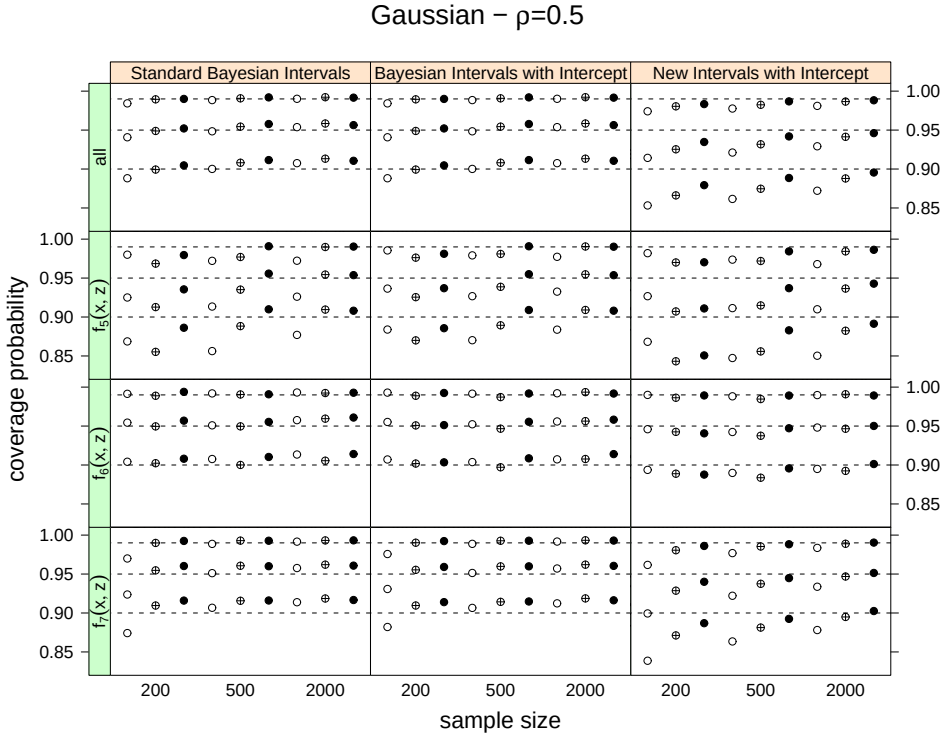

Figure 8: Coverage probability results for Gaussian data generated using $\eta_{2,i}$ as a linear predictor. Covariate correlation was equal to 0.5. Details are given in the caption of Figure 3. Notice the improvement in the performance of the intervals for $f_5(x, z)$, when the intercept is included in the calculations.

when $\rho = 0.9$. The interval performance for $f_3(x)$ and $f_4(x)$ does not change significantly. Concerning $f_3(x)$, straight line estimates and wiggly ones are selected, but always centered on the right level. For this reason, interval performance does not degrade. Regarding $f_4(x)$, the smoothing parameter can still be reasonably estimated as a result of the fact that the degree of complexity of this function is not low and hence heavy oversmoothing is not likely to occur, even when $\rho = 0.9$. Obviously, the Bayesian intervals 'with intercept' exhibit better coverage probabilities. Notice, once again, that as the information contained in the data increases interval performance improves.

The following argument explains the counter-intuitive downward coverage probability pattern that can be observed for a near-straight-line function like $f_2(x)$ (see Figure 7 for example). Let us consider a situation in which a straight line estimate is always selected for $f_2(x)$. The confidence intervals for the case in which $n = 200$ and the signal-to-noise ratio is low will be wider than the ones obtained for the case in which $n = 200$ but the signal-to-noise ratio is high. This means that, for some functions like the one being analyzed here, oversmoothing can become unimportant since the intervals will be so wide that they will do a better job than the ones calculated when more information is contained in the data. In other words, when the complexity of a function is not high and provided data are noisy enough to get very wide confidence intervals, the violation of the assumption that B is less than V can become irrelevant. However, as the signal-to-noise ratio and the sample size increase the intervals become narrower and the conclusions drawn in this results section apply.

