Year: 2015

Objective bayesian model selection in generalized additive models with penalized splines

Sabanés Bové, Daniel; Held, Leonhard; Kauermann, Göran

DOI: https://doi.org/10.1080/10618600.2014.912136

Originally published at:
Sabanés Bové, Daniel; Held, Leonhard; Kauermann, Göran (2015). Objective bayesian model selection in generalized additive models with penalized splines. Journal of Computational and Graphical Statistics, 24(2):394-415.
DOI: https://doi.org/10.1080/10618600.2014.912136

Objective Bayesian Model Selection in Generalised Additive Models with Penalised Splines

Daniel Sabanés Bové[*] Leonhard Held[*] Göran Kauermann[†]

Abstract

We propose an objective Bayesian approach to the selection of covariates and their penalised splines transformations in generalised additive models. The methodology is based on a combination of continuous mixtures of g-priors for model parameters and a multiplicity-correction prior for the models themselves. We introduce our approach in the normal model and extend it to non-normal exponential families. A simulation study and an application with binary outcome is provided. An efficient implementation is available in the `R`-package "`hypergsplines`".

Keywords: variable selection, function selection, g-prior, shrinkage, stochastic search

1 Introduction

Semiparametric regression has achieved an impressive dissemination over the last years. Its central idea is to replace parametric regression functions by smooth, semiparametric components. Following Hastie and Tibshirani (1990), suppose we have p continuous

[*] Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland. E-mail: {daniel.sabanesbove,leonhard.held}@ifspm.uzh.ch

[†] Department of Statistics, Ludwig-Maximilians-Universität München, Germany. E-mail: goeran.kauermann@stat.uni-muenchen.de

covariates $x_1, \dots, x_p$ and use the additive model

$$y = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \epsilon, \tag{1}$$

where the $m_j(\cdot)$, $1 \leq j \leq p$, are smooth but otherwise unspecified functions and $\epsilon \sim \mathrm{N}(0, \sigma^2)$. For identifiability purposes we further assume that $\mathbb{E}\{m_j(X_j)\} = 0$ with respect to the marginal distribution of each covariate X_j. Estimation of the smooth terms in (1) can be carried out in different ways, where we here make use of penalised splines, see *e.g.* Eilers and Marx (2010) or Wood (2006). A general introduction to penalised spline smoothing has been provided by Ruppert, Wand, and Carroll (2003) and the approach has become a popular smoothing technique since then. The general idea is to decompose the functions m_j into a linear and a nonlinear part, where the latter is represented through a spline basis, that is

$$m_j(x_j) = x_j \beta_j + \boldsymbol{Z}_j(x_j)^T \boldsymbol{u}_j. \tag{2}$$

Here $\boldsymbol{Z}_j(x_j)$ is a $K \times 1$ spline basis vector at position x_j and $\boldsymbol{u}_j$ is the corresponding coefficient vector, for $1 \leq j \leq p$. Conveniently one may choose a truncated polynomial basis for $\boldsymbol{Z}_j(\cdot)$ but representation (2) holds in general as well, see Wand and Ormerod (2008). To achieve a smooth fit one imposes a quadratic penalty on the spline coefficient vector $\boldsymbol{u}_j$. Equivalently, one may formulate the penalty as a normal prior

$$\boldsymbol{u}_j \,|\, \sigma^2, \rho_j \sim \mathrm{N}_K(\boldsymbol{0}_K, \sigma^2 \rho_j \boldsymbol{I}_K), \tag{3}$$

where $\boldsymbol{0}_K$ is the all-zeros vector and $\boldsymbol{I}_K$ is the identity matrix of dimension K, which leads together with (1) and (2) to a linear mixed model (see Wand, 2003; Kauermann, Krivobokova, and Fahrmeir, 2009). The variance factor ρ_j plays the role of a smoothing parameter which steers the amount of penalisation (relative to the regression variance σ^2). A larger ρ_j leads to a higher prior variance of the spline coefficients and hence a more wiggly function m_j, while a smaller ρ_j leads to a stronger penalty on $\|\boldsymbol{u}_j\|$ and thus a smoother function m_j. In the extreme case, setting ρ_j to zero imposes $\boldsymbol{u}_j \equiv \boldsymbol{0}_K$ so that $m_j(x_j)$ collapses to a linear term $m_j(x_j) = x_j\beta_j$. Hence the role of ρ_j ($j = 1, \dots, p$) can be seen twofold. For $\rho_j > 0$ it plays the role of a smoothing parameter but with

$\rho_j = 0$ it extends to model selection of (generalised) additive models by separating linear from non-linear effects. We will extend the idea in this paper coherently by proposing a general model selection including variable selection, that is by allowing the alternative $m_j(x_j) \equiv 0$. The central idea is that ρ_j determines uniquely the contribution of the function $m_j(x_j)$ to the overall degrees of freedom of the model (see Ruppert et al., 2003), which is a measure of the complexity of the model. So instead of estimating or drawing inference about ρ_j we draw inference about the corresponding degrees of freedom.

The selection of variables and covariates, respectively, is a central question in statistics. This applies in particular to regression models where the intention is to reduce the variance of effect estimates due to uninformative covariates. The field is wide and many different approaches have been proposed in the last years including the following. Friedman (2001) and Tutz and Binder (2006) describe boosting algorithms, which are extended by Kneib, Hothorn, and Tutz (2009) to geoadditive regression models (Fahrmeir, Kneib, and Lang, 2004). For the same model class, Belitz and Lang (2008) propose to use information-criteria or cross-validation, while Fahrmeir, Kneib, and Konrath (2010) and Scheipl, Fahrmeir, and Kneib (2012) use spike-and-slab priors for variable and function selection (see also Scheipl, Kneib, and Fahrmeir (2013) for simulation studies comparing their approach to the one presented in this paper). Brezger and Lang (2008) adopt the concept of Bayesian contour probabilities (Held, 2004) to decide on the inclusion and form of covariate effects. Cottet, Kohn, and Nott (2008) generalise earlier work by Yau, Kohn, and Wood (2003) to Bayesian double-exponential regression models, which comprise generalised additive models as a special case. Shrinkage approaches are proposed by Wood (2011) and Marra and Wood (2011). Zhang and Lin (2006) use a lasso-type penalised likelihood approach, and Ravikumar, Liu, Lafferty, and Wasserman (2008) and Meier, van de Geer, and Bühlmann (2009) use penalties favouring both sparsity and smoothness of high-dimensional models. Likelihood-ratio testing methods are described by Kauermann and Tutz (2001) and Cantoni and Hastie (2002). This list mirrors the multitude as well as the variety of the different approaches and is, of course, in no way exhaustive.

In this paper we propose a novel objective Bayesian variable and function selection approach based on continuous mixtures of (generalised) g-priors. This type of prior for the parameters in the generalised additive model traces back to the g-prior in the linear model (Zellner, 1986). Its hyper-parameter g acts as an inverse relative prior sample size, and assigning it a hyper-prior solves the information paradox (Liang, Paulo, Molina, Clyde, and Berger, 2008, section 4.1) of the fixed-g prior (Berger and Pericchi, 2001, p. 148) in the linear model. One specific example are the hyper-g priors of Liang et al. (2008, section 3.2), which enjoy a closed form for the marginal likelihood and lead to consistent model selection and model-averaged prediction. We will proceed to use hyper-g priors, because they have been well studied and have shown good frequentist properties in the Gaussian linear model. They have recently been extended to generalised linear models by Sabanés Bové and Held (2011b). We follow the conventional prior approach (Berger and Pericchi, 2001, section 2.1) by using non-informative improper priors for parameters which are common to all models, and default proper hyper-g priors for model-specific parameters.

While hyper-g priors have been discussed extensively in the Bayesian variable selection literature, *e. g.* by Cui and George (2008), Liang et al. (2008), Bayarri, Berger, Forte, and García-Donato (2012) and Celeux, Anbari, Marin, and Robert (2012), this is the first paper to our knowledge that applies hyper-g priors to generalised additive models. The general idea of applying hyper-g priors, originally developed for linear models, to generalised additive models is new. The rationale is that default priors have carefully and exhaustively been constructed for the linear model, so their advantages should be used when drawing inferences about generalised additive models. Moreover, we consider both variable selection and transformation in a coherent Bayesian framework.

The paper is organised as follows. We first describe how to approach additive models in Section 2, including the specification of hyper-g priors in this model class (Section 2.1), and a suitable multiplicity-correction prior as well as a stochastic search procedure on the model space (Section 2.2). We illustrate the performance of the methodology with a simulation study (Section 2.3). We then extend our focus to generalised additive models in Section 3, which is complemented by an application to real data (Section 3.2). Section 4 closes the paper with a discussion.

2 Additive Models

Assume we have observed independent responses y_i at covariate values $x_{i1}, \dots, x_{ip}$, $i = 1, \dots, n$, from the additive normal model (1). For each covariate $j = 1, \dots, p$, we stack the covariate values into the $n \times 1$ vector $\tilde{\boldsymbol{x}}_j = (x_{1j}, \dots, x_{nj})^T$ and the spline basis vectors into the $n \times K$ matrix $\tilde{\boldsymbol{Z}}_j = (\boldsymbol{Z}_j(x_{1j}), \dots, \boldsymbol{Z}_j(x_{nj}))^T$. To achieve orthogonality we apply the Gram-Schmidt process (see Björck, 1967)

$$\boldsymbol{x}_j = \tilde{\boldsymbol{x}}_j - \boldsymbol{1}_n \frac{\boldsymbol{1}_n^T \tilde{\boldsymbol{x}}_j}{\boldsymbol{1}_n^T \boldsymbol{1}_n} = \tilde{\boldsymbol{x}}_j - \boldsymbol{1}_n \bar{x}_j, \tag{4}$$

$$\boldsymbol{Z}_j = \tilde{\boldsymbol{Z}}_j - \boldsymbol{1}_n \frac{\boldsymbol{1}_n^T \tilde{\boldsymbol{Z}}_j}{\boldsymbol{1}_n^T \boldsymbol{1}_n} - \boldsymbol{x}_j \frac{\boldsymbol{x}_j^T \tilde{\boldsymbol{Z}}_j}{\boldsymbol{x}_j^T \boldsymbol{x}_j}, \tag{5}$$

where $\boldsymbol{1}_n$ denotes the all-ones vector of dimension n. This ensures that $\boldsymbol{1}_n$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ are orthogonal to each other, *i. e.* $\boldsymbol{1}_n^T \boldsymbol{x}_j = 0$ and $\boldsymbol{1}_n^T \boldsymbol{Z}_j = \boldsymbol{x}_j^T \boldsymbol{Z}_j = \boldsymbol{0}_K$. The orthogonalisation procedure ensures that we can separate the linear and nonlinear part of m_j, which is a prerequisite for the definition of the degrees of freedom measure below. Note that covariates may still be mutually correlated.

