

EFFICIENT BAYESIAN MULTIVARIATE SURFACE REGRESSION

FENG LI AND MATTIAS VILLANI

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.

1. INTRODUCTION

Flexible models of the regression function $E(y|x)$ 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.

Li (corresponding author): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.
E-mail: feng.li@stat.su.se. Villani: Division of Statistics, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden. E-mail: mattias.villani@liu.se.

The most widely used models assume additivity in the regressors, i.e. $E(y|x_1, \dots, x_q) = \sum_{j=1}^q f_j(x_j)$, where $f_j(x_j)$ is a spline function for the j th 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 (\mathbf{x}) to its q -dimensional knot ($\boldsymbol{\xi}$), e.g. the cubic radial basis $\|\mathbf{x} - \boldsymbol{\xi}\|^3$, where $\mathbf{x} = (x_1, \dots, x_q)'$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_q)'$ 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, see Section 5.4 for some evidence. 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:

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

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

$$\mathbf{x}_{sj}(\boldsymbol{\xi}_{sj}) = \|\mathbf{x}_o - \boldsymbol{\xi}_{sj}\|^2 \ln \|\mathbf{x}_o - \boldsymbol{\xi}_{sj}\|, \quad j = 1, \dots, q_s, \quad (2)$$

where \mathbf{x}_o is one of the original data points and $\boldsymbol{\xi}_{sj}$ is the j th 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

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

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

For a given set of fixed knot locations, the model in (1) is linear in the regression coefficients \mathbf{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 ξ_a and ξ_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 achieve 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 $E(\mathbf{y}|\mathbf{x})$, 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, see Section 5. 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 \mathbf{B}_o , \mathbf{B}_a and \mathbf{B}_s and the covariance matrix Σ , conditional on the knots. The prior for \mathbf{b} and Σ are set as

$$\text{vec} \mathbf{B}_i | \Sigma, \boldsymbol{\lambda}_i \sim N \left(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{1/2} \Sigma \boldsymbol{\Lambda}_i^{1/2} \otimes \mathbf{P}_i^{-1} \right), i \in \{o, a, s\},$$

$$\Sigma \sim \text{IW}(n_0 \mathbf{S}_0, n_0),$$

with prior independence between the \mathbf{B}_i . The prior mean of $\text{vec} \mathbf{B}_i$ is $\boldsymbol{\mu}_i$, which we set to zero in our shrinkage prior. $\boldsymbol{\Lambda}_i = \text{diag}(\boldsymbol{\lambda}_i) = \text{diag}(\lambda_{i,1}, \dots, \lambda_{i,p})$, \mathbf{P}_i is a positive definite symmetric matrix. $\text{IW}(\cdot)$ denotes the inverse Wishart distribution, with location matrix \mathbf{S}_0 and degrees of freedom n_0 . \mathbf{P}_i is typically either the identity matrix or $\mathbf{P}_i = \mathbf{X}'_i \mathbf{X}_i$. 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 $\mathbf{P}_i = \mathbf{I}_{q_i}$ 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 $\mathbf{P}_o = \mathbf{X}'_o \mathbf{X}_o$, $\mathbf{P}_a = \mathbf{I}_{q_a}$ and $\mathbf{P}_s = \mathbf{I}_{q_s}$. 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

coefficients 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 \mathbf{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 $j = 1, 2, \dots, p$ 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 and 5, 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 k th covariate is $c^2(\mathbf{a}'\mathbf{a})^{-1}$, where \mathbf{a} is the k th column of \mathbf{X}_o . Similarly, the prior covariance matrix of a surface knot is $c^2(\mathbf{X}'_o\mathbf{X}_o)^{-1}$. We use $c^2 = n$ as the default setting.

The hyperparameter \mathbf{S}_0 in the IW prior for $\boldsymbol{\Sigma}$ is set equal to the estimated error covariance matrix from the fitted linear model $\hat{\mathbf{Y}} = \mathbf{X}_o \hat{\mathbf{B}}_o$. 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 $\mathbf{b}|\boldsymbol{\Sigma}, \boldsymbol{\lambda} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_b)$ where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_o, \boldsymbol{\lambda}'_a, \boldsymbol{\lambda}'_s)'$, $\boldsymbol{\mu}^* = (\boldsymbol{\mu}'_o, \boldsymbol{\mu}'_a, \boldsymbol{\mu}'_s)'$, $\boldsymbol{\Sigma}_b = (\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Sigma}_K\boldsymbol{\Lambda}^{1/2}) * \mathbf{P}^{-1}$, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\Sigma}_K$ is a three-block diagonal matrix with $\boldsymbol{\Sigma}$ on each block, $\mathbf{P} = \text{diag}(\mathbf{P}_o, \mathbf{P}_a, \mathbf{P}_s)$ is a block diagonal matrix and $\mathbf{A} * \mathbf{C}$ denotes the Khatri-Rao product (Khatri & Rao 1968) which is Kronecker product of the corresponding blocks of matrices \mathbf{A} and \mathbf{C} . It will also be convenient to define $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$.

Note that \boldsymbol{b} and $\boldsymbol{\beta}$ contain the same elements with two different stacking orders. As a result, $\boldsymbol{\beta}|\boldsymbol{\Sigma}, \boldsymbol{\lambda} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ essentially have the same entries as $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ have, respectively (Section A.3).

3. THE POSTERIOR INFERENCE

3.1. The posterior. The posterior distribution can be decomposed as

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

where

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

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

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

where $\tilde{\mathbf{S}} = (\mathbf{Y} - \mathbf{X} \tilde{\boldsymbol{B}})' (\mathbf{Y} - \mathbf{X} \tilde{\boldsymbol{B}}) / n$, $c = 2^{-(n_0+n+q)p/2} \pi^{-p(n+q)/2} \Gamma_p^{-1}(n_0/2) |n_0 \mathbf{S}_0|^{n_0/2}$, $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma[a + (1-j)/2]$ 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 $p(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma} | \mathbf{Y}, \mathbf{X})$ using a three-block Gibbs sampling algorithm with Metropolis-Hastings (MH) updating steps. Draws from $p(\boldsymbol{B} | \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X})$ can subsequently be obtained by direct simulation. The updating steps of the Gibbs sampling algorithm are:

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

In the special case when $p = 1$

$$\Sigma | \xi, \lambda, Y, X \sim \text{IW} \left(n_0 S_0 + n \tilde{S} + \sum_{i \in \{o, a, s\}} \Lambda_i^{-1/2} (\tilde{B}_i - M_i)' P_i (\tilde{B}_i - M_i) \Lambda_i^{-1/2}, n_0 + n \right) \quad (4)$$

where M_i and \tilde{B}_i are the prior and posterior mean of B_i , respectively. Actually, when $p = 1$, Σ is a scalar and the IW density reduces to a scaled χ^2 distribution. When $p > 1$, $p(\Sigma | \xi, \lambda, Y, X)$ 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 ξ and λ 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 θ with posterior $p(\theta | Y)$, which could be a conditional posterior in a Metropolis-within-Gibbs algorithm (e.g. $p(\xi | \lambda, \Sigma, Y, X)$). 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 θ_c in the MCMC sampling towards the mode of $p(\theta | Y)$, to obtain $\hat{\theta}$ and the Hessian at $\hat{\theta}$. Note that $\hat{\theta}$ may not be the mode but is typically close to it already after a few Newton iterations since the previously accepted θ 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 θ_p is now drawn from the multivariate t -distribution with $\nu > 2$ degrees of freedom:

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

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

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

The proposal density at the current point $g(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)$ is a multivariate \mathbf{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, see Section 5.4 for some evidence, 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 $p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X})$ and $p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X})$. Appendix A 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 in Section 5 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

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

where $\tilde{\mathbf{Y}}_d$ is an $(n_d \times p)$ -dimensional matrix containing the n_d observations in the d th testing sample and $\tilde{\mathbf{Y}}_{-d}$ denotes the training observations used for estimation. If we assume that the

observations are independent conditional on $\boldsymbol{\theta}$, then

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

where τ_d is the index set for the observations in $\tilde{\mathbf{Y}}_d$, and the LPDS is easily computed by averaging $\prod_{i \in \tau_d} p(\mathbf{y}_i | \boldsymbol{\theta}, \mathbf{x}_i)$ over the posterior draws from $p(\boldsymbol{\theta} | \tilde{\mathbf{Y}}_{-d})$. This requires sampling from each of the D posteriors $p(\boldsymbol{\theta} | \tilde{\mathbf{Y}}_{-d})$ for $d = 1, \dots, D$, 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 in Section 5.

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 \mathbf{X}_o from a mixture of multivariate normals with five components. The weight for the r th mixture component is $u_r / \sum_{l=1}^5 u_l$, where u_1, \dots, u_5 are independent $U(0, 1)$ variables. The mean of each component

is a draw from $U(-1, 1)$ and the components' variances are all 0.1. We randomly select five observations without replacement from \mathbf{X}_o as the true surface knots $\boldsymbol{\xi}_s$, and then create the basis expanded design matrix \mathbf{X} using the thin-plate radial basis surface spline, see Section 2.1. The coefficients matrix \mathbf{B} is generated by repeating the sequence $\{-1, 1\}$. The error term \mathbf{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.

$$\text{DNL} = \sqrt{n^{-1} \sum_{i=1}^n [f(\mathbf{x}_i) - \hat{g}(\mathbf{x}_i)]^2}. \quad (5)$$

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

$$\text{Loss}(q_s) = \frac{1}{n^*} \sum_{i=1}^{n^*} [f(\mathbf{x}_i) - \tilde{f}(\mathbf{x}_i)]^2 \quad (6)$$

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 $\sqrt{\hat{\epsilon}'\hat{\epsilon}}$ 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 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 $p = 2$.

