Efficient Bayesian Multivariate Surface Regression

Feng Li and Mattias Villani *

November 21, 2013

Abstract

Methods for choosing a fixed set of knot locations in additive spline models are fairly well established in the statistical literature. While most of these methods are in principle directly extendable to non-additive surface models, they are less likely to be successful in that setting because of the curse of dimensionality, especially when there are more than a couple of covariates. We propose a regression model for a multivariate Gaussian response that combines both additive splines and interactive splines, and a highly efficient MCMC algorithm that updates all the knot locations jointly. We use shrinkage priors to avoid overfitting with different estimated shrinkage factors for the additive and surface part of the model, and also different shrinkage parameters for the different response variables. This makes it possible for the model to adapt to varying degrees of nonlinearity in different parts of the data in a parsimonious way. Simulated data and an application to firm leverage data show that the approach is computationally efficient, and that allowing for freely estimated knot locations can offer a substantial improvement in out-of-sample predictive performance.

KEYWORDS: Bayesian inference, Markov chain Monte Carlo, Surface regression, Splines, Free knots.

*Please note this is not the original paper. Please visit `http://dx.doi.org/10.1111/sjos.12022` to obtain the journal version of this paper.

1 Introduction

Flexible models of the regression function $\mathrm{E}(y|x)$ has been an active research field for decades, see e.g. Ruppert et al. (2003) for a recent textbook introduction and further references. Intensive research was initially devoted to kernel regression methods (Nadaraya 1964, Watson 1964, Gasser & Müller 1979), and later followed by a large literature on spline regression modeling. A spline is a linear regression on a set of nonlinear basis functions of the original regressors. Each basis function is defined from a knot in regressor space and the knots determine the points of flexibility of the fitted regression function. This gives rise to a locally adaptable model with continuity at the knots.

The most widely used models assume additivity in the regressors, i.e. $\mathrm{E}(y|x_1, ..., x_q) = \sum_{j=1}^{q} f_j(x_j)$, where $f_j(x_j)$ is a spline function for the jth regressor (Hastie & Tibshirani 1990). Assuming additivity is clearly a very convenient simplification, but it is also somewhat unnatural to make such a strong assumption in an otherwise very flexible model. This has motivated research on surface models with interactions between regressors. One line of research extends the additive models by including higher-order interactions of the spline basis functions, see e.g. the structured ANOVA approach or the tensor product basis in Hastie et al. (2009). The multivariate adaptive regression splines (MARS) introduced in Friedman (1991) is a version of the tensor product spline with interactions sequentially entering the model using a greedy algorithm. Regression trees (Breiman et al. 1984) is another popular class of models, with the BART model in Chipman et al. (2010) as its most prominent Bayesian member. Our paper follows a recent strand of literature that models surfaces using radial basis functions splines, see e.g. Buhmann (2003). A radial basis function is defined in $\mathbb{R}^q$ and has a value that depends only on the distance from a covariate vector ($\boldsymbol{x}$) to its q-dimensional knot ($\boldsymbol{\xi}$), e.g. the cubic radial basis $\|\boldsymbol{x} - \boldsymbol{\xi}\|^3$, where $\boldsymbol{x} = (x_1, ..., x_q)'$, $\boldsymbol{\xi} = (\xi_1, ..., \xi_q)'$ and $\|\cdot\|$ is the Euclidean norm. The model is again linear in the basis expanded space.

The basic challenge in spline regression is the choice of knot locations. This problem is clearly much harder for a general surface than it is for additive models since any manageable set of q-dimensional knots are necessarily sparse in $\mathbb{R}^q$ when q is moderate or large, a manifestation of the curse of dimensionality. The state-of-the-art inferential procedures place the knots at the centroids from a clustering of the regressor observations. The selected knot locations are kept fixed throughout the analysis. To prevent overfitting, Bayesian variable

selection methods are used to automatically remove or downweight the influence of the knots using Markov chain Monte Carlo (MCMC) methods (Smith & Kohn 1996). The reversible jump MCMC (RJMCMC) in for example Denison et al. (2002) treats the number of knots as unknown subject to an upper bound, but the location of the knots are still fixed throughout the analysis.

Using a fixed set of knot locations is impractical when estimating a surface with more than a few regressors. An algorithm that can move the knots rapidly over the regressor space is expected to be a clear improvement. All previous attempts have focused on efficient selection of fixed knots, and have paid little attention to moving the knots. The otherwise very elaborate RJMCMC approaches in Dimatteo et al. (2001), Denison et al. (1998), Gulam Razul et al. (2003) and Holmes & Mallick (2003) all include a very simple MCMC update where a single knot is re-located using a Metropolis random walk step with a proposal variance that is the same for all knots. There are typically strong dependencies between the knots, and local one-knot-at-a-time moves will lead to slow convergence of the algorithm and inability to escape from local modes. This is especially true in the surface case with more than a couple of regressors.

The main contribution in this paper is a highly efficient MCMC algorithm for the Gaussian multivariate surface regression where the locations of all knots are updated jointly. Rapid mixing of the knot locations is obtained from the following two features of our algorithm. First, the knots are simulated from a marginal posterior where the high-dimensional regression coefficients have been integrated out analytically. Second, the knots' proposal distribution is tailored to the posterior distribution using the posterior gradient, which we derive in compact analytical form and evaluate efficiently by a careful use of sparsity. We use a shrinkage prior on the regression coefficients to prevent overfitting, where the shrinkage hyperparameters are treated as unknowns and are estimated in a separate updating step. Also this step is tailored to the posterior using the gradient in analytical form.

Even a highly efficient MCMC algorithm is likely to have problems exploring the joint posterior of many surface knots in a high-dimensional covariate space. To deal with this, our model is decomposed into three parts: i) the original covariates entering in linear form, ii) additive spline basis functions and iii) radial basis functions for capturing the remaining part of the surface and interactions. The idea is to let the additive part of the model capture the

bulk of the nonlinearities so that the radial basis functions can focus exclusively on modeling the interactions. This way we can keep the number of knots in the interaction part of the model to a minimum, which is beneficial for MCMC convergence. We use separate shrinkage priors for the three parts of the model. Moreover, we also allow for separate shrinkage parameters in each response equation. This gives us an extremely flexible yet potentially parsimonious model where we can shrink out e.g. the surface part of the model in a subset of the response equations.