A common measure of model complexity is the degrees of freedom of a model. While in parametric models this is just the number of parameters, for smoothing and mixed models Aerts, Claeskens, and Wand (2002, section 2.2) relate the smoothing parameter ρ_j to the corresponding degrees of freedom through

$$d_j(\rho_j) = \operatorname{tr}\{(\boldsymbol{Z}_j^T \boldsymbol{Z}_j + \rho_j^{-1} \boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \boldsymbol{Z}_j\} + 1 \in (1, K+1) \tag{6}$$

for a smoothly modelled covariate effect m_j. Note that $d_j(\rho_j) = \sum_{k=1}^K \lambda_{jk} / (\lambda_{jk} + \rho_j^{-1})$ is easy to calculate via the (positive) eigenvalues λ_{jk} of $\boldsymbol{Z}_j^T \boldsymbol{Z}_j$. This also shows that $d_j(\rho_j)$ is strictly increasing in ρ_j with derivative $\sum_{k=1}^K \lambda_{jk} / (\rho_j \lambda_{jk} + 1)^2 > 0$. This in turn implies that we may (numerically) invert the function to $\rho_j(d_j)$, which means that we have a one-to-one relation between ρ_j and the degrees of freedom d_j. Note that (6) is an asymptotic approximation of the more commonly used definition of degrees of freedom for linear smoothers (see Aerts et al., 2002) and may thus lead to an imprecise measure of model complexity in small samples.

Subsequently we will restrict the degrees of freedom to take values in a finite set $\mathcal{D} \subset \{0\} \cup [1, K+1)$. In the remainder of this article we will use $\mathcal{D} = \{0, 1, 2, 3, \ldots, K\}$, which determines the size of $\mathcal{D}$ to be $K+1$. In general you may want to pick the grid of degrees of freedom to be finer or with the maximum degrees of freedom less than K (perhaps to be chosen by the user), which might be advantageous in some cases. For $d_j = 0$ we set $m_j(x_j) \equiv 0$ while for $d_j = 1$ we have the linear model $m_j(x_j) = x_j \beta_j$. In general, we translate the structure of model (1) into the index vector $\boldsymbol{d} = (d_1, \ldots, d_p)$ giving the degrees of freedom for each functional component. The objective of the paper is to draw inference about $\boldsymbol{d}$, which we subsequently refer to as the "model". To do so, we look now at the stochastic model for the response based on a specific model $\boldsymbol{d}$.

After combining the $I = \sum_{j=1}^{p} \mathbb{I}(d_j \geq 1)$ vectors $\boldsymbol{x}_j$ to the $n \times I$ linear design matrix $\boldsymbol{X}_{\boldsymbol{d}} = (\boldsymbol{x}_j : d_j \geq 1)$ and the $J = \sum_{j=1}^{p} \mathbb{I}(d_j > 1)$ matrices $\boldsymbol{Z}_j$ to the $n \times JK$ spline design matrix $\boldsymbol{Z}_{\boldsymbol{d}} = (\boldsymbol{Z}_j : d_j > 1)$, and analogously constructing the respective coefficient vectors $\boldsymbol{\beta}_{\boldsymbol{d}}$ and $\boldsymbol{u}_{\boldsymbol{d}}$, the conditional additive model for the response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}}, \sigma^2 \sim \mathrm{N}_n\left(\boldsymbol{1}_n \beta_0 + \boldsymbol{X}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}} + \boldsymbol{Z}_{\boldsymbol{d}} \boldsymbol{u}_{\boldsymbol{d}}, \, \sigma^2 \boldsymbol{I}_n\right). \tag{7}$$

Integrating out the the spline coefficient vector $\boldsymbol{u}_{\boldsymbol{d}} \,|\, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_{JK}(\boldsymbol{0}_{JK}, \sigma^2 \boldsymbol{D}_{\boldsymbol{d}})$, where $\boldsymbol{\rho}_{\boldsymbol{d}} = (\rho_j : d_j > 1)$ and $\boldsymbol{D}_{\boldsymbol{d}}$ is block-diagonal with J blocks $\rho_j \boldsymbol{I}_K$ $(d_j > 1)$, yields the so-called marginal model

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_n\left(\boldsymbol{1}_n \beta_0 + \boldsymbol{X}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}}, \, \sigma^2 \boldsymbol{V}_{\boldsymbol{d}}\right) \tag{8}$$

with $\boldsymbol{V}_{\boldsymbol{d}} = \boldsymbol{I}_n + \boldsymbol{Z}_{\boldsymbol{d}} \boldsymbol{D}_{\boldsymbol{d}} \boldsymbol{Z}_{\boldsymbol{d}}^T$. To illustrate the notation, consider for example $p = 4$ covariates and $K = 3$ knots, and a model with degrees of freedom $d_1 = 0$, $d_2 = 1$ and $d_3, d_4 = 2$. Then $\boldsymbol{X}_{\boldsymbol{d}} = (\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ has $I = 3$ columns, $\boldsymbol{Z}_{\boldsymbol{d}} = (\boldsymbol{Z}_3, \boldsymbol{Z}_4)$ has is composed of $J = 2$ matrices and has $JK = 6$ columns, and $\boldsymbol{D}_{\boldsymbol{d}} = \mathrm{diag}(\rho_3 \boldsymbol{I}_3, \rho_4 \boldsymbol{I}_3)$. This general linear model can be decorrelated into a standard linear model by using the Cholesky decomposition $\boldsymbol{V}_{\boldsymbol{d}} = \boldsymbol{V}_{\boldsymbol{d}}^{T/2} \boldsymbol{V}_{\boldsymbol{d}}^{1/2}$: For the transformed response vector $\tilde{\boldsymbol{y}} = \boldsymbol{V}_{\boldsymbol{d}}^{-T/2} \boldsymbol{y}$ we have

$$\tilde{\boldsymbol{y}} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_n\left(\tilde{\boldsymbol{1}}_n \beta_0 + \tilde{\boldsymbol{X}}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}}, \, \sigma^2 \boldsymbol{I}_n\right) \tag{9}$$

with analogously transformed all-ones vector $\tilde{\mathbf{1}}_n = \boldsymbol{V}_{\boldsymbol{d}}^{-T/2}\mathbf{1}_n$ and design matrix $\tilde{\boldsymbol{X}}_{\boldsymbol{d}} = \boldsymbol{V}_{\boldsymbol{d}}^{-T/2}\mathbf{X}_{d}$. Note that now also $\tilde{\boldsymbol{y}}$ and $\tilde{\mathbf{1}}_n$ depend on the model $\boldsymbol{d}$, but we suppress this dependence for ease of notation.

2.1 Hyper-g Priors for Additive Models

We will now impose priors on the parameters and show how to use hyper-g priors for the parameter components β_0, $\boldsymbol{\beta}_{\boldsymbol{d}}$ and σ^2 in the decorrelated marginal model (9). The hyper-g priors comprise a locally uniform prior $f(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $f(\sigma^2) \propto (\sigma^2)^{-1}$ on the regression variance and the g-prior (Zellner, 1986)

$$\boldsymbol{\beta}_{\boldsymbol{d}} \,|\, g, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_I\left(\mathbf{0}_I,\, g\sigma^2(\tilde{\mathbf{X}}_{\boldsymbol{d}}^T\tilde{\mathbf{X}}_{\boldsymbol{d}})^{-1}\right) \tag{10}$$

on the linear coefficient vector. Note that the prior precision matrix in (10) is proportional to $\sigma^{-2}\tilde{\mathbf{X}}_{\boldsymbol{d}}^T\tilde{\mathbf{X}}_{\boldsymbol{d}} = \sigma^{-2}\mathbf{X}_{\boldsymbol{d}}^T\boldsymbol{V}_{\boldsymbol{d}}^{-1}\mathbf{X}_{\boldsymbol{d}}$, which is the Fisher information matrix of $\boldsymbol{\beta}_{\boldsymbol{d}}$ in model (8). The prior construction is completed with either a uniform hyper-prior on the shrinkage coefficient $g/(1+g)$,

$$\frac{g}{1+g} \sim \mathrm{U}(0,1), \tag{11}$$

leading to the hyper-g prior, or with

$$\frac{g/n}{1+g/n} \sim \mathrm{U}(0,1), \tag{12}$$

leading to the hyper-g/n prior (Liang et al., 2008). We recommend to use the latter, because it also leads to consistent posterior model probabilities if the true model is the null model (Liang et al., 2008, theorem 4), see Table 1 in Section 2.3 for illustration.