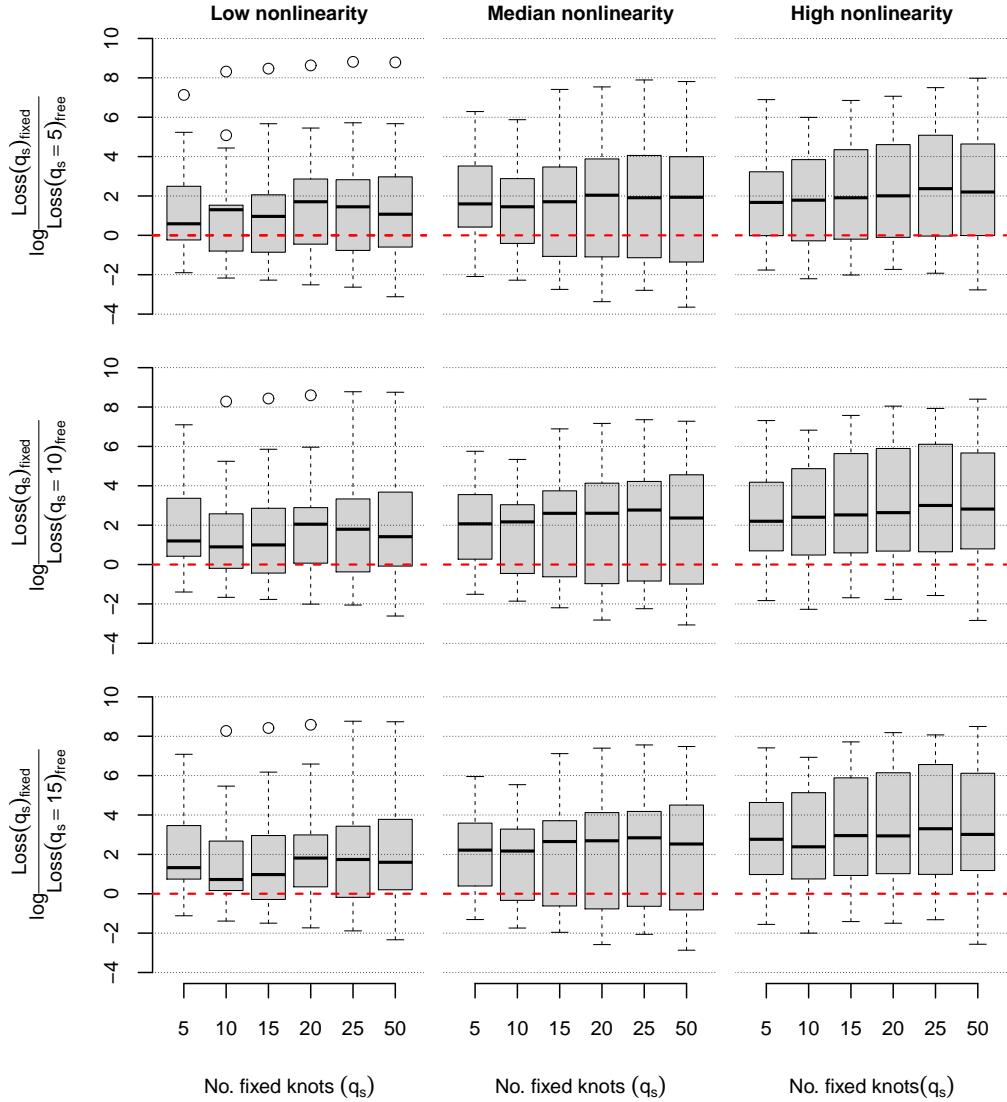


FIGURE 1. Boxplot of the log loss ratio comparing the performance of the fixed knots model with the free knots model for the DGP with $p = 2$ and $n = 200$. The three columns of the figure correspond to different degrees of nonlinearity of the realized datasets, as measured by estimated DNL in (5).

5. APPLICATION TO FIRM CAPITAL STRUCTURE DATA

5.1. The data. The classic paper by Rajan & Zingales (1995) analyze firm leverage ($\text{leverage} = \text{total debt}/(\text{total debt} + \text{book value of equity})$) as a function of its fixed assets ($\text{tang} = \text{tangible assets}/\text{book value of total assets}$), its market-to-book ratio ($\text{market2book} = (\text{book value}$

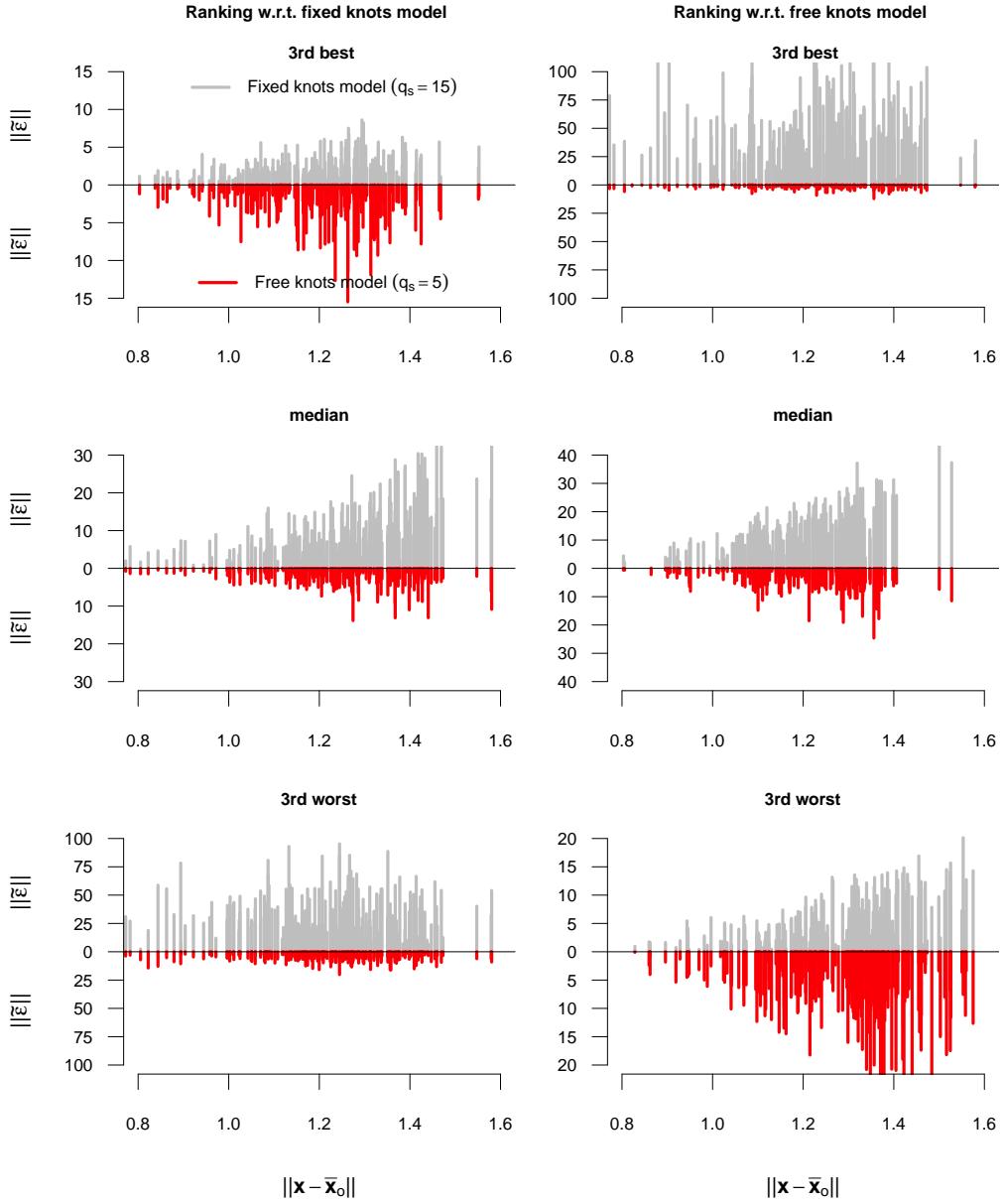


FIGURE 2. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 200$. The lines in each subplot are the normed residuals from the model with 15 fixed surface knots (vertical 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.

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

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

of total assets - book value of equity + market value of equity)/book value of total assets), logarithm of sales (LogSale) and profit (Profit = earnings before interest, taxes, depreciation, and amortization/book value of total assets). Strong nonlinearities seem to be a quite general feature of balance sheet data, but only a handful articles have suggested using nonlinear/nonparametric models, see e.g. Bastos & Ramalho (2010), and Villani et al. (2012). We use a similar data to the one in Rajan & Zingales (1995) which covers 4,405 American non-financial firms with positive sales in 1992 and complete data records and analyze the leverage in terms of total debt. Villani et al. (2012) analyze the same data with a smooth mixture of Beta regressions.