5 Example: Diabetes in Pima women

In this section, we show the results obtained by applying the confidence intervals discussed in the paper to a dataset on diabetes in Pima Indian women.

The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, and are provided with the `MASS` library in `R`. Women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, were tested for diabetes according to World Health Organization criteria. The aim of the analysis was to investigate the relationship between presence of `diabetes` (defined as a binary variable) and different types of risk factors. The covariates considered were `age` (in years), `pedigree` (diabetes pedigree function), `mass` (body mass index, weight in kg/(height in m)^2), `triceps` (triceps skin fold thickness, mm), `pressure` (diastolic blood pressure, mm Hg), `glucose` (plasma glucose concentration in an oral glucose tolerance test), `pregnant` (number of pregnancies). We used the 532 complete records.

Figure 9 shows the smooth function estimates, together with the three kinds of intervals considered in this paper, obtained when fitting a GAM on the Pima Indian dataset. A Bernoulli distribution with a logistic link function between the linear predictor and the mean was employed. Smooth functions were represented using penalized thin regression splines based on second-order derivatives and with basis dimensions equal to 10. The smoothing parameters were chosen by generalized AIC. Before interpreting the empirical findings, it is important to stress that these confidence intervals have close to nominal coverage probabilities when averaged across the observation points, but not necessarily point-wise. This means that care must be taken not to over-interpret point-wise. Moreover, as