Our MCMC scheme is designed for a fixed number of knots, and we select the number of knots by Bayesian cross-validation of the log predictive score using parallel computing, see Section 3.3. This has the disadvantage of not accounting for the uncertainty regarding the number of knots as is done in RJMCMC schemes, but the benefits are substantially more robustness to variations in the prior and improved MCMC efficiency.

We illustrate our algorithm on simulated and real data, and compare the predictive performance of the models using Bayesian cross-validation techniques. We find that the free knots model constantly outperforms the model with fixed knots. Additionally, we find it is easier to obtain better fitting result by combining additive knots and surface knots in the model.

2 Bayesian multivariate surface regression

2.1 The model

Our proposed model is a Gaussian multivariate regression with three sets of covariates:

$$\boldsymbol{Y} = \boldsymbol{X}_o\boldsymbol{B}_o + \boldsymbol{X}_a(\boldsymbol{\xi}_a)\boldsymbol{B}_a + \boldsymbol{X}_s(\boldsymbol{\xi}_s)\boldsymbol{B}_s + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{Y}(n \times p)$ contains n observations on p response variables, and the rows of $\boldsymbol{E}$ are error vectors assumed to be iid $\mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The matrix $\boldsymbol{X}_o(n \times q_o)$ contains the original regressors (first column is a vector of ones for the intercept) and $\boldsymbol{B}_o$ holds the corresponding regression coefficients. The q_a columns of the matrix $\boldsymbol{X}_a(\boldsymbol{\xi}_a)$ are additive splines functions of the covariates in $\boldsymbol{X}_o$. Our notation makes it clear that $\boldsymbol{X}_a$ depends on the knots $\boldsymbol{\xi}_a$. Note that the knots in the additive part of the model are scalars, and that our model allows for

unequal number of knots in the different covariates. Finally, $\boldsymbol{X}_s(\boldsymbol{\xi}_s)$ contains the surface, or interaction, part of the model. The knots in $\boldsymbol{\xi}_s$ are q_o-dimensional vectors. Note how this decomposition makes it possible for the additive part of the model to capture the main part of the nonlinearities so that the number of knots in $\boldsymbol{X}_s$ is kept to a minimum. We will refer to the three different parts of the model as the *linear component*, the *additive component* and the *surface component*, respectively. We will refer to $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_s$ as the additive and surface knots, respectively. Likewise, $\boldsymbol{B}_a$ and $\boldsymbol{B}_s$ are the additive and surface coefficients.

There are a large number of different spline bases that one can use for the additive part of the model. The menu of choices for the surface basis is more limited, see Denison et al. (2002) for a survey of the most commonly used bases. We will use thin-plate splines for illustration, but our approach can be used with any basis with trivial changes, see Section 3 and Appendix A for computational details. The thin-plate spline basis in the surface case is of the form

$$\boldsymbol{x}_{sj}(\boldsymbol{\xi}_{sj}) = \|\boldsymbol{x}_o - \boldsymbol{\xi}_{sj}\|^2 \ln \|\boldsymbol{x}_o - \boldsymbol{\xi}_{sj}\|, \;\; j = 1, ..., q_s, \tag{2}$$

where $\boldsymbol{x}_o$ is one of the original data points and $\boldsymbol{\xi}_{sj}$ is the jth q_o-dimensional surface knot. The univariate thin-plate basis used in the additive part is a special case of the multivariate thin-plate in (2) where both the data point and the knot are one-dimensional.

For notational convenience, we sometimes write model (1) in compact form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{X} = [\boldsymbol{X}_o, \boldsymbol{X}_a, \boldsymbol{X}_s]$ is the $n \times q$ design matrix ($q = q_o + q_a + q_s$) and $\boldsymbol{B} = [\boldsymbol{B}_o', \boldsymbol{B}_a', \boldsymbol{B}_s']'$. Define also $\boldsymbol{b}_i = \text{vec}\boldsymbol{B}_i$ as the vectorization of the coefficients matrix $\boldsymbol{B}_i$, and $\boldsymbol{b} = [\boldsymbol{b}_o', \boldsymbol{b}_a', \boldsymbol{b}_s']'$.

For a given set of fixed knot locations, the model in (1) is linear in the regression coefficients $\boldsymbol{B}$. As explained in the Introduction, the great challenge with spline models is the choice of knot locations. This is especially true in the surface case where the curse of dimensionality makes it really hard to distribute the multi-dimensional knots in $\mathbb{R}^{q_o}$ in an effective way. To get a fair coverage of knots in the covariate space, a recommended approach is to place the knots at the cluster centers from some clustering algorithm, e.g. k-means clustering or using a mixture of multivariate normals, see Smith & Kohn (1996) and Denison et al. (1998). This typically leads to many redundant knots (since the response variables are

not used to aid the clustering) which is a source of overfitting. One solution is to remove (downweight) the knots by Bayesian variable selection (Smith & Kohn 1996), possibly in a RJMCMC approach, see e.g. Dimatteo et al. (2001) and Denison et al. (2002). Nevertheless, using a set of pre-determined knots is unlikely to work well in the surface case with more than a handful of regressors.

We will treat the knot locations in $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_s$ as unknown parameters to be estimated. This is in principle straightforward from a Bayesian point of view, but great care is needed in the actual implementation of the posterior computations. We propose an efficient MCMC scheme for sampling from the joint posterior of the all knot locations and the regression coefficients, see Section 3 for details. The model is clearly highly (over)parametrized and in need of some regularization of the parameters. The two main regularization techniques in Bayesian analysis are shrinkage priors and variable (knot) selection priors. Variable selection can in principle be incorporated in the analysis, but would be computationally demanding since the number of gradient evaluations needed in our MCMC algorithm would increase dramatically. This is important since evaluating the gradient with respect to the knots is time-consuming as the knot locations enter the likelihood in a very complicated nonlinear way; see Section 3.2 for details. Moreover, part of the attraction of variable selection is that they also provide interpretable measures of variable importance; this is much less interesting here since the covariates correspond to knot locations, which are not interesting in themselves. We have therefore chosen to achieving parsimony with shrinkage priors that pull the regression coefficients towards zero (or any other reference point if so desired), see Section 2.2 for details. We allow for separate shrinkage parameters for the linear, additive and surface parts of the model, and separate shrinkage parameters for the p responses within each of the three model parts. The shrinkage parameters are treated as unknowns and estimated, so that, for example, the surface part can be shrunk towards zero if this agrees with the data. Allowing the knots to move freely in covariate space introduces a knot switching problem similar to the well-known label switching problem in mixture models. The likelihood is invariant to a switch of two knot locations and their regression coefficients. This lack of identification is not important if our aim is to model the regression surface $\mathrm{E}(\boldsymbol{y}|\boldsymbol{x})$, without regard to the posterior of the individual knot locations (Geweke 2007). Also, the MCMC draws of the knot locations can also be used to construct heat maps in covariate space to represent the density

of knots in a certain regions. Such heat maps are clearly also immune to the knot switching problem.