Basically all formulae given by Liang et al. (2008) carry over to our setting, since inner products of the response vector $\boldsymbol{y}$, the all-ones vector $\mathbf{1}_n$ and the design matrix $\boldsymbol{X}_{\boldsymbol{d}}$ in model (8) carry over to their transformed counterparts $\tilde{\boldsymbol{y}}$, $\tilde{\mathbf{1}}_n$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{d}}$ in model (9). This is due to

$$\boldsymbol{V}_{\boldsymbol{d}}^{-1} = (\boldsymbol{I}_n + \boldsymbol{Z}_d\boldsymbol{D}_d\boldsymbol{Z}_d^T)^{-1} = \boldsymbol{I}_n - \boldsymbol{Z}_d(\boldsymbol{Z}_{\boldsymbol{d}}^T\boldsymbol{Z}_{\boldsymbol{d}} + \boldsymbol{D}_{\boldsymbol{d}}^{-1})^{-1}\boldsymbol{Z}_{\boldsymbol{d}}^T, \tag{13}$$

which follows from the matrix inversion lemma (see Henderson and Searle, 1981) and leads to $\tilde{\mathbf{1}}_n^T \tilde{\mathbf{1}}_n = \mathbf{1}_n^T \mathbf{1}_n = n$, $\tilde{\mathbf{1}}_n^T \tilde{\boldsymbol{X}}_d = \mathbf{1}_n^T \boldsymbol{X}_d = \mathbf{0}_l$ and $\tilde{\mathbf{1}}_n^T \tilde{\boldsymbol{y}} = \mathbf{1}_n^T \boldsymbol{y}$ by straightforward calculations. A most convenient property of the hyper-g priors is that they yield closed form marginal likelihoods, which need to be computed on the original response scale via the change of variables formula:

$$f(\boldsymbol{y} \,|\, \boldsymbol{d}) \propto f(\tilde{\boldsymbol{y}} \,|\, \boldsymbol{d}) |\boldsymbol{V}_d^{1/2}|^{-1}, \tag{14}$$

where $f(\tilde{\boldsymbol{y}} \,|\, \boldsymbol{d})$ is the marginal likelihood of the transformed response vector $\tilde{\boldsymbol{y}}$ in the standard linear model (9). The closed forms for $f(\tilde{\boldsymbol{y}} \,|\, \boldsymbol{d})$ under the hyper-g priors are given in Appendix A.

For completeness we note that other hyper-priors could be assigned to g as well, but they will typically not lead to a closed form of the marginal likelihood. Examples are the incomplete inverse-gamma prior on $1+g$ (Cui and George, 2008, p. 891), which generalises the above uniform prior on $g/(1+g)$, and an inverse-gamma prior on g, which corresponds to the Cauchy prior of Zellner and Siow (1980). The hyper-g/n prior is a special case of the robust prior proposed by Bayarri et al. (2012), for which a closed form of the marginal likelihood exists. An overview of hyper-g priors is given by Ley and Steel (2012).

Posterior inference in a given model $\boldsymbol{d}$ is based on Monte Carlo estimation of the parameters in model (7). We therefore use the factorisation

$$f(\beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d, \sigma^2, g \,|\, \boldsymbol{y}) = f(\boldsymbol{u}_d \,|\, \beta_0, \boldsymbol{\beta}_d, \sigma^2, \boldsymbol{y}) f(\beta_0, \boldsymbol{\beta}_d \,|\, \sigma^2, g, \boldsymbol{y}) f(\sigma^2 \,|\, \boldsymbol{y}) f(g \,|\, \boldsymbol{y}). \tag{15}$$

Sampling of g, σ^2 and subsequently $\beta_0, \boldsymbol{\beta}_d$ can be done along the lines of Sabanés Bové and Held (2011a, section 2.3): Based on the decorrelated model (9), we sample g using inverse sampling (either with a closed-form quantile function, if the hyper-g prior (11) is used, or with a numerical approximation of the quantile function, if the hyper-g/n prior (12) is used), σ^2 from an inverse-gamma distribution, and finally $\beta_0, \boldsymbol{\beta}_d \,|\, g, \sigma^2$ from a Gaussian distribution. Finally, the spline coefficient vector $\boldsymbol{u}_d$ is sampled from

$$\begin{aligned} f(\boldsymbol{u}_d \mid \beta_0, \boldsymbol{\beta}_d, \sigma^2, \boldsymbol{y}) &\propto f(\boldsymbol{u}_d \mid \sigma^2) f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d, \sigma^2) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left[\boldsymbol{u}_d^T \boldsymbol{D}_d^{-1} \boldsymbol{u}_d + \|\boldsymbol{y} - \boldsymbol{1}_n \beta_0 - \boldsymbol{X}_d \boldsymbol{\beta}_d - \boldsymbol{Z}_d \boldsymbol{u}_d\|^2 \right] \right\} \\ &\propto \mathrm{N}_{JK}\left(\boldsymbol{u}_d \mid \boldsymbol{\Sigma}_d \boldsymbol{Z}_d^T (\boldsymbol{y} - \boldsymbol{X}_d \boldsymbol{\beta}_d), \sigma^2 \boldsymbol{\Sigma}_d \right), \end{aligned} \tag{16}$$

where $\boldsymbol{\Sigma}_d = (\boldsymbol{Z}_d^T \boldsymbol{Z}_d + \boldsymbol{D}_d^{-1})^{-1}$ and β_0 disappears because $\boldsymbol{Z}_d^T \boldsymbol{1}_n = \boldsymbol{0}_{JK}$. A more detailed description of the parameter sampling approach can be found in the supplementary material.

The general intention though is to draw inference about $\boldsymbol{d}$, which is with the prerequisites introduced so far possible as proposed in the next section.

2.2 Model Prior and Stochastic Search

First we propose a prior $f(\boldsymbol{d})$ on the model space $\mathcal{D}^p$ which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of the p covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model (I) is uniformly distributed on $\{0, 1, \ldots, p\}$. The choice of the I covariates is then uniformly distributed on all possible configurations, and their degrees of freedom are independent and uniformly distributed on $\mathcal{D} \setminus \{0\} = \{1, 2, 3, \ldots, K\}$. Altogether, this gives

$$1/f(\boldsymbol{d}) = (p+1) \binom{p}{I} K^I. \tag{17}$$

A nice property of this prior is that it leads to marginal prior probabilities $\mathbb{P}(d_j = 0) = \mathbb{P}(d_j > 0) = 1/2$. Elsewhere this is often achieved by assigning independent priors to the p covariates, which implies that averaged over all models, $I \sim \mathrm{Bin}(p, 1/2)$. It is clear that our uniform prior on I allows the data $\boldsymbol{y}$ to have a maximum effect on the posterior of I because it is the reference prior (Bernardo, 1979). Note that this prior actually favours models with high or low numbers of covariates, as there are fewer such models. This or similar model priors have been used in a number of previous papers, including *e.g.* George and McCulloch (1993) and Ley and Steel (2009).

Alternatively, one might also use a fixed (independent of K) prior probability for a linear effect ($d_j = 1$). This is appropriate for the situation where one explicitly wants to test linearity versus nonlinearity of each effect. Furthermore, a multiplicity correction for these tests can be implemented by assuming that the number of smoothly included covariates (J) is uniformly distributed on $\{0, 1, \ldots, I\}$ and their choice is uniform on all possible choices. This would add one level to the prior hierarchy.

As the model space $\mathcal{D}^p$ grows exponentially in the number of covariates p, only for small values of p all possible models can be evaluated. Otherwise the marginal likelihoods $f(\boldsymbol{y} \mid \boldsymbol{d})$ and posterior model probabilities $f(\boldsymbol{d} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{d}) f(\boldsymbol{d})$ can be computed only for a subset of the model space. Usually this subset is determined by stochastic search procedures (Madigan and York, 1995). Here we propose to use a simple Metropolis-Hastings algorithm with two possible move types in the proposal kernel:

Move Sample a covariate index $j \sim \mathrm{U}\{1, 2, \ldots, p\}$ and decrease or increase d_j to the next adjacent value in $\mathcal{D}$ (with probability $1/2$ each, or deterministically if $d_j = 0$ or $d_j = K$, respectively).

Swap Sample a pair $(i, j) \sim \mathrm{U}\{(1, 1), (1, 2), \ldots, (p, p)\}$ of covariate indices ($i \leq j$) and swap d_i and d_j.

The 'Swap' move is designed to efficiently trace models with high posterior probability even in situations where covariates are almost collinear. For each Metropolis-Hastings iteration, a 'Move' is chosen with some fixed probability (we use $3/4$), and otherwise a 'Swap'. Denote the current model by $\boldsymbol{d}$, then the proposed model $\boldsymbol{d}'$ is accepted with probability

$$\alpha(\boldsymbol{d}' \mid \boldsymbol{d}) = 1 \wedge \frac{f(\boldsymbol{y} \mid \boldsymbol{d}') f(\boldsymbol{d}') q(\boldsymbol{d}' \mid \boldsymbol{d})}{f(\boldsymbol{y} \mid \boldsymbol{d}) f(\boldsymbol{d}) q(\boldsymbol{d} \mid \boldsymbol{d}')}$$

where the calculation of the proposal probability ratio $q(\boldsymbol{d}' \mid \boldsymbol{d}) / q(\boldsymbol{d} \mid \boldsymbol{d}')$ is straightforward (see the supplementary material).

2.3 Simulation Study

In order to study the performance of our approach in identifying the true model, we performed a simulation study. Full details are provided in the supplementary material; Here we summarise the main results. Three different true models were simulated: The first model ("null") was the null model with $p = 20$ nuisance covariates. The second model ("small") also had $p = 20$ covariates of which 3 had a linear effect and 3 had a nonlinear (quadratic, sine, and skew-normal density) effect. Correlations of different strength were generated between some of the covariates. The third model ("large") was identical to the second model, but included additional 80 nuisance covariates, which were independent of the first 20 covariates. For the "small" and "large" models, one covariate was chosen to be a surrogate for the true (quadratic) effect of another covariate. It masks the quadratic effect if only linear effects can be fitted by a variable selection algorithm. For three different sample sizes $n = 50, 100, 1000$, and for the three different true models, we simulated n observations from the Gaussian additive model (1) with $\beta_0 = 0$ and $\sigma^2 = 0.2^2$. This was repeated 50 times for each combination of model and sample size, in order to assess the sampling variability.

We applied the proposed additive model selection approaches to each data set, using the hyper-priors (11) and (12) ("hyper-g splines" and "hyper-g/n splines", respectively). As the computational complexity of the marginal likelihood (14) is cubic in the spline basis dimension K (see the supplementary material), we want to use splines with few, quantile-based knots. Therefore, we choose cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we got basis matrices $\boldsymbol{Z}_j$ with $K = 8$ columns from 6 inner knots at the septiles. We applied the stochastic search algorithm described in Section 2.2 with 10^6 iterations.