Figure 3 plots the response variable **leverage** in both original scale and logit scale ($\ln[y/(1-y)]$) against each of the four covariates. The relationships between the leverage and the covariates are clearly highly nonlinear even when the logit transformation is used. There are also outliers which can be seen from the subplots with respect to covariates **Market2Book** and **Profit**.

5.2. Models with only surface or additive components. We first fit models that either have only a surface component or only an additive component (both types of models also have a linear component). Note that the shrinkage parameters are also estimated in all cases. All four covariates are used in the estimation procedure and we use the logit transformation of the leverage, and standardize each covariate to have zero mean and unit variance.

Figure 4 depicts the LPDS for the surface component model and the additive component model for both the case of fixed and free knots. The LPDS generally improves as the number of knots increases for both the fixed and free knots models, but seems to eventually level off at large number of knots. The free knots model always outperforms the fixed knots model when only a surface component is used (left subplot). For example, the model with 12 free surface

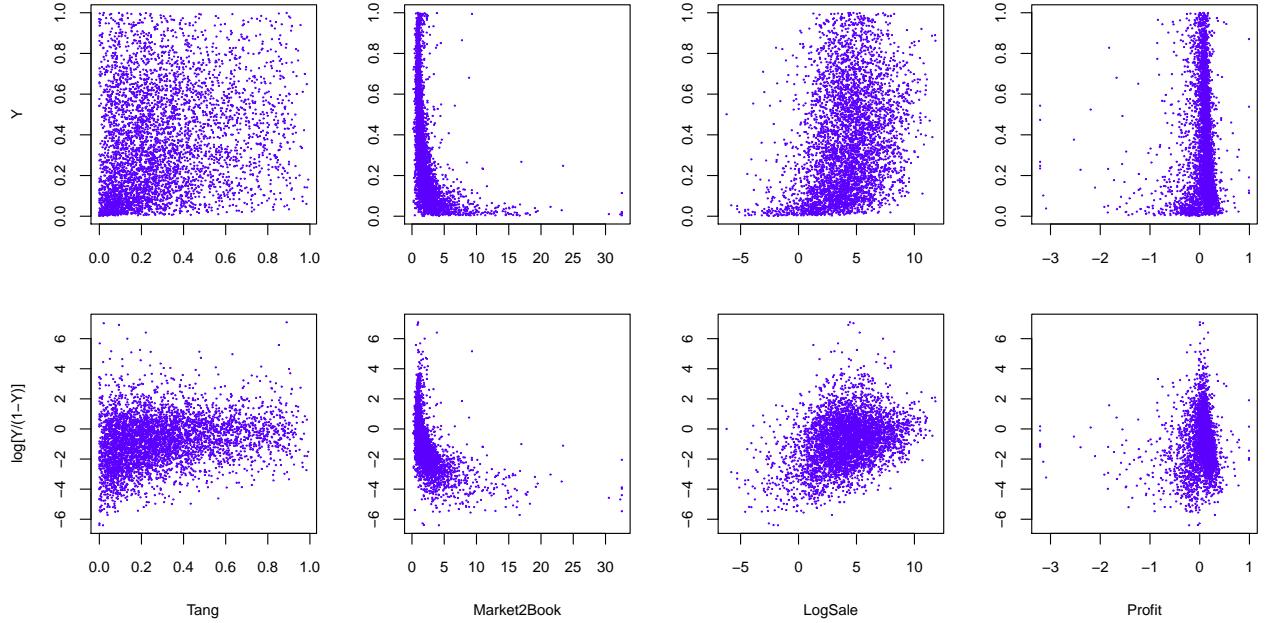


FIGURE 3. Scatter plots of the firm leverage data with leverage (\mathbf{Y}) on both original scale (top subplots) and logit transformed scale (bottom subplots) against each of the four covariates.

knots is roughly 32 LPDS units better than the fixed knots model with the same number of knots. This is a quite impressive improvement in out-of-sample performance considering that the fixed knot locations are chosen with state-of-the-art clustering methods for knot selection. The ability to move the knots clearly also helps to keep the number of knots to a minimum; it takes for example more than 30 fixed surface knots to obtain the same LPDS as a model with 12 free surface knots.

Turning to the strictly additive models in right subplot of Figure 4 we see that the additive models are in general inferior to the models with only surface knots, and that the differences in LPDS between the fixed and free knots approaches are much smaller here, at least for eight knots or more. The improvement in LPDS levels off at roughly 16 knots. It is important to note that the horizontal axis in Figure 4 displays the number of additive knots *in each covariate*, and the fact that we do not overfit bear testimony to the effectiveness of the shrinkage priors.

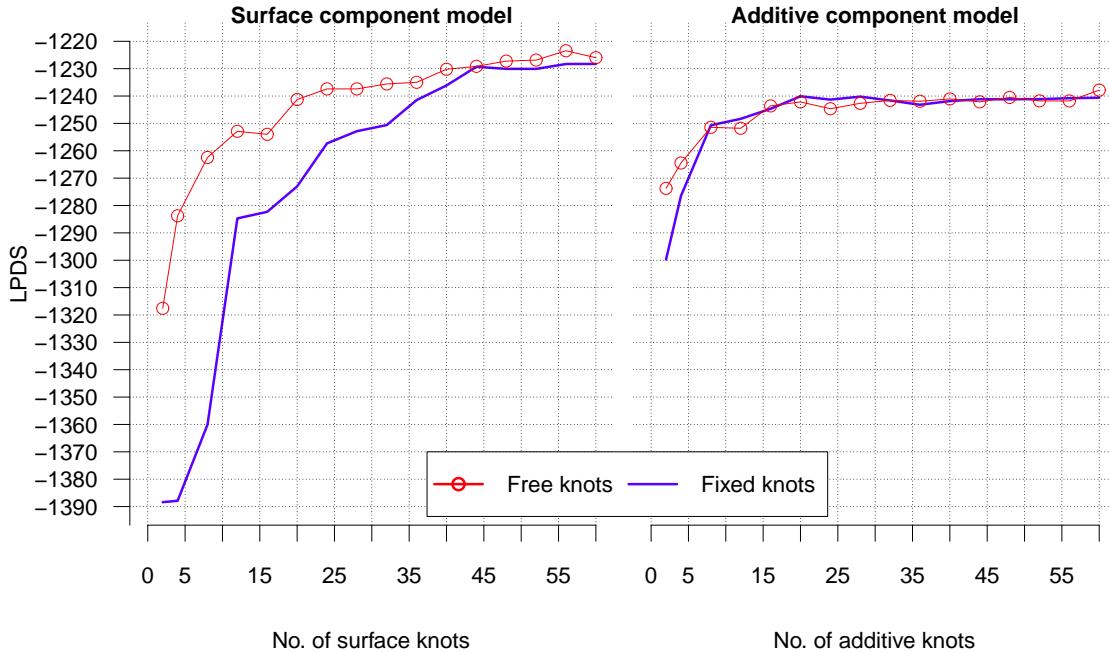


FIGURE 4. LPDS for the firm leverage data with surface component model (left) and additive component model (right). Note that the number of knots in additive model is the number of spline basis functions on each covariate.

5.3. Models with both additive and surface components. We now consider models with both additive and surface components. It is worth mentioning that we draw from the joint posterior distribution of the surface and additive knots, see Section A for MCMC details.

Figure 5 shows that there are generally improvements from using both surface knots and additive knots in the same model. For example, the model with 4 free surface knots has an LPDS of $-1,284$. Adding two free additive knots increases the LPDS to $-1,270$ and adding another two additive knots gives a further increase of 14 LPDS units. Figure 5 also shows strong gains from estimating the knots' locations, but the improvement in LPDS from free knots tends to be less dramatic when more additive knots are used to complement the surface knots. There is little or no improvement in LPDS as the number of surface knots approaches 60. The results in Figure 5 reinforces the evidence in Figure 4 that the shrinkage prior is very effective in mitigating potential problems with overfitting.