2.2 The prior

We now introduce an easily specified shrinkage prior for the three sets of regression coefficients $\boldsymbol{B}_o$, $\boldsymbol{B}_a$ and $\boldsymbol{B}_s$ and the covariance matrix $\boldsymbol{\Sigma}$, conditional on the knots. The prior for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ are set as

$$\begin{aligned} \text{vec}\boldsymbol{B}_i|\boldsymbol{\Sigma},\ \boldsymbol{\lambda}_i &\sim \text{N}\left(\boldsymbol{\mu}_i,\ \boldsymbol{\Lambda}_i^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_i^{1/2} \otimes \boldsymbol{P}_i^{-1}\right),\ i \in \{o, a, s\}, \\ \boldsymbol{\Sigma} &\sim \text{IW}\left(n_0\boldsymbol{S}_0,\ n_0\right), \end{aligned}$$

with prior independence between the $\boldsymbol{B}_i$. The prior mean of $\text{vec}\boldsymbol{B}_i$ is $\boldsymbol{\mu}_i$, which we set to zero in our shrinkage prior. $\boldsymbol{\Lambda}_i = \text{diag}(\boldsymbol{\lambda}_i) = \text{diag}(\lambda_{i,1}, ..., \lambda_{i,p})$, $\boldsymbol{P}_i$ is a positive definite symmetric matrix. IW($\cdot$) denotes the inverse Wishart distribution, with location matrix $\boldsymbol{S}_0$ and degrees of freedom n_0. $\boldsymbol{P}_i$ is typically either the identity matrix or $\boldsymbol{P}_i = \boldsymbol{X}_i'\boldsymbol{X}_i$. The latter choice has been termed a g-prior by Zellner (1986) and has the advantage of automatically adjusting for the different scales of the covariates. Setting $\lambda_i = n$ makes the information content of the prior equivalent to a single data point and is usually called the unit information prior. The choice of $\boldsymbol{P}_i = \boldsymbol{I}_{q_i}$ can prevent the design matrix from falling into singularity problem when some of the basis functions are highly correlated, which can easily happen with many spline knots. See also the discussion in Denison et al. (2002). Our default choice is therefore $\boldsymbol{P}_o = \boldsymbol{X}_o'\boldsymbol{X}_o$, $\boldsymbol{P}_a = \boldsymbol{I}_{q_a}$ and $\boldsymbol{P}_s = \boldsymbol{I}_{q_s}$. Other shrinkage priors on the regression coefficients can be used in our approach, for example the Laplace distribution leading to the popular Lasso (Tibshirani 1996), but they will typically not allow us to integrate out the regression coefficents analytically, see Section 3.1. The optimal choice of shrinkage prior depends on the unknown data generating model (a normal prior is better when all coefficients have roughly the same magnitude; Lasso is better when many coefficients are close to zero, but some are really large etc).

We also estimate the shrinkage parameters, $\boldsymbol{\lambda}_o$, $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_s$ via a Bayesian approach. Note that our prior constructions for $\boldsymbol{B}$ allow for separate shrinkage of the linear, additive and surface components. This gives us automatic regularization/shrinkage of the regression

coefficients and helps to avoid problems with overfitting. Our MCMC scheme in Section 3 allows for a user-specified prior on λ_{ij}, for $i \in \{o, a, s\}$ and $j = 1, 2, ..., p$ of essentially any functional form. However the default prior of λ_{ij} in this paper follows a log normal distribution with mean of $n/2$ and standard deviation of $n/2$ in order to ensure that both tight and flat shrinkages are attainable within one standard deviation in the prior. For computational convenience, we use a log link for λ_{ij} and make inference on $\log(\lambda_{ij})$. As a result the preceding prior on λ_{ij} yields a normal prior for $\log(\lambda_{ij})$ with mean $[\log(n) - 3/2 \cdot \log(2)]$ and variance $\log(2)$.

We use the same number of additive knots for each covariate in the simulations and the application in Section 4, but it should be clear that our approach also permits unequal number of knots in the different covariates. There is no particular requirements for the prior on the knots, but a vague prior should permit the knots to move freely in covariate space. Our default prior assumes independent knot locations following a normal distribution. The mean of the knots comes from the centers of a k-means clustering of the covariates. In the additive case, the prior variance of all the knots in the kth covariate is $c^2(\boldsymbol{a}'\boldsymbol{a})^{-1}$, where $\boldsymbol{a}$ is the kth column of $\boldsymbol{X}_o$. Similarly, the prior covariance matrix of a surface knot is $c^2(\boldsymbol{X}_o'\boldsymbol{X}_o)^{-1}$. We use $c^2 = n$ as the default setting.

The hyperparameter $\boldsymbol{S}_0$ in the IW prior for $\boldsymbol{\Sigma}$ is set equal to the estimated error covariance matrix from the fitted linear model $\hat{\boldsymbol{Y}} = \boldsymbol{X}_o\hat{\boldsymbol{B}}_o$. A small degrees of freedom (n_0) gives diffuse prior on $\boldsymbol{\Sigma}$ and $n_0 = 10$ is set as the default.