We compared the results with those from pure variable selection including only linear functions ("hyper-g linear" and "hyper-g/n linear"), Bayesian fractional polynomials ("Bayesian FPs") (Sabanés Bové and Held, 2011a), spike-and-slab function selection ("Spike-and-slab", Scheipl et al., 2012) and splines knot selection ("Knot selection", Denison, Mallick, and Smith, 1998, using code from chapters 3 and 4 in Denison, Holmes, Mallick, and Smith, 2002).

Concerning the discovery of the true set of influential covariates, the additive model selection procedures introduced in this paper were very competitive with the considered alternative methods, as is illustrated in Table 1. In particular, they showed clear advantages in the case of small and moderate sample sizes. Using splines instead of

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	$83_{(8)}$	$84_{(7)}$	$84_{(9)}$	$49_{(25)}$	$65_{(13)}$	$86_{(14)}$	$2_{(26)}$	$74_{(15)}$	$87_{(16)}$
Hyper-g/n splines	$86_{(10)}$	$91_{(6)}$	$97_{(3)}$	$47_{(24)}$	$68_{(14)}$	$87_{(13)}$	$0_{(24)}$	$75_{(15)}$	$89_{(15)}$
Hyper-g linear	$20_{(7)}$	$21_{(6)}$	$23_{(7)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$
Hyper-g/n linear	$50_{(16)}$	$64_{(15)}$	$90_{(8)}$	$0_{(0)}$	$0_{(1)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$
Bayesian FPs	$37_{(6)}$	$37_{(7)}$	$37_{(6)}$	$2_{(15)}$	$35_{(16)}$	$3_{(23)}$	$0_{(9)}$	$47_{(19)}$	$37_{(28)}$
Spike-and-slab	$89_{(6)}$	$93_{(2)}$	$98_{(0)}$	$3_{(5)}$	$45_{(6)}$	$79_{(2)}$	$0_{(0)}$	$10_{(8)}$	$71_{(5)}$
Knot selection	$92_{(8)}$	$94_{(5)}$	$98_{(2)}$	$0_{(1)}$	$34_{(20)}$	$95_{(6)}$	$0_{(0)}$	$0_{(1)}$	$89_{(9)}$

Table 1 *– Median posterior probability of the true model in percentage, when the true model is defined by correct variable inclusion. Standard deviations (in parentheses) are computed from the 50 replications.*

only linear functions proved essential for the discovery of the masked quadratic effect and hence convergence to the true model. Looking at the standard deviations in the 50 replications, we observe for the hyper-g and hyper-g/n spline methods a relatively high variability for $n = 50$, which decreases then for larger sample sizes. Interestingly the variability is increasing for the Bayesian FPs, and no clear trend is visible for the spike-and-slab and knot selection methods.

Variable inclusion performance did not differ substantively with respect to sensitivity, specificity and area under the ROC curve between the considered methods, with the exception of a slightly worse performance of the two linear methods. However, as shown in Table 2, the hyper-g and hyper-g/n spline methods were clearly better in distinguishing the truly effective covariates from the highly correlated nuisance covariates. Moreover, for small sample sizes, they outperformed the other nonlinear methodologies concerning the discovery of the masked quadratic effect. In this task the merely linear methods obviously failed. With respect to sampling variability, the proposed spline methods are very competitive, with smallest variability among all methods for larger sample sizes.

Concerning the average mean squared errors of the model-averaged posterior mean

	small			large		
	$n = 50,$	100,	1000	$n = 50,$	100,	1000
Hyper-g splines	$75_{(29)}$	$97_{(3)}$	$98_{(3)}$	$26_{(36)}$	$100_{(0)}$	$100_{(0)}$
Hyper-g/n splines	$79_{(26)}$	$97_{(3)}$	$98_{(3)}$	$20_{(50)}$	$100_{(0)}$	$100_{(0)}$
Hyper-g linear	$18_{(17)}$	$44_{(19)}$	$87_{(8)}$	$6_{(11)}$	$26_{(33)}$	$98_{(3)}$
Hyper-g/n linear	$22_{(21)}$	$48_{(22)}$	$90_{(8)}$	$17_{(44)}$	$26_{(33)}$	$98_{(2)}$
Bayesian FPs	$41_{(33)}$	$89_{(12)}$	$68_{(16)}$	$9_{(18)}$	$92_{(19)}$	$81_{(15)}$
Spike-and-slab	$30_{(19)}$	$88_{(5)}$	$97_{(0)}$	$1_{(2)}$	$60_{(19)}$	$97_{(1)}$
Knot selection	$9_{(19)}$	$78_{(22)}$	$99_{(1)}$	$4_{(11)}$	$13_{(20)}$	$99_{(3)}$

Table 2 – *Average difference* $\frac{1}{2}(P_{16} + P_{17}) - \frac{1}{3}(P_{18} + P_{19} + P_{20})$ *of inclusion probabilities* $P_j = \mathbb{P}\{m_j(x_j) \neq 0 \,|\, \boldsymbol{y}\}$ *(in percentage points) between the truly effective covariates* x_{16} *and* x_{17} *and the nuisance covariates* x_{18}, x_{19}, x_{20}*, which had correlation* 0.8 *with* x_{16} *and* x_{17}*. (The optimal value is* 100*, the worst value is* -100*.) Standard deviations (in parentheses) are computed from the 50 replications.*

function estimates $\hat{m}_j(x_j)$, the proposed additive model selection procedures were very competitive. They performed well or better than the best compared methods each, as is shown in Table 3. It is interesting that the hyper-g splines were slightly but consistently better than the hyper-g/n splines. We also investigated the coverage rates of pointwise 95% credible intervals for the functions, and found that the two proposed methods were slightly conservative.

Finally, the average computational effort of the two proposed additive model selection procedures ranged between one minute for $n = 100$ in a "null" data set to about 50 minutes for $n = 50$ in a "large" data set (times to be expected on a 2.8 GHz single-core CPU, see the supplementary material for more details).

Average MSE	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	344.38	114.14	22.43	39.15	10.32	1.68	30.42	1.88	0.33
Hyper-g/n splines	462.92	71.72	2.17	47.82	18.33	3.20	784.44	2.78	0.61
Hyper-g linear	7586.25	1378.49	137.06	158.10	133.55	121.97	45.11	32.26	24.36
Hyper-g/n linear	2155.39	182.62	6.78	189.57	169.00	120.96	378.07	36.23	26.09
Bayesian FPs	1424.17	283.76	19.20	16837.92	3026.61	29.51	76.78	356.30	5.80
Spike-and-slab	19038.78	18224.91	5660.40	80.94	14.00	2.09	45.45	8.71	0.81
Knot selection	337.77	36.79	0.65	180.03	35.29	2.07	47.23	29.33	0.78

Standard deviation	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1268.26	341.24	114.91	36.32	3.49	0.26	17.89	0.51	0.05
Hyper-g/n splines	1435.00	169.16	8.59	28.22	6.05	1.20	1720.04	0.88	0.19
Hyper-g linear	20871.36	2197.63	208.82	23.28	20.47	5.28	8.15	5.19	1.10
Hyper-g/n linear	6172.70	446.47	33.01	27.02	28.45	4.55	589.21	4.61	1.59
Bayesian FPs	5760.39	973.94	38.44	118315.78	21154.38	5.63	248.39	2471.83	0.97
Spike-and-slab	8768.03	5619.08	1762.70	30.05	6.43	0.40	5.13	3.60	0.09
Knot selection	1734.24	126.34	2.35	39.28	28.25	0.40	6.89	4.59	0.32

Table 3 – *Average mean squared errors (top table, in 10^{-8} units for the "null" model, and 10^{-4} units for the "small" and "large" models) and corresponding standard deviations (bottom table, same units as in top table) of function estimates. Numbers are averaged over all covariates and the 50 replications, standard deviations are computed from the 50 replications.*

3 Generalised Additive Models

Now we extend the above setting and assume that the covariate effects $m_j(x_j)$ enter additively into the linear predictor

$$\eta = \beta_0 + \sum_{j=1}^{p} m_j(x_j) \tag{18}$$

of an exponential family distribution with canonical parameter θ, mean $\mathbb{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\operatorname{Var}(y) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see McCullagh and Nelder, 1989). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w. For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ is

$$\boldsymbol{\eta} = \mathbf{1}_n\beta_0 + \boldsymbol{X}_d\boldsymbol{\beta}_d + \boldsymbol{Z}_d\boldsymbol{u}_d \tag{19}$$

and the likelihood is

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d) \propto \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i}\right\}. \tag{20}$$

The main challenge for the derivation of a generalised g-prior is that the marginal density $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_d)$, which results from integrating out the spline coefficient vector

$$\boldsymbol{u}_d \mid \boldsymbol{\rho}_d \sim \mathrm{N}_{JK}(\mathbf{0}_{JK}, \boldsymbol{D}_d) \tag{21}$$

from (20), has no closed form. In particular, it is not Gaussian, in contrast to (8).

Calculation of the degrees of freedom $d_j(\rho_j)$ for a smoothly modelled term m_j can be carried out with a reasonable generalisation of (6), that is (see Ruppert et al., 2003, section 11.4)

$$d_j(\rho_j) = \operatorname{tr}\{(\boldsymbol{Z}_j^T\widehat{\boldsymbol{W}}\boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1}\boldsymbol{Z}_j^T\widehat{\boldsymbol{W}}\boldsymbol{Z}_j\} + 1, \tag{22}$$

which uses a fixed weight matrix $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\mathbf{1}_n\widehat{\beta}_0)$, where $\boldsymbol{W}(\boldsymbol{\eta}) = \operatorname{diag}\{(dh(\eta_i)/d\eta)^2/\operatorname{Var}(y_i)\}_{i=1}^n$ is the usual generalised linear model weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model $\boldsymbol{d} = \mathbf{0}_p$. This definition avoids dependence of $\rho_j(d_j)$ on the model $\boldsymbol{d}$ under consideration and serves as simplification. In particular it again allows to invert $d_j(\rho_j)$

to obtain the variance component ρ_j for a given degree d_j. As a consequence, we next need to generalise the orthogonalisation of the original covariate vector $\tilde{\boldsymbol{x}}_j$ and spline basis matrix $\tilde{\boldsymbol{Z}}_j$ from (4) and (5) to

$$\boldsymbol{x}_j = \tilde{\boldsymbol{x}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{x}}_j}{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \mathbf{1}_n} \tag{23}$$

$$\text{and} \quad \boldsymbol{Z}_j = \tilde{\boldsymbol{Z}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{Z}}_j}{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \mathbf{1}_n} - \boldsymbol{x}_j \frac{\boldsymbol{x}_j^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{Z}}_j}{\boldsymbol{x}_j^T \widehat{\boldsymbol{W}} \boldsymbol{x}_j}, \tag{24}$$

implying that $\mathbf{1}_n$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ are orthogonal to each other with respect to the inner product in terms of $\widehat{\boldsymbol{W}}$. This ensures again that (22) correctly captures only the degrees of freedom associated with the nonlinear part of m_j.