To simplify the graphical presentation of the results, we choose to illustrate the posterior inference of the knot locations in a model with only the two covariates `Market2Book` and

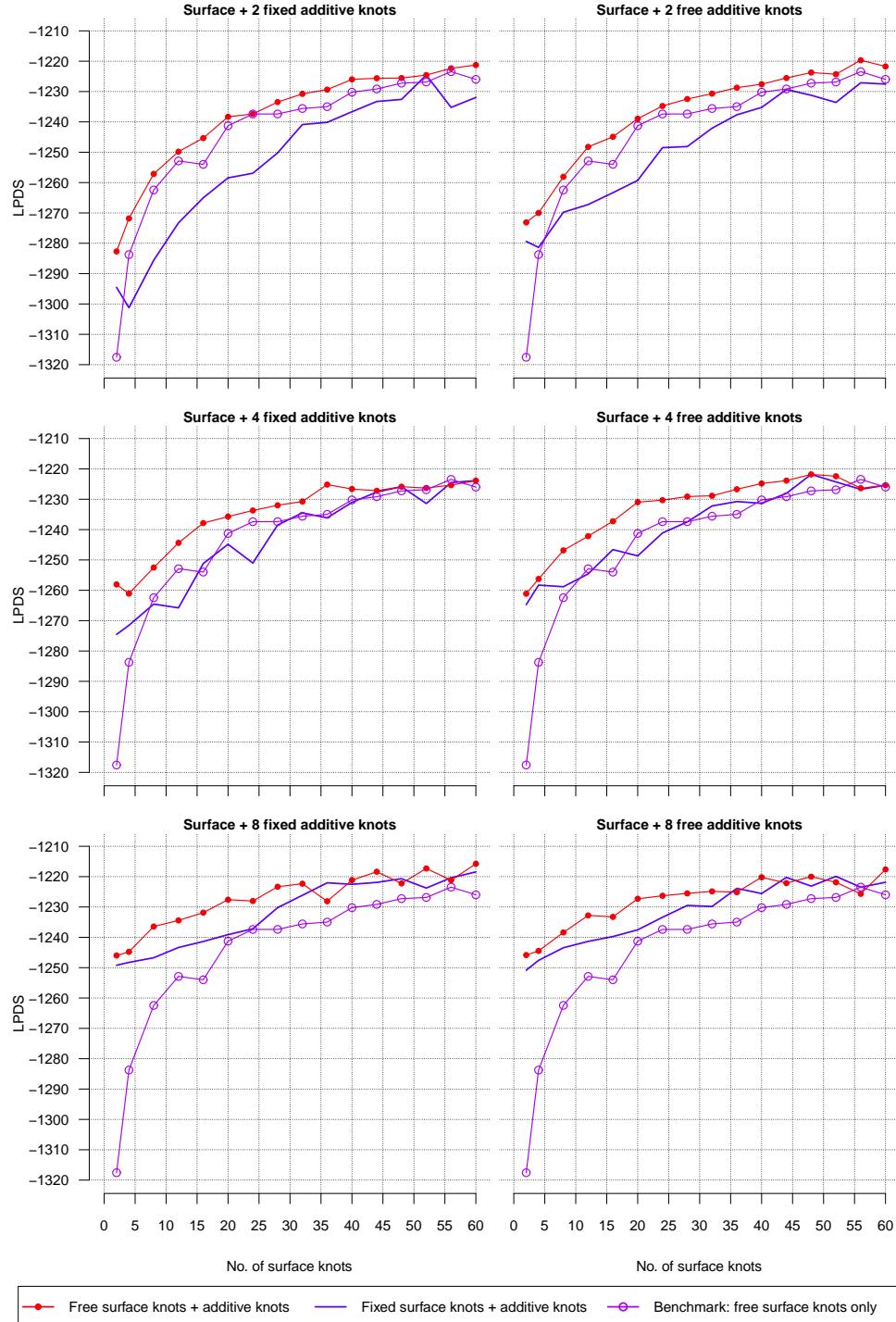


FIGURE 5. LPDS for the firm leverage data for the free and fixed knots models with varying number of surface and additive knots.

Profit. We use 20 surface knots and 4 additive knots in each covariate. The mean acceptance probabilities for the knot locations and the shrinkage parameters in Metropolis-Hastings algorithm are 0.73 and 0.64, respectively, which are exceptionally large considering that all $2 \times 20 + 2 \times 4 = 48$ knot location parameters are proposed jointly, as are all the shrinkage parameters. The acceptance probability in the updating step for Σ is 1 since we are proposing directly from the exact conditional posterior when $p = 1$. Because of the knot switching problem (see Section 3), it does not make much sense to display the posterior distribution of the knot locations directly. We instead choose to partition the covariate space into small rectangular regions, count the frequency of knots in each region over the MCMC iterations, and use heat maps to visualize the density of knots in different regions of covariate space. Figure 6 displays this knot density heat map. As expected, the estimated knot locations are mostly concentrated in the data dense regions, particularly in regions where the relation between the covariates and response in the data is most nonlinear, which is seen by comparing Figure 6 and Figure 3.

Finally, we present the posterior surface for the firm leverage data in Figure 7. To enhance the visual representation, the graphs zoom in on the region with the majority of the data observations. Figure 7 plots the mean (left) and the standard deviation (right) of the posterior surface. The latter object is for brevity sometimes referred to as the *posterior standard deviation surface*. Figure 7 (right) also displays the covariate observations to give a sense of where the data observations are located. The Supporting Information to this article investigates the robustness of the posterior results to variations in both the prior mean and variance of the knot locations. The posterior heat map of the knot locations are affected by the fairly dramatic variations in the prior mean of the knots, and to a lesser extent by changes in the prior variance of the knot locations, but the posterior mean and standard deviation surfaces are robust to variations in the prior on the knots, especially in data dense regions. The Supporting Information also shows that the posterior is robust to changes in the prior on the shrinkage factors.

5.4. MCMC efficiency in the updating of the knot locations. In order to study the efficiency of our algorithm for sampling the knot locations, we compare three types of MCMC updates of the knots: i) one-knot-at-a-time updates using a random walk Metropolis proposal with tuned variance (SRWM), ii) one-knot-at-a-time updates with the tailored Metropolis-Hastings step (SMH) in Section 3.2, and iii) full block updating of all knots using the tailored

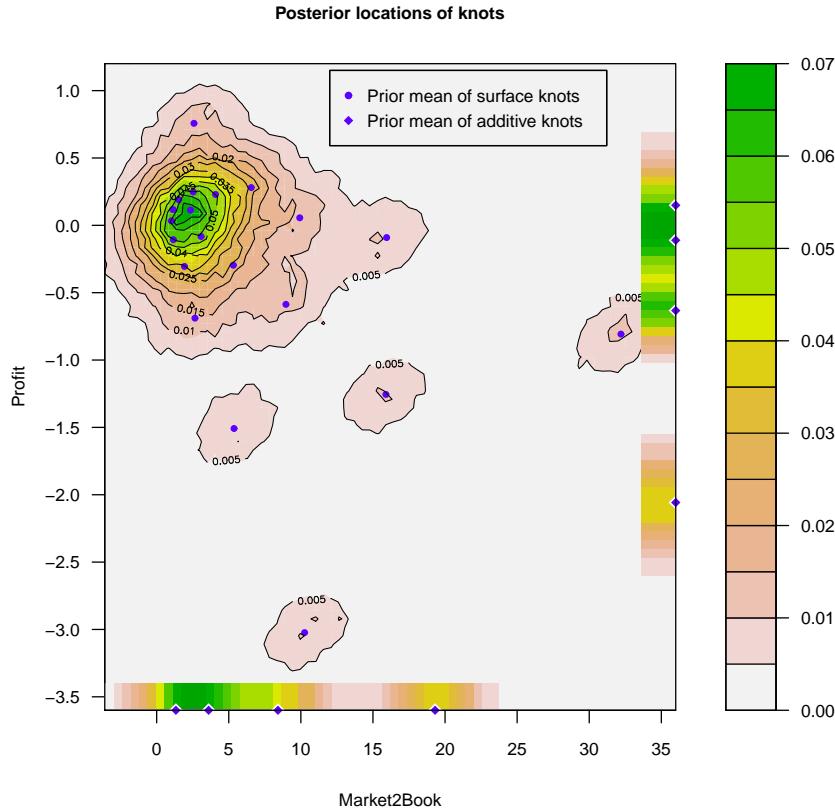


FIGURE 6. Heat map to visualize the posterior density of the knot locations in covariate space for model with 4 free additive knots and 20 free surface knots for the firm leverage dataset. The plot is constructed by partitioning the covariate space into 70×70 rectangular regions and counting the number of surface knots in each rectangle over the MCMC draws. The posterior density of the locations of the additive knots is constructed in a similar fashion and separate heat maps for the additive knots in each covariate are shown just above the horizontal axis and vertical axis, respectively.

Metropolis-Hastings step (BMH) in Section 3.2. SRWM moves are used in state-of-the-art RJMCMC approaches such as Dimatteo et al. (2001) and Gulam Razul et al. (2003). Note that we are not studying the performance of a complete RJMCMC scheme; we are here interested in isolating this particular updating step and comparing it to our tailored proposal. We use the inefficiency factor (IF) (Geweke 1992) to measure the efficiency of MCMC. The IF is a measure of the number of draws needed to obtain the equivalent of a