For notational convenience and further computational implementation, we write the prior for the regression coefficients in condensed form as $\boldsymbol{b}|\boldsymbol{\Sigma}, \boldsymbol{\lambda} \sim \mathrm{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_{\boldsymbol{b}})$ where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_o', \boldsymbol{\lambda}_a', \boldsymbol{\lambda}_s')'$, $\boldsymbol{\mu}^* = (\boldsymbol{\mu}_o', \boldsymbol{\mu}_a', \boldsymbol{\mu}_s')'$, $\boldsymbol{\Sigma}_{\boldsymbol{b}} = (\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Sigma}_K\boldsymbol{\Lambda}^{1/2}) \circledast \boldsymbol{P}^{-1}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\Sigma}_K$ is a three-block diagonal matrix with $\boldsymbol{\Sigma}$ on each block, $\boldsymbol{P} = \mathrm{diag}(\boldsymbol{P}_o, \boldsymbol{P}_a, \boldsymbol{P}_s)$ is a block diagonal matrix and $\boldsymbol{A} \circledast \boldsymbol{C}$ denotes the Khatri-Rao product (Khatri & Rao 1968) which is Kronecker product of the corresponding blocks of matrices $\boldsymbol{A}$ and $\boldsymbol{C}$. It will also be convenient to define $\boldsymbol{\beta} = \mathrm{vec}\boldsymbol{B}$. Note that $\boldsymbol{b}$ and $\boldsymbol{\beta}$ contain the same elements with two different stacking orders. As a result, $\boldsymbol{\beta}|\boldsymbol{\Sigma}, \boldsymbol{\lambda} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ essentially have the same entries as $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ have, respectively.

3 The posterior inference

3.1 The posterior

The posterior distribution can be decomposed as

$$p(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\xi}, \boldsymbol{\lambda} | \boldsymbol{Y}, \boldsymbol{X}) = p(\boldsymbol{B} | \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X}) p(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \boldsymbol{X}),$$

where

$$\text{vec}\boldsymbol{B} | \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X} \sim \text{N}(\tilde{\boldsymbol{\beta}},\ \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}),$$

$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} = [\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}]^{-1}$, $\tilde{\boldsymbol{\beta}} = \text{vec}\tilde{\boldsymbol{B}} = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}[\text{vec}(\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}]$ (Zellner 1971), and

$$\begin{aligned} p\left(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \boldsymbol{X}\right) = \ & c \times p(\boldsymbol{\xi}, \boldsymbol{\lambda}) \times |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2} |\boldsymbol{\Sigma}|^{-(n+n_0+p+1)/2} |\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}|^{-1/2} \\ & \times \exp\left\{ -\frac{1}{2} \left[\text{tr}\boldsymbol{\Sigma}^{-1} \left(n_0 \boldsymbol{S}_0 + n\tilde{\boldsymbol{S}} \right) + \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu} \right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu} \right) \right] \right\} \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{S}} = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{B}})'(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{B}})/n$, $c = 2^{-(n_0+n+q)p/2} \pi^{-p(n+q)/2} \Gamma_p^{-1}(n_0/2) |n_0 \boldsymbol{S}_{\boldsymbol{0}}|^{n_0/2}$, $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left[a + (1-j)/2\right]$ is the multivariate gamma function. It is important to note that it is in general not possible to integrate out $\boldsymbol{\Sigma}$ analytically in our model. This is a consequence of using different shrinkage factors for the different responses and on the original, additive and surface parts of the model (the prior covariance matrix of $\boldsymbol{B}$ does not have a Kronecker structure). Only in the special case with a univariate response ($p = 1$) can we integrate out $\boldsymbol{\Sigma}$ analytically, since $\boldsymbol{\Sigma}$ is then a scalar. To obtain a uniform treatment of the models and their gradients, we have chosen to not integrate out $\boldsymbol{\Sigma}$ even for the case $p = 1$. The next subsection proposes an MCMC algorithm for sampling from the joint posterior distribution of all parameters.

3.2 The MCMC algorithm

Our approach is to sample from $p\left(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \boldsymbol{X}\right)$ using a three-block Gibbs sampling algorithm with Metropolis-Hastings (MH) updating steps. Draws from $p(\boldsymbol{B} | \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$ can subsequently be obtained by direct simulation. The updating steps of the Gibbs sampling algorithm are:

1. Simulate $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma}|\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{X})$.

2. Simulate $\boldsymbol{\xi}$ from $p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$.

3. Simulate $\boldsymbol{\lambda}$ from $p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$.

In the special case when $p = 1$

$$\boldsymbol{\Sigma}|\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{X} \sim \text{IW}\left(n_0\boldsymbol{S}_0 + n\tilde{\boldsymbol{S}} + \sum\nolimits_{i\in\{o,a,s\}} \boldsymbol{\Lambda}_i^{-1/2}(\tilde{\boldsymbol{B}}_i - \boldsymbol{M}_i)'\boldsymbol{P}_i(\tilde{\boldsymbol{B}}_i - \boldsymbol{M}_i)\boldsymbol{\Lambda}_i^{-1/2},\ n_0 + n\right) \tag{4}$$

where $\boldsymbol{M}_i$ and $\tilde{\boldsymbol{B}}_i$ are the prior and posterior mean of $\boldsymbol{B}_i$, respectively. Actually, when $p = 1$, $\boldsymbol{\Sigma}$ is a scalar and the IW density reduces to a scaled χ^2 distribution. When $p > 1$, $p(\boldsymbol{\Sigma}|\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{X})$ is no longer IW, but the distribution in (4) is an excellent approximation and can be used as a very efficient MH proposal density.

The conditional posterior distributions for $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ in Steps (2) and (3) above are highly non-standard and we update these parameters using Metropolis-Hastings steps with a tailored proposal, which we now describe for a general parameter vector $\boldsymbol{\theta}$ with posterior $p(\boldsymbol{\theta}|\boldsymbol{Y})$, which could be a conditional posterior in a Metropolis-within-Gibbs algorithm (e.g. $p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$). This method was originally proposed by Gamerman (1997) and later extended by Nott & Leonte (2004) and Villani et al. (2012). All of these three articles are confined to a generalized linear model (GLM) or GLM-like context where the parameters enter the likelihood function through a scalar-valued link function. A contribution of our paper is to show that the algorithm can be extended to models without such a nice structure and that it retains its efficiency even when the parameters are high-dimensional and enter the model in a highly nonlinear way. The way the knot locations and the shrinkage parameters are buried deep in the marginal posterior (see Equation 3.1 above) makes the necessary gradients (see below) much more involved and numerically challenging (see Appendix A).