3.1 Hyper-g Priors for Generalised Additive Models

We will now derive a generalised g-prior analogous to (10) for the linear coefficient vector $\boldsymbol{\beta}_d$ in the generalised additive model. The idea is to apply the iterative weighted least squares (IWLS) approximation to the non-normal likelihood (20) to obtain an approximate normal model of the form (7) and then derive the resulting g-prior (10). With a slight abuse of notation, *e.g.* $h(\boldsymbol{\eta}) = (h(\eta_1), \dots, h(\eta_n))^T$, let

$$\boldsymbol{z}_0 = \boldsymbol{\eta}_0 + \operatorname{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}\{\boldsymbol{y} - h(\boldsymbol{\eta}_0)\} \tag{25}$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\boldsymbol{y})$ around $\boldsymbol{y} = h(\boldsymbol{\eta}_0)$. Then

$$\boldsymbol{z}_0 \,|\, \beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d \overset{\text{approx}}{\sim} \mathrm{N}\big(\mathbf{1}_n\beta_0 + \boldsymbol{X}_d\boldsymbol{\beta}_d + \boldsymbol{Z}_d\boldsymbol{u}_d,\, \boldsymbol{W}_0^{-1}\big) \tag{26}$$

with $\boldsymbol{W}_0 = \boldsymbol{W}(\boldsymbol{\eta}_0)$ is the working normal model (see *e.g.* McCullagh and Nelder, 1989, p. 40). The IWLS algorithm iteratively updates $\boldsymbol{\eta}_0$ by weighted least squares estimation of the coefficients in (26). Here, we fix $\boldsymbol{\eta}_0 = \mathbf{0}_n$, which is the value expected *a priori*. Then we rewrite (26) using $\tilde{\boldsymbol{z}}_0 = \boldsymbol{W}_0^{1/2}\boldsymbol{z}_0$, $\tilde{\mathbf{1}}_n = \boldsymbol{W}_0^{1/2}\mathbf{1}_n$, $\tilde{\boldsymbol{X}}_d = \boldsymbol{W}_0^{1/2}\boldsymbol{X}_d$ and $\tilde{\boldsymbol{Z}}_d = \boldsymbol{W}_0^{1/2}\boldsymbol{Z}_d$ as

$$\tilde{\boldsymbol{z}}_0 \,|\, \beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d \overset{\text{approx}}{\sim} \mathrm{N}(\tilde{\mathbf{1}}_n\beta_0 + \tilde{\boldsymbol{X}}_d\boldsymbol{\beta}_d + \tilde{\boldsymbol{Z}}_d\boldsymbol{u}_d,\, \boldsymbol{I}_n), \tag{27}$$

which brings us back to a normal model of the form in (7). By computing the corresponding g-prior (10), we arrive at the generalised g-prior

$$\boldsymbol{\beta}_d \,|\, g, \boldsymbol{\rho}_d \sim \mathrm{N}_I(\mathbf{0}_I, g\boldsymbol{J}_0^{-1}) \tag{28}$$

with prior precision matrix proportional to

$$\begin{aligned} \boldsymbol{J}_0 &= \tilde{\boldsymbol{X}}_d^T(\boldsymbol{I}_n + \tilde{\boldsymbol{Z}}_d\boldsymbol{D}_d\tilde{\boldsymbol{Z}}_d^T)^{-1}\tilde{\boldsymbol{X}}_d \\ &= \boldsymbol{X}_d^T\boldsymbol{W}_0^{1/2}(\boldsymbol{I}_n + \boldsymbol{W}_0^{1/2}\boldsymbol{Z}_d\boldsymbol{D}_d\boldsymbol{Z}_d^T\boldsymbol{W}_0^{1/2})^{-1}\boldsymbol{W}_0^{1/2}\boldsymbol{X}_d. \end{aligned} \tag{29}$$

An appealing feature of this prior is that it directly generalises the g-prior proposed by Sabanés Bové and Held (2011b) for generalised linear models, to which it reduces when there are no spline effects in the model, *i.e.* $\boldsymbol{J}_0 = \boldsymbol{X}_d^T\boldsymbol{W}_0\boldsymbol{X}_d$. An alternative and more rigorous derivation of (29) as the Fisher information obtained from a Laplace approximation to the marginal model $f(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_d)$ is provided in Appendix B.

The generalised hyper-g prior

$$f(\beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d, g) = f(\beta_0)f(\boldsymbol{\beta}_d \,|\, g, \boldsymbol{\rho}_d)f(g)f(\boldsymbol{u}_d) \tag{30}$$

is defined to comprise the locally uniform prior $f(\beta_0) \propto 1$ on the intercept β_0, the generalised g-prior (28) on the linear coefficient vector $\boldsymbol{\beta}_d$, the penalty prior (21) on the spline coefficient vector $\boldsymbol{u}_d$, and some proper hyper-prior $f(g)$ on the hyper-parameter g. Posterior inference under this prior can be implemented as outlined in the following. The efficient R-package "`hypergsplines`" for this and all other computations in this paper is available from R-Forge at `http://hypergsplines.r-forge.r-project.org/`. For installation, just type `install.packages("hypergsplines",repos="http://r-forge.r-project.org")` into R.

Let $\boldsymbol{X}_a = (\mathbf{1}_n, \boldsymbol{X}_d, \boldsymbol{Z}_d)$ and $\boldsymbol{\beta}_a = (\beta_0, \boldsymbol{\beta}_d^T, \boldsymbol{u}_d^T)^T$ denote the grand design matrix and regression coefficient vector, respectively, such that $\boldsymbol{\eta} = \boldsymbol{X}_a\boldsymbol{\beta}_a$. The prior for $\boldsymbol{\beta}_a$ conditional on g has a Gaussian form with mean zero and singular precision matrix $\mathrm{diag}(0, g^{-1}\boldsymbol{J}_0, \boldsymbol{D}_d^{-1})$. Thus, the Gaussian approximation of $f(\boldsymbol{\beta}_a \,|\, \boldsymbol{y}, g, \boldsymbol{d})$, which is necessary for the Laplace approximation of $f(\boldsymbol{y} \,|\, g, \boldsymbol{d})$, can be obtained by the Bayesian IWLS algorithm (West, 1985). Afterwards, an approximation of the marginal likelihood of model $\boldsymbol{d}$,

$$f(\boldsymbol{y} \mid \boldsymbol{d}) = \int_0^\infty f(\boldsymbol{y} \mid g, \boldsymbol{d}) f(g) \, dg, \tag{31}$$

is obtained by numerical integration of the Laplace approximation $\tilde{f}(\boldsymbol{y} \mid g, \boldsymbol{d})$. For small sample sizes, using a higher order Laplace approximation can be useful, see Sabanés Bové and Held (2011b, section 3.1). Note that integrated Laplace approximations have successfully been applied in a more general context (Rue, Martino, and Chopin, 2009). Finally, we can use a tuning-free Metropolis-Hastings algorithm to sample from the joint posterior of $\boldsymbol{\beta}_a$ and g in a specific model $\boldsymbol{d}$: Values g are sampled on the log-scale from a proposal density obtained by linear interpolation of pairs $\{z_j, \tilde{f}(z_j, \boldsymbol{y} \mid \boldsymbol{d})\}$, $j = 1, \dots, 20$, which are already used for the above numerical integration of the Laplace approximation. Here $\tilde{f}(z, \boldsymbol{y} \mid \boldsymbol{d}) = \tilde{f}(\boldsymbol{y} \mid g, \boldsymbol{d}) f(g) g$ is the approximated unnormalised posterior density of $z = \log(g)$. Note that this sampling scheme for g can be interpreted as an approximate griddy Gibbs sampler (Ritter and Tanner, 1992). Conditional on the proposed value of g, a Gaussian proposal density for $\boldsymbol{\beta}_a$ is obtained by performing one or more IWLS steps from the previous state of $\boldsymbol{\beta}_a$ (Gamerman, 1997). See Sabanés Bové and Held (2011b, section 3), on which this implementation is based on, for more details on the computations.

3.2 Application

We now apply the generalised additive model selection approach to the logistic regression of $p = 7$ potential risk factors on the presence of diabetes in $n = 532$ women of Pima Indian heritage (Ripley, 1996; Frank and Asuncion, 2010), see Table 4 for details. We use cubic O'Sullivan splines with 4 inner knots at the quintiles and the hyper-prior (12), and explore the model space of dimension $7^7 = 823\,543$ with 10^6 iterations of the stochastic search algorithm. Note that the most complex model spends $4 \cdot 7 = 28$ degrees of freedom. Considering the recommendation that a parametric logistic regression model should contain at least 10 events (successes or failures) for each independent explanatory variable (Peduzzi, Concato, Kemper, Holford, and Feinstein, 1996), this most complex model would be large because we only have 177 successes in this data set. This rule easily extends to nonparametric logistic regression by replacing the number of explanatory variables by the total degrees of freedom. From this perspective it is not recommended to use more knots for the spline bases. More knots also do not change the results in this example, as we have seen when using 9 inner knots at the deciles.

The computational complexity is higher than for the normal response case, with 95 minutes required for the evaluation of the 39 081 models found. We validated the results with an exhaustive evaluation of all models, requiring 33 hours. Indeed, the stochastic search found 99% of the posterior probability mass and the 733 top models.

Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m^2]
x_6	Diabetes pedigree function
x_7	Age [years]

Table 4 – *Description of the variables in the Pima Indian diabetes data set. Note that the original dataset has $n = 768$ observations and $p = 8$ explanatory variables, but several missing values. We dropped the variable `insulin` with the highest proportion of missing values and removed the remaining rows with missing data to perform a complete case analysis.*

In Table 5 the marginal posterior probabilities for linear and smooth inclusion of the covariates are shown. There is clear evidence for inclusion of the covariates x_2, x_5, x_6 and x_7, which have posterior inclusion probabilities over 96%. For the other three covariates, the inclusion probability is below 30%. Smooth modelling of the effects of x_5, x_6 and x_7 seems to be necessary, while this is not so clear for x_2.

In order to examine the mixing properties of the stochastic search algorithm proposed

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
not included ($d_j = 0$)	0.74	0.00	0.88	0.91	0.00	0.04	0.01
linear ($d_j = 1$)	0.07	0.48	0.06	0.04	0.11	0.26	0.00
smooth ($d_j > 1$)	0.19	0.52	0.06	0.05	0.89	0.70	0.99

***Table 5** – Marginal posterior inclusion probabilities in the Pima Indian diabetes data set.*

in Section 2.2, we compared the results based on starting the MCMC chain from the full model with $d_j = 4$ instead of the previously used null model with $d_j = 0$ ($j = 1, \dots, p$). The results are very close: for example, the entries in Table 5 differ by at most $2.28 \cdot 10^{-4}$, and the top 500 models which were visited by the chains are identical. These results are an indication that slow mixing is not a problem for the presented stochastic search algorithm for this example. It is recommended to perform similar checks for all applications.

Figure 1 shows the estimated covariate effects in the *maximum a posteriori* (MAP) model which features a linear term for x_2 and smooth terms for x_5, x_6 and x_7. The estimates are obtained from 10 000 MCMC samples (every 2nd sample after burning the first 1000 iterations of the Markov chain). Using two IWLS steps per proposal yielded an acceptance rate of 67%. Note that for linear functions m_j, the pointwise credible intervals coincide with the simultaneous credible intervals (Besag, Green, Higdon, and Mengersen, 1995, p. 30). This is because all straight lines samples intersect in one point, which is due to the centring of the covariates in (23). Furthermore, we observe that the Chib and Jeliazkov (2001) estimate (−240.924, MCMC standard error 0.008) of the log marginal likelihood of the MAP model, which was also computed, is quite close to the integrated Laplace approximation (−241.01). This indicates that the integrated Laplace approximation is fairly accurate.

When the main interest lies in variable selection, multiple models which feature the same covariates can be summarised into a single meta-model as follows: The posterior probabilities of the sub-models are summed up to give the posterior probability of the

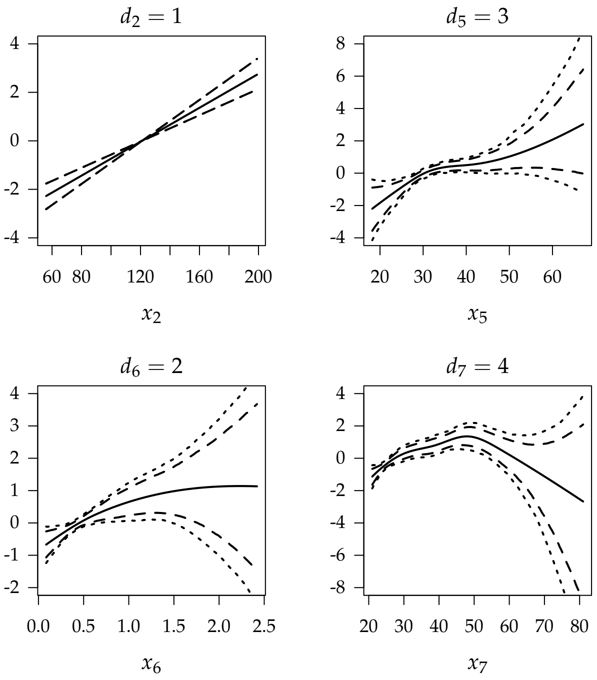

Figure 1 *– Estimated covariate effects in the MAP model for the Pima Indian diabetes data set, based on 10 000 MCMC samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

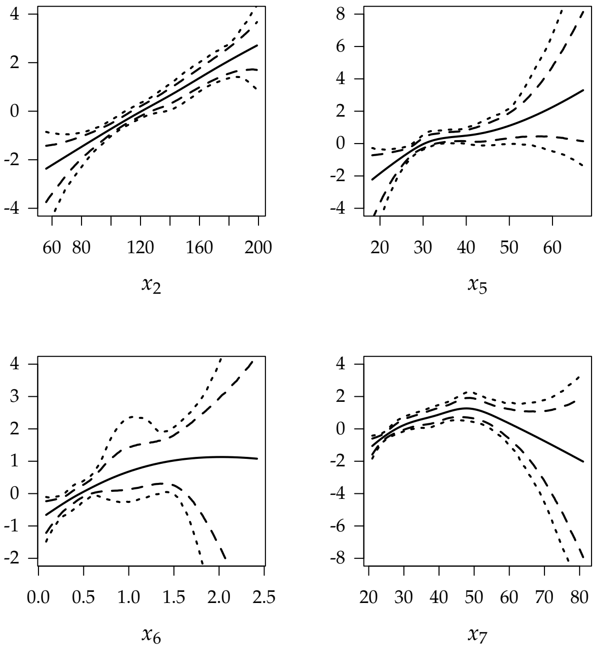

Figure 2 – *Estimated covariate effects in the best meta-model (and median probability meta-model) for the Pima Indian diabetes data, based on 20 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

meta-model, and estimates in the meta-model are obtained by averaging the sub-models with weights proportional to their posterior probabilities (see *e. g.* Hoeting, Madigan, Raftery, and Volinsky, 1999, for model averaging). Here the best meta-model includes x_2, x_5, x_6 and x_7 and has posterior probability 0.598. The corresponding estimates of the covariate effects are shown in Figure 2. This best meta-model happens to be identical with the median probability meta-model, which features all covariates having marginal posterior inclusion probability greater than 50% (Barbieri and Berger, 2004), *cp*. Table 5. Similarly, it could be interesting to summarise models which only differ in the degrees of freedom for smooth terms. This would correspond to the situation of testing linearity versus nonlinearity of covariate effects (*cp*. Section 2.2).

In summary, the results are qualitatively similar to those obtained with a FP modelling approach by Sabanés Bové and Held (2011b, section 5) and with a cubic smoothing spline approach by Cottet et al. (2008, section 3.2). It is interesting that in the earlier work by Yau et al. (2003, section 5.2), a very low posterior inclusion probability (0.07) for x_6 was reported for a different subset of the original Pima Indian diabetes data set. If pure variable selection without covariate transformation is considered, as in Holmes and Held (2006, section 2.6) and Sabanés Bové and Held (2011b, section 4), the strong nonlinear effect of x_7 is missed completely, and instead x_1 gets a higher posterior inclusion probability. This may be a case of a masked nonlinear effect, as was simulated in Section 2.3, and highlights the importance of allowing for nonlinear covariate effects.

4 Discussion

Our Bayesian approach to simultaneous variable and function selection in generalised regression is based on fixed-dimensional spline bases and penalty-parameter smoothness control. In this way it is coherent and differs from knot-selection approaches such as Smith and Kohn (1996) and Denison et al. (1998). We found that fixed-dimensional spline bases based on a small number of knots are flexible enough to capture the functional forms we expect (see *e.g.* Abrahamowicz, MacKenzie, and Esdaile, 1996). Furthermore, at least in the example from Section 3.2, the results are very robust to increasing the number of knots. In the interest of computation times we thus recommend to use only a small number of knots. Moreover, by using fixed-dimensional smooth components we can constrain a covariate effect to be exactly linear. This enables us to look at posterior probabilities of linear *versus* smooth inclusion of covariates. Approaches which use variable-dimensional smooth components and select knots, as Denison et al. (1998), cannot fit linear functions.

We are only considering roughness penalties on a fixed grid of values, which scales automatically for each covariate via the degrees of freedom transformation. We found that it is a very useful approximation of a continuous scale. One possibility for checking the quality of the discrete approximation is to optimise the marginal likelihood of the MAP model with respect to the degrees of freedom of the covariates included. That is, an optimisation of $f(\boldsymbol{y} \mid \boldsymbol{d})$ over the continuous range $1 < d_j < K + 1$ is performed for all covariates included in the MAP model. For example, the MAP configuration for the

Pima Indian diabetes data is $(0, 1, 0, 0, 3, 2, 4)$ and the resulting optimised configuration is $(0, 1, 0, 0, 3.42, 2.1, 3.74)$, which increases the log marginal likelihood from -241.01 to -240.86. Although d_5 and d_7 changed considerably in the optimisation, the resulting function estimates are very similar to those from the MAP model in Figure 1. In all examples we have looked at, the resulting optimised models yielded very similar results compared to the MAP model, which indicates that the fixed grid approximation is good enough. In this regard, our approach is close to many popular Lasso-type proposals, which optimise the tuning-parameters on a fixed grid via cross-validation (*e. g.* Zou and Hastie, 2005). Cantoni and Hastie (2002) propose a likelihood-ratio-type test statistic to compare additive models with different degrees of freedom. Fong, Rue, and Wakefield (2010) use a similar scaling to examine the prior on the degrees of freedom implied by the prior on the variance component in a generalised linear mixed model. They also use O'Sullivan spline bases as we did in our applications, but they do not consider variable selection.

In a frequentist setting, Marra and Wood (2011, section 2.1) propose to use an additional penalty on the linear part of the spline function in order to shrink it adaptively to zero. To include variable selection, a lower threshold for the effective degrees of freedom must be chosen. Our generalised g-prior (28) also shrinks the linear parts of the spline functions to zero, where the prior covariance matrix takes the correlations between the covariates into account. Incorporating the covariates correlation in the coefficients prior allows for better discrimination between influential and correlated nuisance covariates. Empirical results from our simulation study in Section 2.3 support this. Furthermore, we explicitly ex- or include covariates and then compare the resulting models based on their posterior probabilities. This avoids *ad-hoc* choices of a threshold and leads to a coherent variable selection procedure.