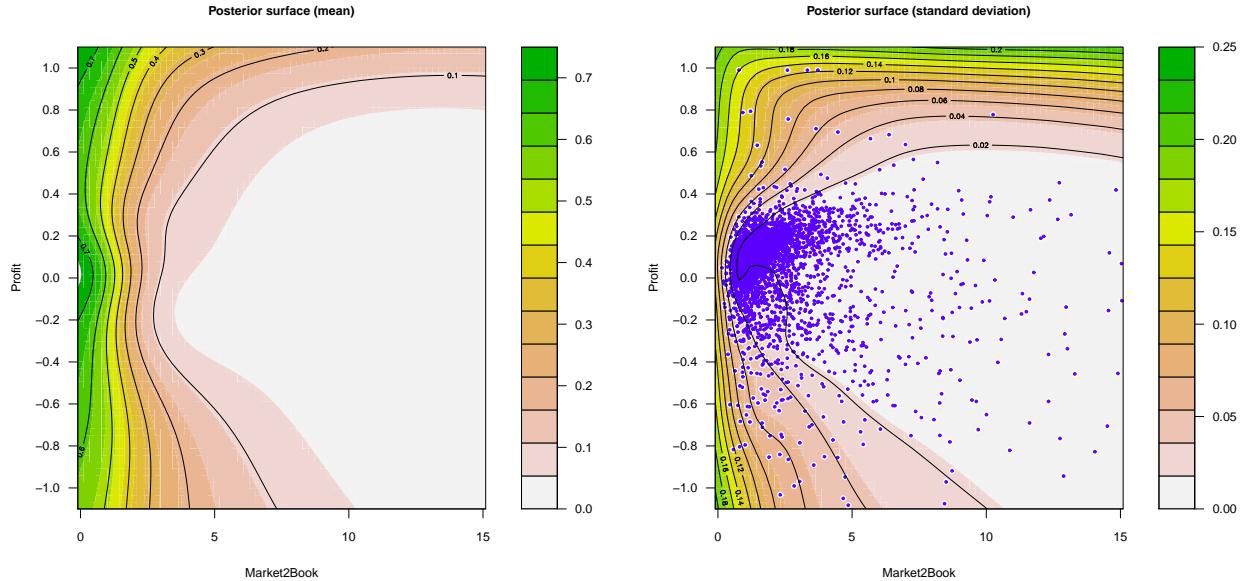


FIGURE 7. The posterior mean (left) and standard deviation (right) of the posterior surface for the model with 4 free additive knots and 20 free surface knots for firm leverage data. The subplot to the right also shows an overlay of the covariate observations.

single independent draw. It is defined as $IF = 1 + 2 \sum_{i=1}^{\infty} \rho_i$ where ρ_i is the autocorrelation of the MCMC trajectory at lag i . We also document the effective sample size per minute, i.e. (number of MCMC draws)/(IF \times computing time) to measure the overall efficiency of the MCMC.

Table 2 shows the efficiency of the three knot sampling algorithms in a model with 20 free surface knots and 4 additive knots in each covariate on the firm leverage data. The inefficiency factor in Table 2 is the average inefficiency of the posterior mean surface in 1000 random chosen points in covariate space. There is some gain from tailoring the proposal for each knot separately, but the really striking observation from Table 2 is the massive efficiency and speed gains from updating all the blocks jointly using a tailored proposal; the effective sample size per minute is roughly 70 times larger when our BMH algorithm is used instead of simple SRWM updates.

TABLE 2. Comparison of algorithms for updating the knot locations in a model with 20 free surface knots and 4 additive knots in each covariate. Firm leverage data.

	SRWM	SMH	BMH
Mean IF for the posterior mean surface	29.63	2.70	1.16
Mean acceptance probability	0.26	0.62	0.88
Computing time (<i>min</i>)	388.21	1716.07	141.72
Effective sample size per minute	0.87	2.16	60.83

6. 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.

7. 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.

APPENDIX 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.

A.1. Gradient w.r.t. the knot locations.

$$\begin{aligned} \frac{\partial \ln p(\boldsymbol{\xi} | \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X})}{\partial \boldsymbol{\xi}'} &= \frac{\partial \log p(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}'} - \frac{p}{2} \sum_{i \in \{o, a, s\}} (\text{vec} \mathbf{P}_i)' \frac{\partial \text{vec} \mathbf{P}_i}{\partial \boldsymbol{\xi}'} \\ &\quad - \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu} \right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\xi}'} - \frac{1}{2} (\text{vec} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}})' \frac{\partial \text{vec} [\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}' \mathbf{X}]}{\partial \boldsymbol{\xi}'} \\ &\quad - \frac{1}{2} (\text{vec} \boldsymbol{\Sigma}^{-1})' (\mathbf{I}_p + \mathbf{K}_{p,p}) \left\{ (\mathbf{I}_p \otimes \tilde{\mathbf{E}}' \mathbf{X}) \frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\xi}'} + (\tilde{\mathbf{B}}' \otimes \tilde{\mathbf{E}}') \frac{\partial \text{vec} \mathbf{X}}{\partial \boldsymbol{\xi}'} \right\} \\ &\quad - \frac{1}{2} \left\{ \text{vec} \left[\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu} \right) \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu} \right)' + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} \right] \right\}' \frac{\partial \text{vec} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\xi}'}, \end{aligned}$$

where $\tilde{\mathbf{E}} = \mathbf{Y} - \mathbf{X} \tilde{\mathbf{B}}$, \mathbf{I}_p is the identity matrix, $\mathbf{K}_{p,p}$ is the commutation matrix and

$$\frac{\partial \text{vec} [\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}' \mathbf{X}]}{\partial \boldsymbol{\xi}'} = (\mathbf{I}_p \otimes \mathbf{K}_{q,p} \otimes \mathbf{I}_q) (\text{vec} \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{q^2}) (\mathbf{I}_{q^2} + \mathbf{K}_{q,q}) (\mathbf{I}_q \otimes \mathbf{X}') \frac{\partial \text{vec} \mathbf{X}}{\partial \boldsymbol{\xi}'},$$

$$\begin{aligned} \frac{\partial \tilde{\beta}}{\partial \xi'} &= \Sigma_{\tilde{\beta}} \left[[\Sigma^{-1} \mathbf{Y}' \otimes \mathbf{I}_q] \mathbf{K}_{n,q} \frac{\partial \text{vec} \mathbf{X}}{\partial \xi'} + (\boldsymbol{\mu}' \otimes \mathbf{I}_{pq}) \frac{\partial \text{vec} \Sigma_{\beta}^{-1}}{\partial \xi'} \right] \\ &\quad - \left[\left\{ [\text{vec}(\mathbf{X}' \mathbf{Y} \Sigma^{-1}) + \Sigma_{\beta}^{-1} \boldsymbol{\mu}]' \Sigma_{\tilde{\beta}} \right\} \otimes \Sigma_{\tilde{\beta}} \right] \left[\frac{\partial \text{vec}[\Sigma^{-1} \otimes \mathbf{X}' \mathbf{X}]}{\partial \xi'} + \frac{\partial \text{vec} \Sigma_{\beta}^{-1}}{\partial \xi'} \right]. \end{aligned}$$

We can decompose the gradient for the design matrix w.r.t the knots as

$$\frac{\partial \text{vec} \mathbf{X}}{\partial \xi'} = \begin{bmatrix} \mathbf{0}_{(nq_o \times l_s)} & \mathbf{0}_{(nq_o \times l_a)} \\ \partial \text{vec} \mathbf{X}_s / \partial \text{vec}(\xi'_s)' & \mathbf{0}_{(nq_s \times l_a)} \\ \mathbf{0}_{(nq_a \times l_s)} & \partial \text{vec} \mathbf{X}_a / \partial \xi'_a \end{bmatrix}$$

where l_s and l_a are numbers of parameters in the knots locations for surface and additive component, respectively. This decomposition makes user-specified basis functions for different components possible and one may update the locations in a parallel mode (efficient for small models) or batched mode (for models with many parameters). In particular for the thin-plate spline, we have

$$\frac{\partial \text{vec} \mathbf{X}_i}{\partial \xi_i'} = - \begin{bmatrix} (1 + 2 \ln \|\mathbf{x}_i - \xi_{ij}\|)(\mathbf{x}_i - \xi_{ij}) \\ \vdots \\ (1 + 2 \ln \|\mathbf{x}_i - \xi_{ij}\|)(\mathbf{x}_i - \xi_{ij}) \end{bmatrix}_{\substack{i \in \{a,s\}, \\ j \in \{1, \dots, q_i\}}}.$$

Note that the gradient can be obtained efficiently by applying Lemma A.1 and Algorithm A.1 in Section A.3 below whenever $\partial \text{vec} \Sigma_{\beta}^{-1} / \partial \xi'$ and the commutation matrix appear.

A.2. Gradient w.r.t. the shrinkage parameters.

$$\begin{aligned} \frac{\partial \ln p(\boldsymbol{\lambda} | \xi, \Sigma, \mathbf{Y}, \mathbf{X})}{\partial \boldsymbol{\lambda}'} &= \frac{\partial \log p(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}'} - \frac{1}{2} [q_o \boldsymbol{\lambda}_o', q_s \boldsymbol{\lambda}_s', q_a \boldsymbol{\lambda}_a'] - (\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu})' \Sigma_{\beta}^{-1} \frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} \\ &\quad - \frac{1}{2} (\text{vec} \Sigma^{-1})' (\mathbf{I}_p + \mathbf{K}_{p,p}) (\mathbf{I}_p \otimes \tilde{\mathbf{E}}' \mathbf{X}) \frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} \\ &\quad - \frac{1}{2} \text{vec} \left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}) (\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu})' + \Sigma_{\tilde{\beta}} \right]' \frac{\partial \text{vec} \Sigma_{\beta}^{-1}}{\partial \boldsymbol{\lambda}'}, \end{aligned}$$

where

$$\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} = \left\{ \left[(\text{vec}(\mathbf{X}' \mathbf{Y} \Sigma^{-1}) + \Sigma_{\beta}^{-1} \boldsymbol{\mu})' \Sigma_{\tilde{\beta}} \right] \otimes \Sigma_{\beta} - \boldsymbol{\mu}' \otimes \Sigma_{\tilde{\beta}} \right\} \frac{\partial \text{vec} \Sigma_{\beta}^{-1}}{\partial \boldsymbol{\lambda}'},$$

and $\partial \text{vec} \Sigma_{\beta}^{-1} / \partial \boldsymbol{\lambda}'$ can be obtained efficiently by applying Lemma A.1 in Section A.3 and by

$$\begin{aligned} \frac{\partial \text{vec}[(\Lambda_i^{-1/2} \Sigma^{-1} \Lambda_i^{-1/2}) \otimes \mathbf{P}_i]}{\partial \boldsymbol{\lambda}_i'} &= (\mathbf{I}_p \otimes \mathbf{K}_{q_i, p} \otimes \mathbf{I}_{q_i}) (\mathbf{I}_{p^2} \otimes \text{vec} \mathbf{P}_i) (\mathbf{I}_{p^2} + \mathbf{K}_{p, p}) \\ &\times \left(\mathbf{I}_p \otimes [\Lambda_i^{-1/2} \Sigma^{-1}] \right) \frac{\partial \text{vec} \Lambda_i^{-1/2}}{\partial \boldsymbol{\lambda}_i'}, \quad i \in \{a, s\}. \end{aligned}$$

where $\partial \text{vec} \Lambda_i / \partial \boldsymbol{\lambda}_i'$ is $p^2 \times p$ matrix with elements $\nabla_{j(p+1)-p, j} = -1/2 \lambda_{i,j}^{-3/2}$ for $j = 1, \dots, p$ and zero elsewhere.