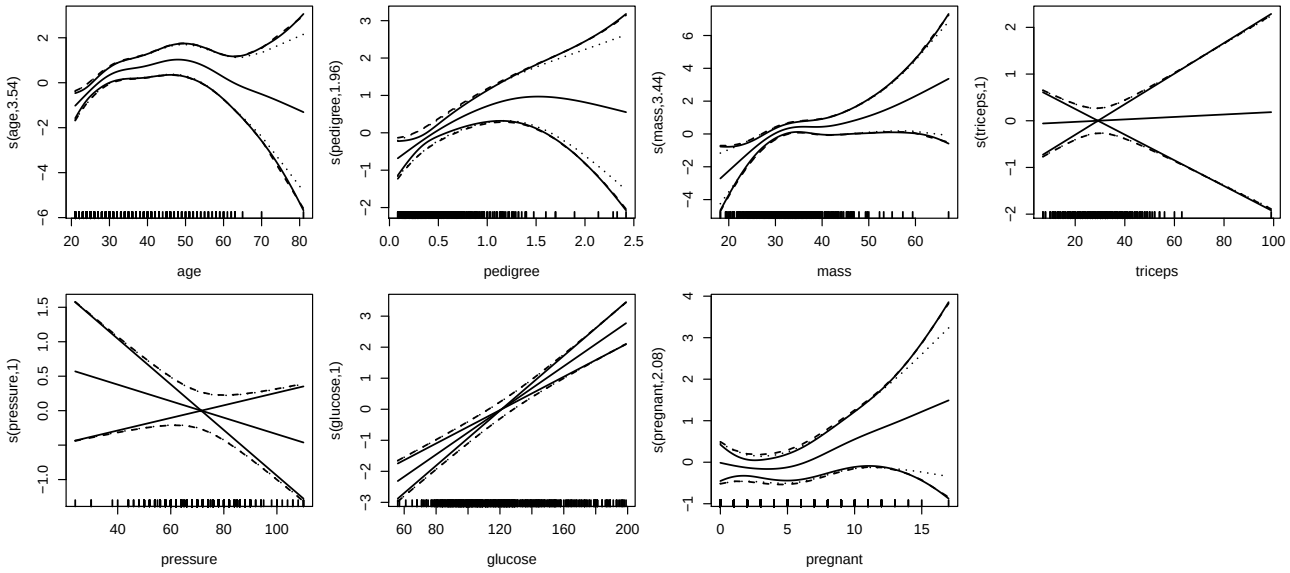

Figure 9: Smooth function estimates (solid middle lines) obtained fitting a GAM on the Pima Indian dataset, with corresponding 95% standard Bayesian intervals (solid lines), Bayesian intervals with intercept (dashed lines), and fully frequentist intervals with intercept (dotted lines). The results are reported on the scale of the linear predictor. The numbers in brackets in the y-axis captions are the estimated degrees of freedom of the smooth curves. The 'rug plot', at the bottom of each graph, shows the covariate values. Note that in some cases the Bayesian and frequentist intervals with intercept differ only minimally, hence they can hardly be distinguished. Further details are given in Section 5.

pointed out, e.g., by Hastie and Tibshirani (1990, p. 62), a confidence band is limited in the amount of information it provides, in that the functions in the confidence set might exhibit features that are not necessarily evident from the confidence band.

The intervals in Figure 9 all suggest no significant role for `triceps` and `pressure`, with no real difference in interpretation between the alternative intervals (as we would expect for effects that are really in the penalty null space). Inference about the strong, non-linear effects of `age` and `mass` are essentially unchanged between the different intervals (as is expected for effects that are not 'close to' the penalty null space). Interpretation of the remaining three covariates highlights the practical difference in using the 'with intercept' intervals: for all three the interpretation changes somewhat between the constrained intervals and the others, although the effects are not dramatic. For `glucose` the 'with intercept' intervals admit the possibility of a mildly non-linear effect, as opposed to the strictly linear effect of the constrained interval. For `pregnant` we see some weakening of the evidence for any effect at all, in moving to the with intercept intervals. Similarly the evidence for a strong initial positive effect of `pedigree`, is weakened, although there is still good evidence that the effect is real. This example illustrates what we would expect to see in general: by using better calibrated intervals, we are unlikely to see dramatic changes in overall interpretation, but will be somewhat less likely to over-interpret.

6 Discussion

We have shown by simulation and extension of Nychka's (1988) analysis, that the Wahba/Silverman type Bayesian intervals for the components of a penalized regression spline based GAM have gen-

erally close to nominal frequentist properties, across-the-function. As simulation evidence and our theoretical arguments suggest, the exception occurs when components estimated subject to identifiability constraints have interval widths vanishing somewhere as a result of heavy smoothing. Coverage probabilities can be improved if intervals are only obtained for unconstrained quantities, such as a smooth component plus the model intercept. The theoretical results also allow us to define alternative intervals from a purely frequentist approach, and these appear to perform almost as well as the Bayesian intervals.

The results make a novel contribution in extending Nychka's argument to the GAM component case, thereby pinpointing the circumstances in which the intervals will and will not work, and explaining the role of smoothness selection as well as smoothing parameter uncertainty.

The findings are backed up with quite extensive simulation testing of the finite sample performance of the three types of confidence intervals considered here. Specifically, our Monte Carlo investigation allowed us to compare the component-wise intervals under a wide variety of setting. In this respect, our simulation results show under which circumstances these intervals can reliably represent smooth term uncertainty, given the smoothing parameter selection methods employed here, and may provide applied researchers with some guidance about the practical performance of the intervals.

An obvious open question is whether the theory can be made more rigorous. Our theoretical arguments clearly explain why the Bayesian intervals have close to nominal coverage properties, but do so without actually proving that the intervals will always do so. As detailed in section 3, our argument relies on some plausible assumptions, but substantial further work would be needed to establish the exact conditions under which these will hold.