At any given MCMC iteration we use Newton's method to iterate R steps from the current point $\boldsymbol{\theta}_c$ in the MCMC sampling towards the mode of $p(\boldsymbol{\theta}|\boldsymbol{Y})$, to obtain $\hat{\boldsymbol{\theta}}$ and the Hessian at $\hat{\boldsymbol{\theta}}$. Note that $\hat{\boldsymbol{\theta}}$ may not be the mode but is typically close to it already after a few Newton iterations since the previously accepted $\boldsymbol{\theta}$ is used as the initial value; setting $R = 1, 2$ or 3 is therefore usually sufficient. This makes the algorithm very fast. Having obtained good approximations of the posterior mode and covariance matrix from the Newton iterations, the

proposal $\boldsymbol{\theta}_p$ is now drawn from the multivariate $\boldsymbol{t}$-distribution with $\nu > 2$ degrees of freedom:

$$\boldsymbol{\theta}_p|\boldsymbol{\theta}_c \sim \boldsymbol{t}\left[\hat{\boldsymbol{\theta}},\ -\left(\frac{\partial^2 \ln p(\boldsymbol{\theta}|\boldsymbol{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)^{-1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},\ \nu\right],$$

where the second argument of the density is the covariance matrix and $\hat{\boldsymbol{\theta}}$ is the terminal point of the R Newton steps. The Metropolis-Hastings acceptance probability is

$$a\left(\boldsymbol{\theta}_c \to \boldsymbol{\theta}_p\right) = \min\left[1,\ \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_p)p(\boldsymbol{\theta}_p)g(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}{p(\boldsymbol{Y}|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c)g(\boldsymbol{\theta}_p|\boldsymbol{\theta}_c)}\right].$$

The proposal density at the current point $g(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)$ is a multivariate $\boldsymbol{t}$-density with mode $\tilde{\boldsymbol{\theta}}$ and covariance matrix equal to the negative inverse Hessian evaluated at $\tilde{\boldsymbol{\theta}}$, where $\tilde{\boldsymbol{\theta}}$ is the point obtained by iterating R steps with the Newton algorithm, *this time starting from* $\boldsymbol{\theta}_p$. The need to iterate backwards from $\boldsymbol{\theta}_p$ is clearly important to fulfill the reversibility of the Metropolis-Hastings algorithm. When the number of parameters in $\boldsymbol{\theta}$ is large one can successively apply the algorithm to smaller blocks of parameters in $\boldsymbol{\theta}$.

The tailored proposal distribution turns out to be hugely beneficial for MCMC efficiency, but a naive implementation can easily make the gradient and Hessian evaluations an insurmountable bottleneck in the computations, and a source of numerical instability. We have found the outer product of gradients approximation of the Hessian to work very well, so all we need to implement efficiently are the gradient vector of $p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$ and $p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X})$. The appendix gives compact analytical expression for these two gradient vectors, and shows how to exploit sparsity to obtain fast and stable gradient evaluations. Our gradient evaluations can easily be orders of magnitudes faster than state-of-the-art numerical derivatives, and substantially more stable numerically. For example, already in a relatively small-dimensional model with only four covariates, 20 surface knots and 4 additive knots, the analytical gradient for the knot parameters are more than 40 times faster compared to a numerical gradient with tolerance of 10^{-3}. Since the gradient evaluations accounts for 70-90% of total computing time, this is clearly an important advantage.

3.3 Model comparison

The number of knots is determined via the D-fold out-of-sample log predictive density score (LPDS), defined as

$$\frac{1}{D}\sum_{d=1}^{D}\ln p(\tilde{\boldsymbol{Y}}_d|\tilde{\boldsymbol{Y}}_{-d},\boldsymbol{X}),$$

where $\tilde{\boldsymbol{Y}}_d$ is an $(n_d \times p)$-dimensional matrix containing the n_d observations in the dth testing sample and $\tilde{\boldsymbol{Y}}_{-d}$ denotes the training observations used for estimation. If we assume that the observations are independent conditional on $\boldsymbol{\theta}$, then

$$p(\tilde{\boldsymbol{Y}}_d|\tilde{\boldsymbol{Y}}_{-d},\boldsymbol{X}) = \int\prod_{i\in\tau_d} p(\boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{x}_i)p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})\mathrm{d}\boldsymbol{\theta},$$

where τ_d is the index set for the observations in $\tilde{\boldsymbol{Y}}_d$, and the LPDS is easily computed by averaging $\prod_{i\in\tau_d} p(\boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{x}_i)$ over the posterior draws from $p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})$. This requires sampling from each of the D posteriors $p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})$ for $d = 1, ..., D$, but these MCMC runs can all be run in isolation from each other and are therefore ideal for straightforward parallel computing on widely available multi-core processors. The main advantage for choosing LPDS instead of the marginal likelihood is that the LPDS is not nearly as sensitive to the choice of prior as the marginal likelihood, see e.g. Kass (1993) and Richardson & Green (1997) for a general discussion. The marginal likelihood can also lead to poor predictive inference when the true data generating process is not included in the class of compared models, see e.g. Geweke & Amisano (2011) for an illuminating perspective. The main disadvantage of using the LPDS for selecting the number of knots is that, unlike the marginal likelihood and RJMCMC, there is no rigorous way of including the uncertainty regarding the number of knots in the final inferences. The dataset is systematically partitioned into five folds in our firm leverage application.

4 Simulations

As discussed in the Introduction, the most commonly used approach for spline regression modeling is to use a large number of fixed knots and to use shrinkage priors or Bayesian variable selection to avoid overfitting (Denison et al. 2002). We compare the performance of the

traditional fixed knots approach to our approach with freely estimated knot locations using simulated data with different number of covariates and for varying degrees of nonlinearity in the true surface. We use shrinkage priors with estimated shrinkage both for the fixed and free knot models, but no variable selection. Models with univariate and multivariate response variables are both investigated.

4.1 Simulation setup

We consider data generating processes (DGP) with both univariate ($p = 1$) and bivariate ($p = 2$) responses, and datasets with $q_o = 10$ regressors and two sample sizes, $n = 200$ and $n = 1000$. We first generate the covariate matrix $\boldsymbol{X}_o$ from a mixture of multivariate normals with five components. The weight for the rth mixture component is $u_r / \sum_{l=1}^{5} u_l$, where $u_1, ..., u_5$ are independent $\mathrm{U}(0, 1)$ variables. The mean of each component is a draw from $\mathrm{U}(-1, 1)$ and the components' variances are all 0.1. We randomly select five observations without replacement from $\boldsymbol{X}_o$ as the true surface knots $\boldsymbol{\xi}_s$, and then create the basis expanded design matrix $\boldsymbol{X}$ using the thin-plate radial basis surface spline, see Section 2.1. The coefficients matrix $\boldsymbol{B}$ is generated by repeating the sequence $\{-1, 1\}$. The error term $\boldsymbol{E}$ is from multivariate normal distribution with mean zero, variance 0.1 and covariance 0.1. These settings guarantee a reasonable signal-to-noise ratio.