A.3. Computational remarks. The computational implementation of gradients in Section A.1 and Section A.2 is straightforward but the sparsity of some of the matrices can be exploited in moderate to large datasets. We now present a lemma and an algorithm that can dramatically speed up the computations. It is convenient to define $\mathbf{A}(\mathbf{i}, :)$ and $\mathbf{A}(:, \mathbf{j})$ as matrix operations that reorders the rows and columns of matrix \mathbf{A} with indices \mathbf{i} and \mathbf{j} . Therefore, $\boldsymbol{\beta} = \mathbf{b}(\mathbf{c}, :)$, $\boldsymbol{\mu} = \boldsymbol{\mu}^*(\mathbf{c}, :)$ and $\Sigma_{\beta} = \Sigma_{\mathbf{b}}(\mathbf{c}, \mathbf{c})$ for proper indices \mathbf{c} , and $|\Sigma_{\mathbf{b}}| = |\Sigma_{\beta}|$ since permuting two rows or columns changes the sign but not the magnitude of the determinant.

Lemma A.1. *Given matrix \mathbf{C} and the indexing vector \mathbf{z} such that $(\text{vec} \Sigma_{\mathbf{b}})(\mathbf{z}, :) = \text{vec} \Sigma_{\beta}$ holds, we can decompose the following gradient as*

$$\mathbf{C} \frac{\partial \text{vec}[\Sigma_{\beta}^{-1}(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}'} = \left[\mathbf{C}_s \frac{\partial \text{vec}[(\Lambda_s^{-1/2} \Sigma^{-1} \Lambda_s^{-1/2}) \otimes \mathbf{P}_s]}{\partial \boldsymbol{\theta}_s'}, \mathbf{C}_a \frac{\partial \text{vec}[(\Lambda_a^{-1/2} \Sigma^{-1} \Lambda_a^{-1/2}) \otimes \mathbf{P}_a]}{\partial \boldsymbol{\theta}_a'} \right]$$

where $\boldsymbol{\theta}$ is any parameter vector of the covariance matrix Σ_{β} , $\mathbf{C}_s = \{[\mathbf{C}(:, \mathbf{z})](:, \mathbf{h}_s)\}(:, \mathbf{z}_s \neq 0)$, $\mathbf{h}_s = [(p^2 q q_o + 1), (p^2 q q_o + 2), \dots, p^2 q (q_o + q_s)]'$, $\mathbf{z}_s = \text{vec}([\mathbf{0}_{pq_s \times pq_o}, \mathbf{1}_{pq_s \times pq_s}, \mathbf{0}_{pq_s \times pq_a}]')$, $\mathbf{C}_a = \{[\mathbf{C}(:, \mathbf{z})](:, \mathbf{h}_a)\}(:, \mathbf{z}_a \neq 0)$, $\mathbf{h}_a = [(p^2 q (q_o + q_s) + 1), (p^2 q (q_o + q_s) + 2), \dots, p^2 q^2]'$ and $\mathbf{z}_a = \text{vec}([\mathbf{0}_{pq_a \times p(q_o + q_s)}, \mathbf{1}_{pq_a \times pq_a}]')$.

Algorithm A.1. *An efficient algorithm to calculate $\mathbf{K}_{m,n} \mathbf{Q}$ (or $\mathbf{Q} \mathbf{K}_{m,n}$) where $\mathbf{K}_{m,n}$ is the commutation matrix and \mathbf{Q} is any dense matrix that is conformable to $\mathbf{K}_{m,n}$.*

- (1) Create an $m \times n$ (or $n \times m$) matrix \mathbf{T} and fill it by columns with the sequence $\{1, 2, \dots, nm\}$.
- (2) Obtain the indexing vector $\mathbf{t} = \text{vec}(\mathbf{T}')$.
- (3) Return $\mathbf{Q}(\mathbf{t}, :)$ (or $\mathbf{Q}(:, \mathbf{t})$).

REFERENCES

- Bastos, J. & Ramalho, J. (2010), ‘Nonparametric models of financial leverage decisions’, *CEMAPRE Working Papers*. Available at: <http://cemapre.iseg.utl.pt/archive/preprints/426.pdf>.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and regression trees*, Chapman and Hall/CRC, New York.
- Buhmann, M. (2003), *Radial basis functions: theory and implementations*, Cambridge University Press, Cambridge.
- Chipman, H., George, E. & McCulloch, R. (2010), ‘BART: Bayesian additive regression trees’, *The Annals of Applied Statistics* 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’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 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’, *The Annals of Statistics* 19(1), 1–67.
- Gamerman, D. (1997), ‘Sampling from the posterior distribution in generalized linear mixed models’, *Statistics and Computing* 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. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in J. M. Bernardo, J. O. Berger, A. P. David & A. F. M. Smith, eds, ‘Bayesian Statistics 4’, Oxford University Press, Oxford, pp. 169–193.
- Geweke, J. (2007), ‘Interpretation and inference in mixture models: Simple MCMC works’, *Computational Statistics & Data Analysis* 51(7), 3529–3550.
- Geweke, J. & Amisano, G. (2011), ‘Optimal prediction pools’, *Journal of Econometrics* 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’, *Journal of the American Statistical Association* **98**(462), 352–368.
- Kass, R. (1993), ‘Bayes factors in practice’, *The Statistician* **42**(5), 551–560.
- Khatri, C. & Rao, C. (1968), ‘Solutions to some functional equations and their applications to characterization of probability distributions’, *Sankhyā: The Indian Journal of Statistics, Series A* **30**(2), 167–180.
- Leslie, D., Kohn, R. & Nott, D. (2007), ‘A general approach to heteroscedastic linear regression’, *Statistics and Computing* **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’, *Theory of Probability and its Applications* **9**, 141–142.
- Nott, D. & Leonte, D. (2004), ‘Sampling schemes for Bayesian variable selection in generalized linear models’, *Journal of Computational and Graphical Statistics* **13**(2), 362–382.
- Rajan, R. & Zingales, L. (1995), ‘What do we know about capital structure? Some evidence from international data’, *Journal of Finance* **50**(5), 1421–1460.
- Richardson, S. & Green, P. (1997), ‘On Bayesian analysis of mixtures with an unknown number of components (with discussion)’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **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’, *Journal of Econometrics* **75**(2), 317–343.
- Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.
- Villani, M., Kohn, R. & Nott, D. J. (2012), ‘Generalized Smooth Finite Mixtures’, *Journal of Econometrics* . Forthcoming, available at: <http://dx.doi.org/10.1016/j.jeconom>.

2012.06.012.

- Watson, G. (1964), ‘Smooth regression analysis’, *Sankhyā: The Indian Journal of Statistics, Series A* **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.

SUPPORTING INFORMATION FOR EFFICIENT BAYESIAN MULTIVARIATE SURFACE REGRESSION

FENG LI AND MATTIAS VILLANI

1. PRIOR ROBUSTNESS

This section explores the sensitivity of the posterior inferences with respect to variations in the prior. There are clearly many aspects of the prior to explore, but we will here focus on the sensitivity with respect to the prior on the shrinkage factors and the prior on the knot locations, which are the most influential priors for the model. Since our model is very flexible and richly parametrized it is natural to expect, or even desireable, that the posterior responds to variations in the prior hyperparameters. But since the prior in complex models is always hard to specify, it is hoped that moderate changes in the prior should at least not overturn the posterior inferences.

1.1. The prior on the shrinkage parameters. Figure S1 displays the posterior sensitivity of the knot locations, the posterior mean and standard deviation surfaces to changes in the prior variance on the shrinkage factors. The posterior and predictive results are clearly very robust to changes in this particular aspect of the prior.