Finally note that several other approaches have also been proposed to produce inferential tools for spline type smoothers and GAMs. Hastie and Tibshirani (1990) suggested using simple frequentist approximations to produce approximate confidence intervals. Other frameworks include the use of bootstrap methods as well as the fully Bayesian approach. For example, Härdle and Bowman (1988) and Härdle and Marron (1991) used bootstrap to construct pointwise and simultaneous confidence intervals. An under-smoothing approach has also been taken within the bootstrap framework, as a devise for avoiding smoothing bias (Hall, 1992; Kaurmann and Opsomer, 2003). Direct bootstrapping has been employed as well, as in Härdle et al. (2004) who make use of 'Wild' bootstrap methods. As an alternative, Fahrmeir et al. (2004) and Fahrmeir and Lang (2001) adopted the fully Bayesian approach, which employs MCMC for practical computations. Wang and Wahba (1995) compared the Wahba/Silverman type confidence intervals with those derived using several variations of the bootstrap approach. They found that the bootstrap framework can yield intervals that are comparable to the Bayesian ones in terms of across-the-function coverage properties. However, they are computationally intensive, a problem which may affect the fully Bayesian approach as well.

Acknowledgements

We wish to thank one anonymous referee for the well thought suggestions and constructive criticism which helped to improve the presentation of the paper considerably.

Appendix: The relative magnitude of B and V

Our analysis suggests that the intervals will only work well if B is of substantially smaller magnitude than V. Nychka (1988) provided simulation evidence that this would usually be the case for univariate spline smoothing. This appendix provides some simulation evidence that this also holds in the component wise case, as well as providing a limited theoretical exploration of the issue.

Empirical insight into the relative magnitude of B and V can be gained by examining the percentage mean squared bias and mean variance of the smooth components of a GAM. Table 3 reports these two quantities, calculated according to the definitions in Section 3.2 with $C_i^{-1} = [\mathbf{V}_{\mathbf{f}_j}]_{ii}$ for each smooth component j, for some of the cases considered in our simulation study. Overall, the mean squared bias is of substantially smaller magnitude than the mean variance, except for f_2 where the opposite happens. As explained in Section 3.4, when a true function is close to a term in the null space of the component's penalty, and the corresponding smooth function is estimated as a straight line but subject to an identifiability constraint, the assumption that B is less than V will fail. These results are consistent with our simulations, where poor coverage probabilities were obtained for f_2. Notice that this is not the result of failing to meet the assumption that the mean squared bias of the parameters of a smooth is less than its mean variance. The fact that B is greater than V is due to an 'artifact' induced by the identifiability constraint, an issue which can be explored only if using the extended Nychka argument of this paper, where some constants C_i have to be used to derive non-constant width intervals for GAM components. Recall that as a remedy, improved coverages are obtained if each term's interval is computed as if it alone were unconstrained, and identifiability was obtained by constraints on the other model terms.

	binomial		*gamma*		*Gaussian*		*Poisson*	
function	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$
f_1	8.6	91.4	12.8	87.2	9.0	91.0	17.0	83.0
f_2	76.5	23.5	90.4	9.6	62.5	37.5	94.9	5.1
f_3	0.9	99.1	0.1	99.9	0.1	99.9	0.3	99.7
f_4	15.4	84.6	26.5	73.5	13.4	86.6	17.5	82.5

Table 3: Percentage mean squared bias ($\bar{b^2}^*$) and mean variance ($\bar{v^2}^*$) results for the smooth components of GAMs fitted to data simulated from four error models at medium noise level. Covariate correlation and sample size were 0.5 and 200 (see Section 4 for further details). $\bar{b^2}^* = \bar{b^2}/(\bar{b^2} + \bar{v^2}) * 100$ and $\bar{v^2}^* = \bar{b^2}/(\bar{b^2} + \bar{v^2}) * 100$, where $\bar{b^2}$ and $\bar{v^2}$ were calculated following the definitions in Section 3.2, with $C_i^{-1} = [\mathbf{V}_{\mathbf{f}_j}]_{ii}$ for each smooth component j. Notice that the $B < V$ assumption is comfortably met for all terms except for f_2, which is the problematic case in the first columns of Figures 3 - 7.