Following Wood et al. (2002), we measure the degrees of nonlinearity (DNL) in the DGP by the distance between the true surface $f(\cdot)$ and the plane $\hat{g}(\cdot)$ fitted by ordinary least squares without any knots in the model, i.e.

$$\mathrm{DNL} = \sqrt{n^{-1} \sum\nolimits_{i=1}^{n} [f(\boldsymbol{x}_i) - \hat{g}(\boldsymbol{x}_i)]^2}. \tag{5}$$

A larger DNL indicates a DGP with stronger nonlinearity.

We generate 100 datasets and for each dataset we fit the fixed knots model with 5, 10, 15, 20, 25 and 50 surface knots, and also the free knots model with 5, 10, and 15 surface knots. All fitted models have only linear and surface components. The knot locations are determined by k-means clustering. We compare the models with respect to the mean squared loss

$$\mathrm{Loss}(q_s) = \frac{1}{n^*} \sum\nolimits_{i=1}^{n^*} [f(\boldsymbol{x}_i) - \tilde{f}(\boldsymbol{x}_i)]^2 \tag{6}$$

where $f(\cdot)$ is the true surface and $\tilde{f}(\cdot)$ is the posterior mean surface of a given model with q_s surface knots. The Loss in (6) is evaluated over a new sample of n^* covariate vectors, and it therefore measures out-of-sample performance of the posterior mean surface. We will here set $n^* = n$. Note that the shrinkages and the covariance matrix of the error terms are also estimated in both the fixed and free knots models.

4.2 Results

We present the results for $p = 2$ and $n = 200$. The results for $p = 1$ and $n \in \{200, 1000\}$, and $p = 2$ and $n = 1000$ are qualitatively similar and are available upon request. The Supporting Information documents the results for $p = 2$ and $n = 1000$ for a few different model configurations. Figure 1 displays boxplots for the log ratio of the mean squared loss in (6). The columns of the figure represents varying degrees of nonlinearity in the generated datasets according to the estimated DNL measure in equation (5). Each boxplot shows the relative performance of a fixed knots model with a certain number of knots compared to the free knots model with 5 (top row), 10 (middle row) and 15 (bottom row) surface knots, respectively. The short summary of Figure 1 is that the free knots model outperforms the fixed knots model in the large majority of the datasets. This is particularly true when the data are strongly nonlinear. The performance of the fixed knots model improves somewhat when we add more knots, but the improvement is not dramatic. Having more fixed knots clearly improves the chances of having knots close to the true ones, but more knots also increase the risk of overfitting.

The aggregate results in Figure 1 do not clearly show how strikingly different the fixed and free knots models can perform on a given dataset. We will now show that models with free rather than fixed knots are much more robust across different datasets. Figure 2 displays the Euclidean distance of the multivariate *out-of-sample* predictive residuals $\sqrt{\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}$ for a few selected datasets as a function of the distance between the covariate vector and the sample mean of the covariates. The normed residuals depicted in the leftmost column are from datasets chosen with respect to the ranking of the out-of-sample performance of the fixed knots model. For example, the upper left subplot shows the predictive residuals of both the model with 15 fixed knots (vertical bars above the zero line) and the model with 5 free knots (vertical bars below the zero line) on one of the datasets where the fixed knot models