1.2. The prior on the knot locations. Figure S2 displays the effect on the posterior knot density from changes in both the mean (columns) and the variance (rows) of prior on the knot locations. While there are some differences in the posterior knot densities when the prior variance changes (changes across rows), there is much larger difference in the posterior of the knots when the prior mean of the knot locations change. This is partly explained by fact that the differences between the three ways of placing the prior means are rather dramatic, but it is clear that the prior mean of the knot locations are affecting where the knots are located a posteriori. Considering the complexity in the inference on the knot locations and the fact that many of the knots probably correspond to regression coefficients that are close to zero, this is perhaps not too surprising.

The posterior inference of the knot locations is typically not of interest. What matters is the inferences on the conditional predictive distribution $p(\mathbf{y}|\mathbf{x})$. Figure S3 and S4 investigate the sensitivity of the posterior mean and standard deviation surfaces to changes in the prior mean and variance of the knot locations. Here the robustness to variations in the prior is much larger. Both the predictive mean and the predictive standard deviation remain largely unchanged, considering

Li (corresponding author): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: feng.li@stat.su.se. Villani: Division of Statistics, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden. E-mail: mattias.villani@liu.se.

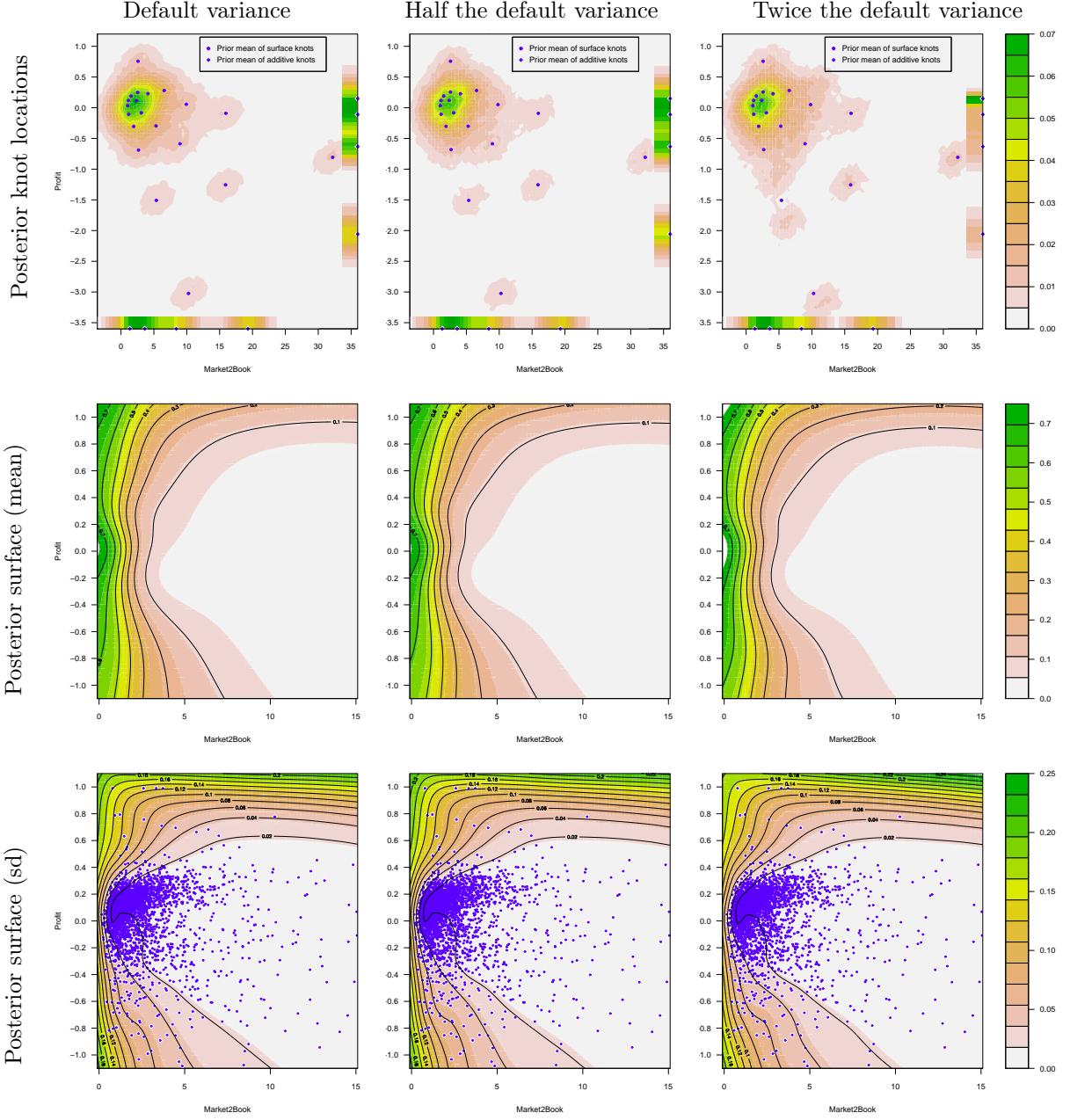


FIGURE S1. Posterior sensitivity with respect to changes in the prior variance of the shrinkage factors. The first column shows the posterior of the knot density (top row), the posterior mean surface (middle row) and the posterior standard deviation surface (bottom row) using the default prior in the paper. The second and third columns demonstrates the effect on the posterior inferences when the prior variance is half of the default value (second column) and twice the default value (third column).

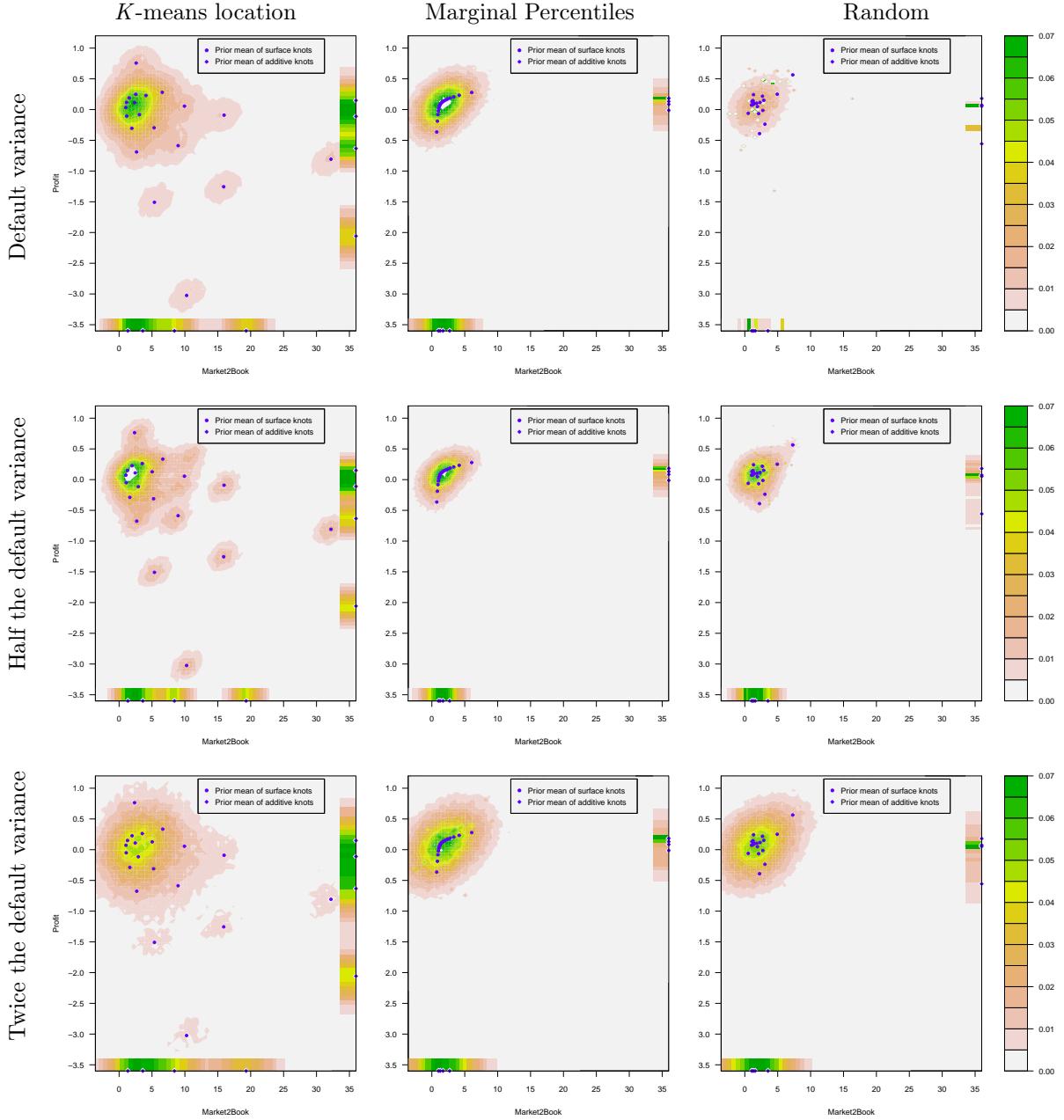


FIGURE S2. Sensitivity of the posterior knot density with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

the magnitude of the changes in the prior. The main differences in the prior mean surface occur in points of covariate space where the uncertainty in the predictive mean is large.