We propose a conventional prior for the intercept and the linear coefficients, which directly generalises the hyper-g priors in the linear model (Liang et al., 2008) and in the generalised linear model (Sabanés Bové and Held, 2011b). Pauler (1998) proposes a related unit-information prior for the fixed effects in linear mixed models, but fixes $g = n$ in (10). Overstall and Forster (2010) propose a unit-information prior for the fixed effects in generalised linear mixed models, but the information matrix is based on the first-stage likelihood and not on the integrated likelihood as in our approach. Also, no hyper-prior on the parameter g is considered, because it is fixed at $g = n$. As they use an inverse-Wishart prior on the covariance matrix of the random effects, their approach is perhaps better suited to generic random effects models. Forster, Gill, and Overstall (2012) propose a novel reversible-jump MCMC algorithm to infer the corresponding posterior model probabilities. We are confident that our proposed generalised additive model selection procedure, which can be used with any of the various well-explored default priors in the linear model, is a competitive alternative to other approaches.

Appendix

Appendix A gives details on the closed form of the marginal likelihood (14) for normal additive models. In Appendix B, an alternative derivation of the prior precision matrix (29) in the generalised g-prior is presented.

A Closed Forms of Marginal Likelihood in Additive Models

Under the hyper-g prior (11), the marginal likelihood of the transformed response vector is (Liang et al., 2008)

$$f(\tilde{\boldsymbol{y}} \,|\, \boldsymbol{d}) \propto \|\boldsymbol{V}_{\boldsymbol{d}}^{-T/2}(\boldsymbol{y} - \boldsymbol{1}_n\bar{y})\|^{-(n-1)}(I+2)^{-1}{}_2\mathrm{F}_1\left(\frac{n-1}{2};1;\frac{I+4}{2};\tilde{R}^2_{\boldsymbol{d}}\right) \tag{32}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, ${}_2\mathrm{F}_1$ is the Gaussian hypergeometric function (Abramowitz and Stegun, 1964, p. 558) and $\tilde{R}^2_{\boldsymbol{d}}$ is the classical coefficient of determination in model (8). Under the hyper-g/n prior (12), the marginal likelihood in the standard linear model is

(Forte, 2011, p. 155)

$$f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d}) \propto n^{-I/2}(1-\tilde{R}_{\boldsymbol{d}}^2)^{-(n-1)/2}\frac{2}{I+2} \\ \times \mathrm{AF}_1\left(\frac{I}{2}+1; \frac{I+1-n}{2}; \frac{n-1}{2}; \frac{I}{2}+2; \frac{n-1}{n}, \frac{n-(1-\tilde{R}_{\boldsymbol{d}}^2)^{-1}}{n}\right), \qquad (33)$$

where AF_1 is the Appell hypergeometric function of the first kind (Appell, 1925). Colavecchia and Gasaneo (2004) provide Fortran code for computing this special function, which is accessible in R via the package "appell" (Sabanés Bové, 2012). For large sample sizes $(n > 100)$ or when the numerical computations of the special functions in (32) or (33) fail, we instead use Laplace approximations as described by Liang et al. (2008, Appendix A). See the supplementary material for details on efficient computation of $\tilde{R}_{\boldsymbol{d}}^2$.

B Approximate Fisher Information in Generalised Additive Models

In this section, we present a formal derivation of formula (29) as the approximate Fisher information obtained from a Laplace approximation to $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}})$. For ease of notation we restrict the presentation to canonical response functions where $\eta = \theta$ and omit subscripts where they are not necessary for understanding. With $\boldsymbol{\Phi} = \operatorname{diag}\{\phi/w_i\}_{i=1}^n$, we can then rewrite the likelihood (20) as

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{\boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta} - \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta})\right\}. \qquad (34)$$

We will now use the Laplace approximation to integrate (34) over $\boldsymbol{u}$ with respect to the prior $\boldsymbol{u} \mid \boldsymbol{\rho} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$.

We first need to maximise the unnormalised log posterior of $\boldsymbol{u}$,

$$\begin{aligned} l(\boldsymbol{u}) &= \log\{f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u})\} + \log\{f(\boldsymbol{u})\} \\ &= \boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta} - \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{D}^{-1}\boldsymbol{u} + \text{const}, \end{aligned} \qquad (35)$$

where β_0 and $\boldsymbol{\beta}$ in $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$ are considered to be fixed. The corresponding score vector is

$$\begin{aligned}\frac{d}{d\boldsymbol{u}} l(\boldsymbol{u}) &= \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} \boldsymbol{y} - \boldsymbol{Z}^T \operatorname{diag}\{b'(\boldsymbol{\eta})\} \boldsymbol{\Phi}^{-1} \mathbf{1} - \boldsymbol{D}^{-1} \boldsymbol{u} \\ &= \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{D}^{-1} \boldsymbol{u},\end{aligned}$$

where $\boldsymbol{\mu} = b'(\boldsymbol{\eta})$, and the corresponding Hessian is

$$\begin{aligned}\frac{d}{d\boldsymbol{u}} \frac{d}{d\boldsymbol{u}^T} l(\boldsymbol{u}) &= \frac{d}{d\boldsymbol{u}} \left\{ (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Phi}^{-1} \boldsymbol{Z} - \boldsymbol{u}^T \boldsymbol{D}^{-1} \right\} \\ &= -\boldsymbol{Z}^T \boldsymbol{W}(\boldsymbol{\eta}) \boldsymbol{Z} - \boldsymbol{D}^{-1}.\end{aligned}$$

Making one Newton-Raphson step from the starting point $\boldsymbol{u} = \mathbf{0}$, we get the approximate mode $\boldsymbol{u}^*$ of $l(\boldsymbol{u})$:

$$\begin{aligned}\boldsymbol{u}^* &= \mathbf{0} - \left(\frac{d}{d\boldsymbol{u}} \frac{d}{d\boldsymbol{u}^T} l(\mathbf{0}) \right)^{-1} \frac{d}{d\boldsymbol{u}} l(\mathbf{0}) \\ &= \left(\boldsymbol{Z}^T \boldsymbol{W}(\boldsymbol{\eta}_L) \boldsymbol{Z} + \boldsymbol{D}^{-1} \right)^{-1} \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_L), \qquad (36)\end{aligned}$$

where $\boldsymbol{\eta}_L = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\mu}_L = b'(\boldsymbol{\eta}_L)$. Note that this corresponds to the result of a second-order Taylor expansion of $l(\boldsymbol{u})$ around $\boldsymbol{u} = \mathbf{0}$. Hence, the Laplace approximation of $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})$ is

$$\begin{aligned}\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}) &\propto \exp(l(\boldsymbol{u}^*)) (2\pi)^{JK/2} \left| -\frac{d}{d\boldsymbol{u}} \frac{d}{d\boldsymbol{u}^T} l(\boldsymbol{u}^*) \right|^{-1/2} \\ &= \exp\left(\boldsymbol{y}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}^* - \mathbf{1}^T \boldsymbol{\Phi}^{-1} b(\boldsymbol{\eta}^*) - \frac{1}{2} \boldsymbol{u}^{*T} \boldsymbol{D}^{-1} \boldsymbol{u}^* \right) \\ &\quad \times (2\pi)^{JK/2} \left| \boldsymbol{Z}^T \boldsymbol{W}(\boldsymbol{\eta}^*) \boldsymbol{Z} + \boldsymbol{D}^{-1} \right|^{-1/2}, \qquad (37)\end{aligned}$$

where JK is the dimension of $\boldsymbol{u}$.

In order to derive the approximate Fisher information of $\boldsymbol{\beta}$ from $\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})$, we make two additional simplifying assumptions: First, we assume that $\boldsymbol{W}(\boldsymbol{\eta})$ does not vary much in $\boldsymbol{\beta}$, so that we can ignore the determinant in (37), for example. This is a common simplification, suggested *e.g.* in Breslow and Clayton (1993). Second, we approximate $b(\boldsymbol{\eta}^*)$ by a second-order Taylor expansion of $b(\boldsymbol{\eta})$ around $\boldsymbol{\eta}_L$, yielding

$$\mathbf{1}^T \boldsymbol{\Phi}^{-1} b(\boldsymbol{\eta}^*) \approx \mathbf{1}^T \boldsymbol{\Phi}^{-1} b(\boldsymbol{\eta}_L) + \boldsymbol{\mu}_L^T \boldsymbol{\Phi}^{-1} \boldsymbol{Z} \boldsymbol{u}^* + \frac{1}{2} \boldsymbol{u}^{*T} \boldsymbol{Z}^T \boldsymbol{W}_L \boldsymbol{Z} \boldsymbol{u}^*,$$

where $\boldsymbol{W}_L = \boldsymbol{W}(\boldsymbol{\eta}_L)$. Using these two simplifications and plugging in (36), we arrive at the expression

$$\begin{aligned}\log\{\tilde{f}(\boldsymbol{y}\,|\,\beta_0,\boldsymbol{\beta})\} &= \boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}_L - \boldsymbol{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}_L)\\ &\quad + (\boldsymbol{y}-\boldsymbol{\mu}_L)^T\boldsymbol{\Phi}^{-1}\boldsymbol{Z}\boldsymbol{u}^* - \frac{1}{2}\boldsymbol{u}^{*T}(\boldsymbol{Z}^T\boldsymbol{W}_L\boldsymbol{Z}+\boldsymbol{D}^{-1})\boldsymbol{u}^*\\ &= \boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}_L - \boldsymbol{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}_L)\\ &\quad + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_L)^T\boldsymbol{\Phi}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{W}_L\boldsymbol{Z}+\boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Phi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_L) \qquad (38)\end{aligned}$$

for the approximate marginal log-likelihood of β_0 and $\boldsymbol{\beta}$. From (38) we can finally approximate the Fisher information $\boldsymbol{J}(\beta_0,\boldsymbol{\beta}) = -\frac{d}{d\boldsymbol{\beta}}\frac{d}{d\boldsymbol{\beta}^T}\log\{f(\boldsymbol{y}\,|\,\beta_0,\boldsymbol{\beta})\}$ as

$$\begin{aligned}\tilde{\boldsymbol{J}}(\beta_0,\boldsymbol{\beta}) &= -\frac{d}{d\boldsymbol{\beta}}\frac{d}{d\boldsymbol{\beta}^T}\log\{\tilde{f}(\boldsymbol{y}\,|\,\beta_0,\boldsymbol{\beta})\}\\ &= \boldsymbol{X}^T\boldsymbol{W}_L^{1/2}\left(\boldsymbol{I}-\boldsymbol{W}_L^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{W}_L\boldsymbol{Z}+\boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{W}_L^{1/2}\right)\boldsymbol{W}_L^{1/2}\boldsymbol{X} \qquad (39)\\ &= \boldsymbol{X}^T\boldsymbol{W}_L^{1/2}(\boldsymbol{I}+\boldsymbol{W}_L^{1/2}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{W}_L^{1/2})^{-1}\boldsymbol{W}_L^{1/2}\boldsymbol{X}. \qquad (40)\end{aligned}$$