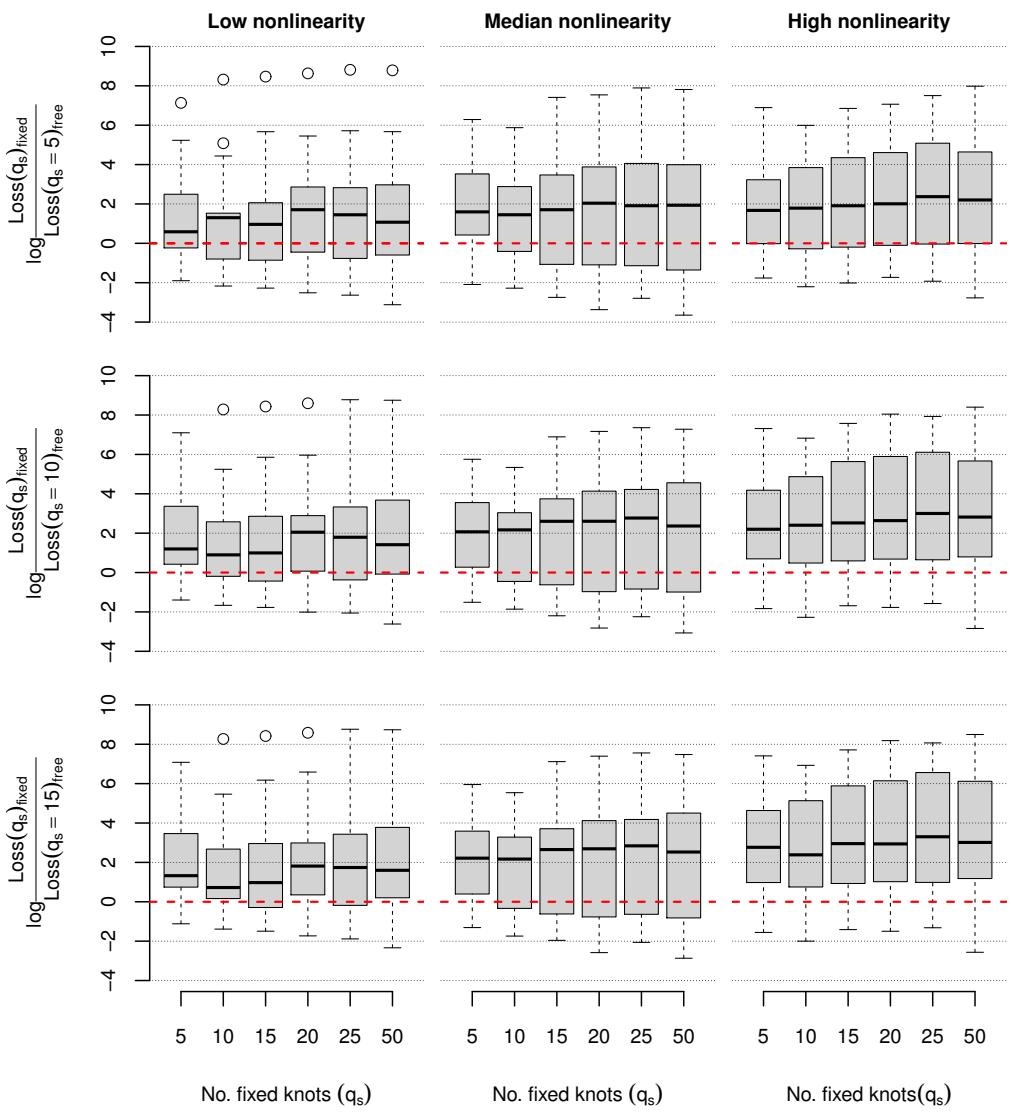

Figure 1: Boxplot of the log loss ratio comparing the performance of the fixed knots model with the free knots model for the DGP with $p = 2$ and $n = 200$. The three columns of the figure correspond to different degrees of nonlinearity of the realized datasets, as measured by estimated DNL in (5).

Table 1: Elapsed computing time (in minutes) for 5,000 iterations with a single dataset of 10 covariates.

	$n = 200$		$n = 1000$	
No. of free surface knots	$p = 1$	$p = 2$	$p = 1$	$p = 2$
2	9	9	16	17
5	13	14	23	26
10	17	18	42	45
15	24	27	61	75

outperform the free knots model by largest margin (3rd best Loss in favor of fixed knots model). It is seen from this subplot that even in this very favorably situation for the fixed knots model, the free knots model is not generating much larger predictive residuals. Moving down to the last row in the left hand column of Figure 2, we see the performance of the two models when the fixed knots model performs very poorly (3rd worse Loss with respect to the fixed knots model). On this particular dataset, the free knots model does well while the fixed knots model is a complete disaster (note the different scales on the vertical axes of the subplots). The column to the right in Figure 2 shows the same analysis, but this time the datasets are chosen with respect to the ranking of the Loss of the free knots model. Overall, Figure 2 clearly illustrates the superior robustness of models with free knots: the free knots model never does much worse than the fixed knots model, but using fixed rather than free knots can lead to a dramatically inferior predictive performance on individual datasets.

4.3 Computing time

The program is written in native R code and all the simulations were performed on a Linux desktop with 2.8 GHz CPU and 4 GB RAM on single instance (without parallel computing). Table 1 shows the computing time in minutes for a single dataset. In general the computing time increases as the size of the design matrix increases, but it increases only marginally as we go from $p = 1$ to $p = 2$.

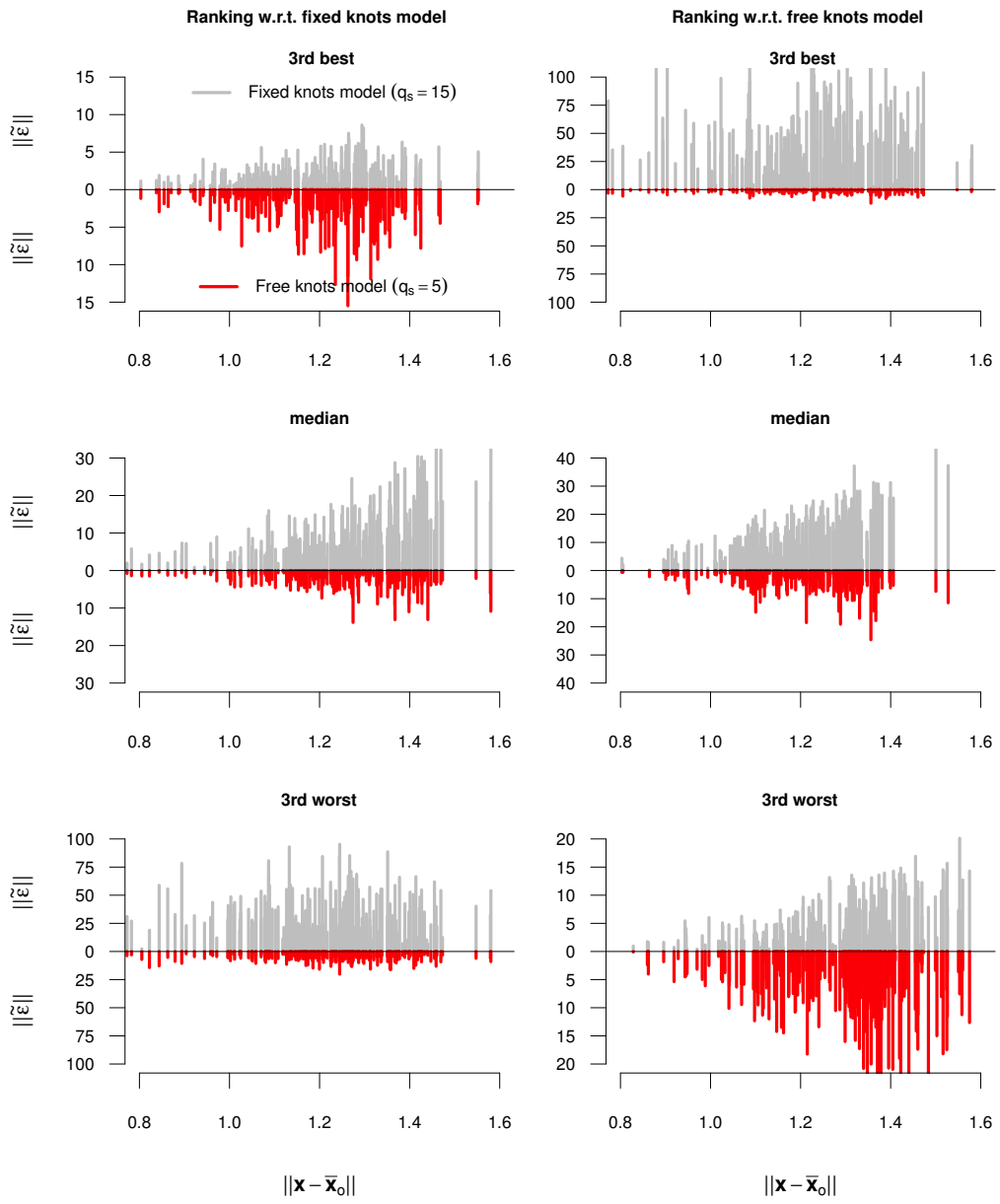

Figure 2: Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 200$. The lines in each subplot are the normed residuals from the model with 15 fixed surface knots (vertical bars above the zero line), and the model with 5 free knots (vertical bars below the zero line). The column to the left shows the results for three datasets chosen when performance is ranked according to the fixed knots model, and the right column displays the results for three datasets chosen when performance is ranked according to the free knots model.