Some limited theoretical insight into the relative magnitude of B and V can be gained by examining the mean squared bias and mean variance of the parameters of a smooth, in the Demmler-Reinsch (or 'natural') parameterization as in Section 3.1. For simplicity, and without loss of generality, let us consider the case of a smooth estimated by a penalized least squares fit to data with variance σ^2. In the new parameterization, provided the model is a reasonable fit, it is easy to find expressions for the

mean variance and mean squared bias of the coefficients,

$$\bar{v^2} = \frac{1}{p}\sum_k \frac{\sigma^2}{(1+D_{kk})^2}$$

and

$$\bar{b^2} = \frac{1}{p}\sum_k \frac{\sigma^2 D_{kk}}{(1+D_{kk})^2}.$$

If M is the dimension of the null space of the smooth penalty then M of the D_{kk} will be zero. These unpenalized coefficients contribute $M\sigma^2/p$ to the mean variance, and nothing at all to the mean squared bias. So for $\bar{b^2}$ to exceed $\bar{v^2}$ we require the remaining terms in the $\bar{b^2}$ to tend to exceed the corresponding terms in $\bar{v^2}$ and to be substantial relative to $M\sigma^2/p$. This is difficult to achieve. The largest term in $\bar{b^2}$ is bounded above by $\sigma^2/(4p)$, and is of the same size as the corresponding term in $\bar{v^2}$. Later terms in $\bar{b^2}$ do become larger than the corresponding terms in $\bar{v^2}$, but at the same time they rapidly become very small relative to $M\sigma^2/p$. Given that different smoothers will have different eigen spectra, it is difficult to make this argument more precise, but it does go some way to explaining the simulation results, and also makes the interesting prediction that the B less than V assumption will tend to hold more strongly as the dimension of the penalty null space increases.

References

[1] 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, 267–281.

[2] Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik*, 31, 377–403.

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

[4] Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models based on Markov Random Field Priors," *Journal of the Royal Statistical Society Series C*, 50, 201–220.

[5] Gentle, J. E. (2003), *Random Number Generation and Monte Carlo Methods*, London: Springer-Verlag.

[6] Gu, C. (1992), "Penalized Likelihood Regression - A Bayesian Analysis," *Statistica Sinica*, 2, 255–264.

[7] Gu, C. (2002), *Smoothing Spline ANOVA Models*, London: Springer-Verlag.

[8] Gu, C., and Wahba, G. (1993), "Smoothing Spline ANOVA with Component-Wise Bayesian Confidence Intervals," *Journal of Computational and Graphical Statistics*, 2, 97-117.

[9] Hall, P. (1992), "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *Annals of Statistics*, 20, 675–694.

[10] Härdle, W., and Bowman, A. W. (1988), "Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands," *Journal of the American Statistical Association*, 83, 102–110.

[11] Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004), "Bootstrap Inference in Semiparametric Generalized Additive Models," *Econometric Theory*, 20, 265–300.

[12] Härdle W., and Marron, J. S. (1991), "Bootstrap Simultaneous Error Bands for Nonparametric Regression," *The Annals of Statistics*, 19, 778–796.

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

[14] Kauermann, G., and Opsomer, J. D. (2003), "Local Likelihood Estimation in Generalized Additive Models," *Scandinavian Journal of Statistics*, 30, 317–337.

[15] McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall.

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

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

[18] 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, 1–52.

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

[20] Wang, Y., and Wahba, G. (1995), "Bootstrap Confidence Intervals for Smoothing Splines and Their Comparison to Bayesian Confidence Intervals". *Journal of Statistical Computation and Simulation*, 51, 263–279.

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

[22] Wood, S. N. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society Series B*, 65, 95–114.

[23] Wood, S. N. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–86.

[24] Wood, S. N. (2006a), *Generalized Additive Models: An Introduction with R*, London: Chapman & Hall.

[25] Wood, S. N. (2006b), "On Confidence Intervals for Generalized Additive Models based on Penalized Regression Splines," *Australian & New Zealand Journal of Statistics*, 48, 445–64.

[26] Wood, S. N. (2008), "Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models," *Journal of the Royal Statistical Society Series B*, 70, 495–518.