Evaluating the approximate Fisher information at $\beta_0 = 0, \boldsymbol{\beta} = \boldsymbol{0}$, such that $\boldsymbol{W}_L = \boldsymbol{W}(\boldsymbol{0})$, we recognise that $\tilde{\boldsymbol{J}}(0,\boldsymbol{0})$ from (40) is identical to $\boldsymbol{J}_0$ in formula (29). Note that the representation (39) can be better suited for computation: the second paragraph of Section 2.1 in the supplementary material applies here after replacing $\boldsymbol{Z}_d$ with $\boldsymbol{W}_L^{1/2}\boldsymbol{Z}$.

References

Abrahamowicz, M., MacKenzie, T., and Esdaile, J. M. (1996), “Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis,” *Journal of the American Statistical Association*, 91, 1432–1439.

Abramowitz, M. and Stegun, I. A. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover, ninth Dover printing, tenth GPO printing ed.

Aerts, M., Claeskens, G., and Wand, M. P. (2002), “Some theory for penalized spline generalized additive models,” *Journal of Statistical Planning and Inference*, 103, 455–470.

Appell, M. P. (1925), “Sur les fonctions hypergéométriques de plusieurs variables, les polynomes d’Hermite et autres fonctions spheriques dans l’hyperespace,” *Mémorial des sciences mathématiques*, 3, 1–75.

Barbieri, M. M. and Berger, J. O. (2004), “Optimal predictive model selection,” *Annals of Statistics*, 32, 870–897.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012), “Criteria for Bayesian model choice with application to variable selection,” *Annals of Statistics*, 40, 1550–1577.

Belitz, C. and Lang, S. (2008), “Simultaneous selection of variables and smoothing parameters in structured additive regression models,” *Computational Statistics and Data Analysis*, 53, 61–81.

Berger, J. O. and Pericchi, L. R. (2001), “Objective Bayesian methods for model selection: introduction and comparison,” in *Model Selection*, ed. Lahiri, P., Beachwood, OH: Institute of Mathematical Statistics, vol. 38 of *IMS Lecture Notes*, pp. 135–207.

Bernardo, J. M. (1979), “Reference posterior distributions for Bayesian inference,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 113–147.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), “Bayesian computation and stochastic systems (with discussion),” *Statistical Science*, 10, 3–66.

Björck, Å. (1967), “Solving linear least squares problems by Gram-Schmidt orthogonalization,” *BIT Numerical Mathematics*, 7, 1–21.

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

Brezger, A. and Lang, S. (2008), “Simultaneous probability statements for Bayesian P-splines,” *Statistical Modelling*, 8, 141–168.

Cantoni, E. and Hastie, T. (2002), “Degrees-of-freedom tests for smoothing splines,” *Biometrika*, 89, 251–263.

Celeux, G., Anbari, M. E., Marin, J.-M., and Robert, C. P. (2012), "Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation," *Bayesian Analysis*, 7, 477–502.

Chib, S. and Jeliazkov, I. (2001), "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, 96, 270–281.

Colavecchia, F. and Gasaneo, G. (2004), "f1: a code to compute Appell's F1 hypergeometric function," *Computer Physics Communications*, 157, 32–38.

Cottet, R., Kohn, R. J., and Nott, D. J. (2008), "Variable selection and model averaging in semiparametric overdispersed generalized linear models," *Journal of the American Statistical Association*, 103, 661–671.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. fully Bayes variable selection," *Journal of Statistical Planning and Inference*, 138, 888–900.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, Chichester: Wiley.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 333–350.

Eilers, P. H. C. and Marx, B. D. (2010), "Splines, knots, and penalties," *Wiley Interdisciplinary Reviews Computational Statistics*, 2, 637–653.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010), "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection," *Statistics and Computing*, 20, 203–219.

Fahrmeir, L., Kneib, T., and Lang, S. (2004), "Penalized structured additive regression for space-time data: A Bayesian perspective," *Statistica Sinica*, 14, 715–745.

Fong, Y., Rue, H., and Wakefield, J. (2010), "Bayesian inference for generalized linear mixed models," *Biostatistics*, 11, 397–412.

Forster, J., Gill, R., and Overstall, A. (2012), "Reversible jump methods for generalised linear models and generalised linear mixed models," *Statistics and Computing*, 22, 107–120.

Forte, A. (2011), "Objective Bayes Criteria for Variable Selection," Ph.D. thesis, Universitat de València, available at https://www.educacion.gob.es/teseo/imprimirFicheroTesis.do?fichero=22234.

Frank, A. and Asuncion, A. (2010), "UCI Machine Learning Repository," available at http://archive.ics.uci.edu/ml.

Friedman, J. H. (2001), "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29, 1189–1232.

Gamerman, D. (1997), "Sampling from the posterior distribution in generalized linear mixed models," *Statistics and Computing*, 7, 57–68.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

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

Held, L. (2004), "Simultaneous posterior probability statements from Monte Carlo output," *Journal of Computational and Graphical Statistics*, 13, 20–35.

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, 23, 53–60.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: a tutorial," *Statistical Science*, 14, 382–417.

Holmes, C. C. and Held, L. (2006), "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, 1, 145–168.

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

Kauermann, G. and Tutz, G. (2001), "Testing generalized linear and semiparametric models against smooth alternatives," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63, 147–166.

Kneib, T., Hothorn, T., and Tutz, G. (2009), "Variable selection and model choice in geoadditive regression models," *Biometrics*, 65, 626–634.

Ley, E. and Steel, M. F. (2009), "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression," *Journal of Applied Econometrics*, 24, 651–674.

— (2012), "Mixtures of g-priors for Bayesian model averaging with economic applications," *Journal of Econometrics*, 171, 251–266.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Madigan, D. and York, J. (1995), "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.

Marra, G. and Wood, S. N. (2011), "Practical variable selection for generalized additive models," *Computational Statistics and Data Analysis*, 55, 2372–2387.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, no. 37 in Monographs on Statistics and Applied Probability, New York: Chapman and Hall, 2nd ed.

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-dimensional additive modeling," *Annals of Statistics*, 37, 3779–3821.

Overstall, A. M. and Forster, J. J. (2010), "Default Bayesian model determination methods for generalised linear mixed models," *Computational Statistics and Data Analysis*, 54, 3269–3288.

Pauler, D. K. (1998), "The Schwarz criterion and related methods for normal linear models," *Biometrika*, 85, 13–27.

Peduzzi, P., Concato, J., Kemper, E., Holford, T., and Feinstein, A. (1996), "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, 49, 1373–1379.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008), "SpAM: Sparse additive models," in *Advances in Neural Information Processing Systems 20*, eds. Platt, J., Koller, D., Singer, Y., and Roweis, S., Cambridge, MA: MIT Press, pp. 1201–1208.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71, 319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press.

Sabanés Bové, D. (2012), *appell: Compute Appell's F1 hypergeometric function*, R package version 0.0-3, available at http://cran.r-project.org/web/packages/appell/.

Sabanés Bové, D. and Held, L. (2011a), "Bayesian fractional polynomials," *Statistics and Computing*, 21, 309–324.

— (2011b), “Hyper-g priors for generalized linear models,” *Bayesian Analysis*, 6, 387–410.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), “Spike-and-slab priors for function selection in structured additive regression models,” *Journal of the American Statistical Association*, 107, 1518–1532.

Scheipl, F., Kneib, T., and Fahrmeir, L. (2013), “Penalized likelihood and Bayesian function selection in regression models,” *Advances in Statistical Analysis*, to appear.

Scott, J. G. and Berger, J. O. (2010), “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem,” *Annals of Statistics*, 38, 2587–2619.

Smith, M. and Kohn, R. (1996), “Nonparametric regression using Bayesian variable selection,” *Journal of Econometrics*, 75, 317–343.

Tutz, G. and Binder, H. (2006), “Generalized additive modeling with implicit variable selection by likelihood-based boosting,” *Biometrics*, 62, 961–971.

Wand, M. P. (2003), “Smoothing and mixed models,” *Computational Statistics*, 18, 223–249.

Wand, M. P. and Ormerod, J. T. (2008), “On semiparametric regression with O’Sullivan penalized splines,” *Australian & New Zealand Journal of Statistics*, 50, 179–198.

West, M. (1985), “Generalized linear models: scale parameters, outlier accommodation and prior distributions,” in *Bayesian Statistics 2*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Amsterdam: North-Holland, pp. 531–558.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Boca Raton: Chapman & Hall/ CRC.

— (2011), “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73, 3–36.

Yau, P., Kohn, R. J., and Wood, S. (2003), “Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression,” *Journal of Computational and Graphical Statistics*, 12, 23–54.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. Goel, P. K. and Zellner, A., Amsterdam: North-Holland, vol. 6 of *Studies in Bayesian Econometrics and Statistics*, chap. 5, pp. 233–243.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Valencia: University of Valencia Press, pp. 585–603.

Zhang, H. H. and Lin, Y. (2006), "Component selection and smoothing for nonparametric regression in exponential families," *Statistica Sinica*, 16, 1021–1041.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 301–320.