5 Concluding remarks

We have presented a general Bayesian approach for fitting a flexible surface model for a continuous multivariate response using a radial basis spline with freely estimated knot locations. Our approach uses shrinkage priors to avoid overfitting. The locations of the knots and the shrinkage parameters are treated as unknown parameters and we propose a highly efficient MCMC algorithm for these parameters with the coefficients of the multivariate spline integrated out analytically. An important feature of our algorithm is that all knot locations are sampled jointly using a Metropolis-Hastings proposal density tailored to the conditional posterior, rather than the one-knot-at-a-time random walk proposals used in previous literature. The same applies to the block of shrinkage parameters. Both a simulation study and a real application on firm leverage data show that models with free knots have a better out-of-sample predictive performance than models with fixed knots. Moreover, the free knots model is also more robust in the sense that it performs consistently well across different datasets. We also found that models that mix surface and additive spline basis functions in the same model perform better than models with only one of the two basis types.

Our approach can be directly used with other splines basis functions, other priors, and it is at least in principle straightforward to augment the model with Bayesian variable selection. Also, the assumption of Gaussian error distribution could be easily removed by using a Dirichlet process mixture (DPM) prior. We would still be able to integrate out the regression coefficients if we assume a Gaussian base measure in the DPM, see Leslie et al. (2007) for details in the univariate case.

6 Acknowledgements

The authors are grateful to Paolo Giordani and Robert Kohn for stimulating discussions and constructive suggestions. The authors thank two anonymous referees for the helpful comments that improved the contents and presentation of the paper. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2011229.

A Details of the MCMC algorithm

In this section we briefly address the MCMC details and related computational issues. For details on matrix manipulations and derivatives, see e.g. Lütkepohl (1996). Our MCMC algorithm in Section 3.2 only requires the gradient of the conditional posteriors w.r.t. each parameter. Since users can always use their own prior on the knots and shrinkages, we will not document the gradient of any particular prior. In particular for the normal prior, one can directly find the results in e.g. Mardia et al. (1979). We now present the full gradients for the knot locations and the shrinkage parameters.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and regression trees*, Chapman and Hall/CRC, New York.

Buhmann, M. D. (2003), *Radial basis functions: theory and implementations*, Cambridge University Press, Cambridge.

Chipman, H., George, E. & McCulloch, R. (2010), 'BART: Bayesian additive regression trees', **4**(1), 266–298.

Denison, D., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Jone Wiley & Sons, Chichester.

Denison, D., Mallick, B. & Smith, A. (1998), 'Automatic Bayesian curve fitting', **60**(2), 333–350.

Dimatteo, I., Genovese, C. & Kass, R. (2001), 'Bayesian curve-fitting with free-knot splines', *Biometrika* **88**(4), 1055–1071.

Friedman, J. (1991), 'Multivariate adaptive regression splines', **19**(1), 1–67.

Gamerman, D. (1997), 'Sampling from the posterior distribution in generalized linear mixed models', **7**(1), 57–68.

Gasser, T. & Müller, H. (1979), Kernel estimation of regression functions, *in* T. Gasser & M. Rosenblatt, eds, 'Smoothing Techniques for Curve Estimation', Vol. 757, Springer, New York, pp. 23–68.

Geweke, J. (2007), 'Interpretation and inference in mixture models: Simple MCMC works', **51**(7), 3529–3550.

Geweke, J. & Amisano, G. (2011), 'Optimal prediction pools', **164**(1), 130–141.

Gulam Razul, S., Fitzgerald, W. & Andrieu, C. (2003), 'Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **497**(2-3), 492–510.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman & Hall/CRC, New York.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Holmes, C. & Mallick, B. (2003), 'Generalized nonlinear modeling with multivariate free-knot regression splines', **98**(462), 352–368.

Kass, R. (1993), 'Bayes factors in practice', **42**(5), 551–560.

Khatri, C. & Rao, C. (1968), 'Solutions to some functional equations and their applications to characterization of probability distributions', **30**(2), 167–180.

Leslie, D., Kohn, R. & Nott, D. (2007), 'A general approach to heteroscedastic linear regression', **17**(2), 131–146.

Lütkepohl, H. (1996), *Handbook of matrices*, John Wiley & Sons, Chichester.

Mardia, K., Kent, J., & Bibby, J. (1979), *Multivariate analysis*, Academic Press, London.

Nadaraya, E. A. (1964), 'On estimating regression', **9**, 141–142.

Nott, D. & Leonte, D. (2004), 'Sampling schemes for Bayesian variable selection in generalized linear models', **13**(2), 362–382.

Richardson, S. & Green, P. (1997), 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)', **59**(4), 731–792.

Ruppert, D., Wand, M. & Carroll, R. (2003), *Semiparametric regression*, Cambridge University Press, Cambridge.

Smith, M. & Kohn, R. (1996), 'Nonparametric regression using Bayesian variable selection', **75**(2), 317–343.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', pp. 267–288.

Villani, M., Kohn, R. & Nott, D. J. (2012), 'Generalized smooth finite mixtures', **171**(2), 121–133.

Watson, G. (1964), 'Smooth regression analysis', **26**(4), 359–372.

Wood, S., Jiang, W. & Tanner, M. (2002), 'Bayesian mixture of splines for spatially adaptive nonparametric regression', *Biometrika* **89**(3), 513.

Zellner, A. (1971), *An introduction to Bayesian inference in econometrics*, John Wiley & Sons, New York.

Zellner, A. (1986), 'On assessing prior distributions and Bayesian regression analysis with g-prior distributions', *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* **6**, 233–243.