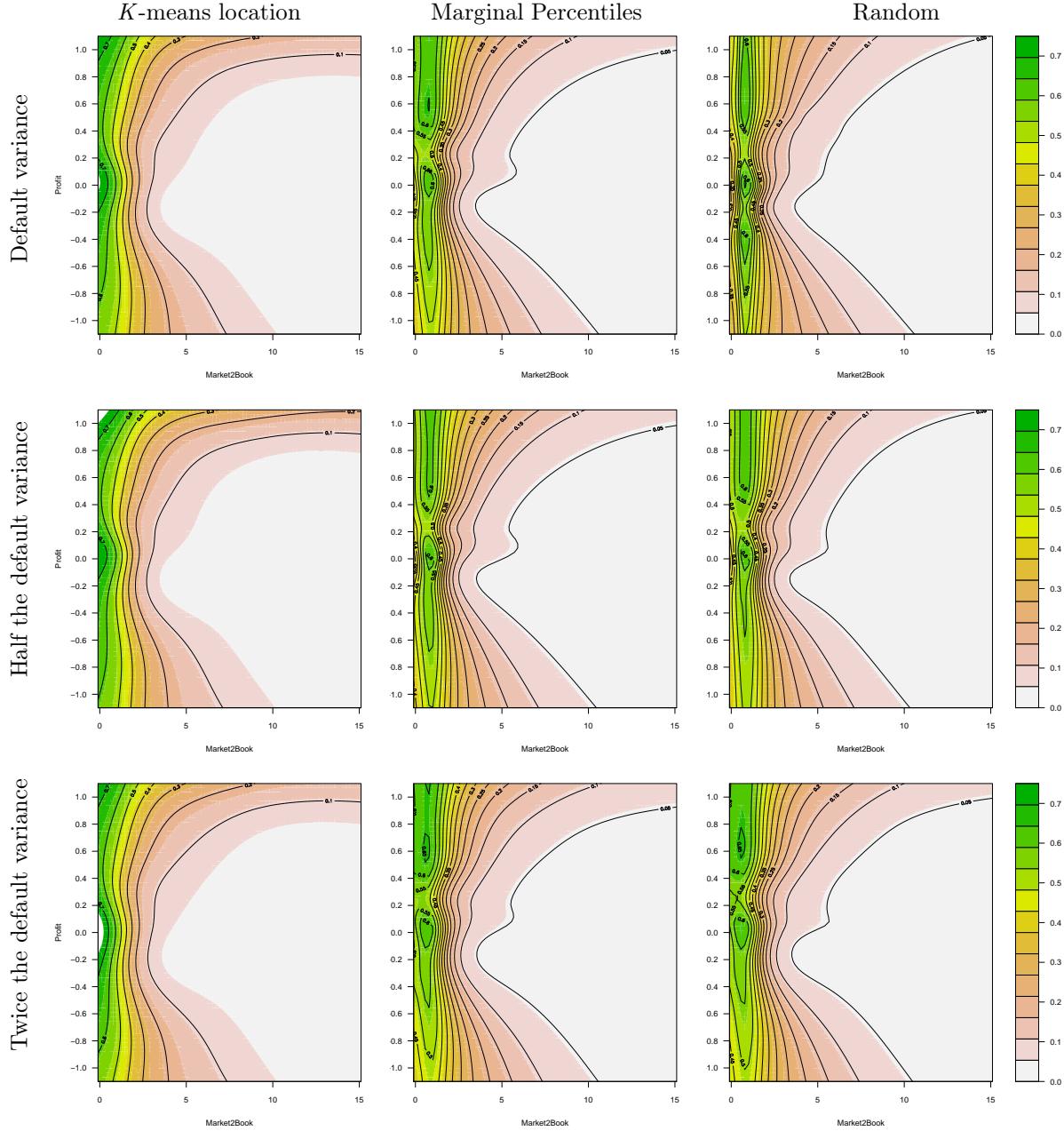


FIGURE S3. Sensitivity of the posterior mean surface with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

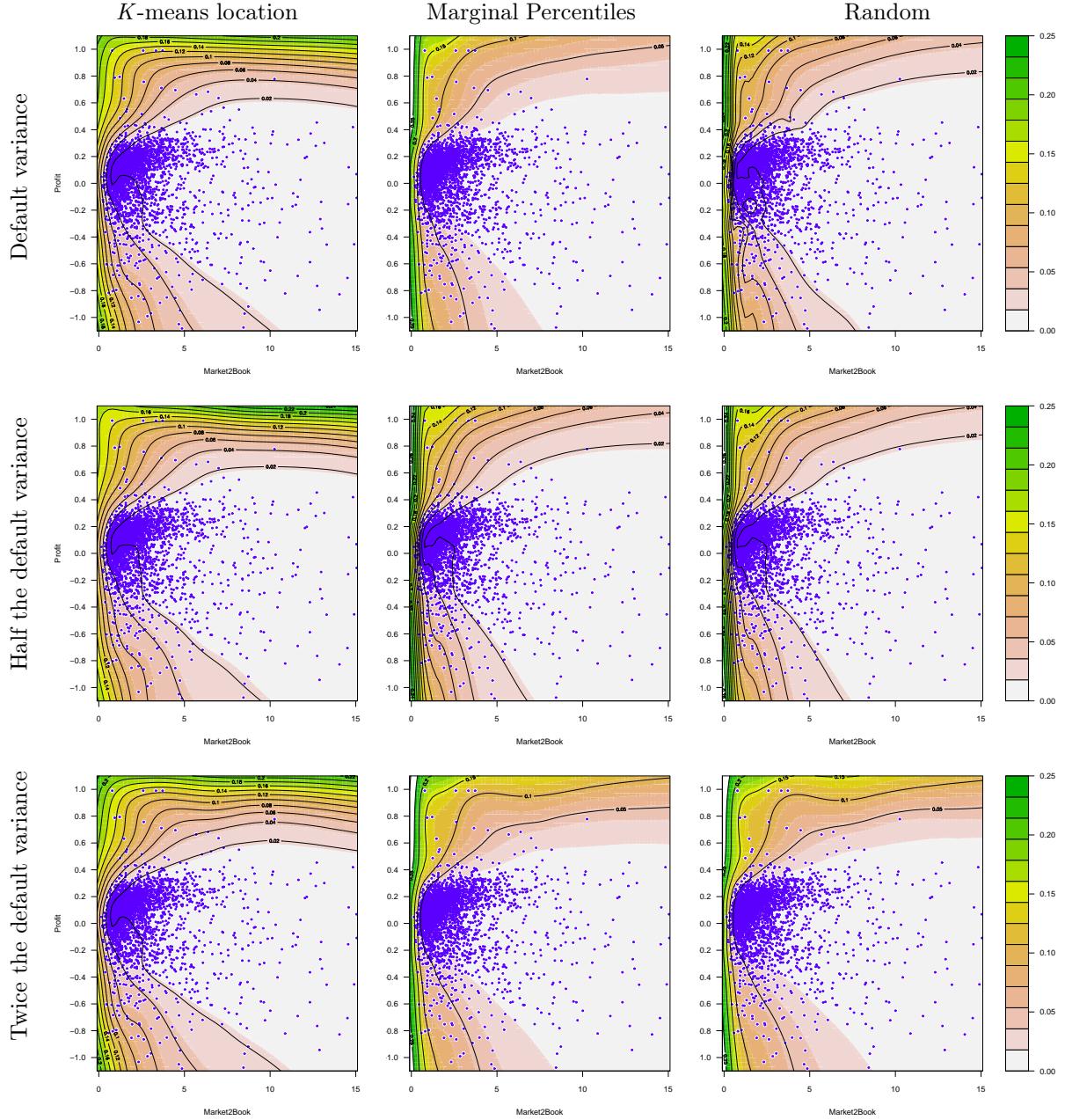


FIGURE S4. Sensitivity of the posterior standard deviation surface with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

2. FURTHER SIMULATION RESULTS

2.1. Additional results from the simulation study in Section 4 of the paper. This section documents the simulation results for the simulation setup with $p = 2$ and $n = 1000$. The simulation setup is identical to the one in Section 4 of the paper with the exception that number of data points is increased from $n = 200$ to $n = 1000$. Figure S5 compares the estimation loss from using fixed and free knots, respectively. Figure S6 compares the out-of-sample predictive residuals from a models with 15 fixed surface knots to a model with 5 free knots, and Figure S7 does the same type of comparison for a model with 20 fixed surface knots to a model with 10 free knots.

2.2. Simulation results from a situation where none of the two models are correct. In this section, we briefly describe a simple simulation example where the true model is not nested in any of the two estimated models. We generate Gaussian data around the following mean surface

$$(S1) \quad y = 42.659[0.1 + (x_1 - 0.5)(0.05 + (x_1 - 0.5)^4 \\ - 10(x_1 - 0.5)^2(x_2 - 0.5)^2 + 5(x_2 - 0.5)^2)]$$

The function in Equation (S1) is called a radial function. We generate 100 datasets using $N(0, 0.1)$ disturbances around the mean. The number of observations are 1000 in each dataset. We use linear and surface components. The number of knots used in the free knot models is 10, 20, and 40, and the number of knots used in the fixed knots model is 10, 20, 40, 60, 80 and 100.

Figure S8 displays boxplots of the losses for the different number of knots in each model. The free knots model outperforms the fixed knots model, but the improvement from using free knots are not that large here since the covariate space is only two-dimensional, which is small enough for the fixed knots to have a decent coverage.

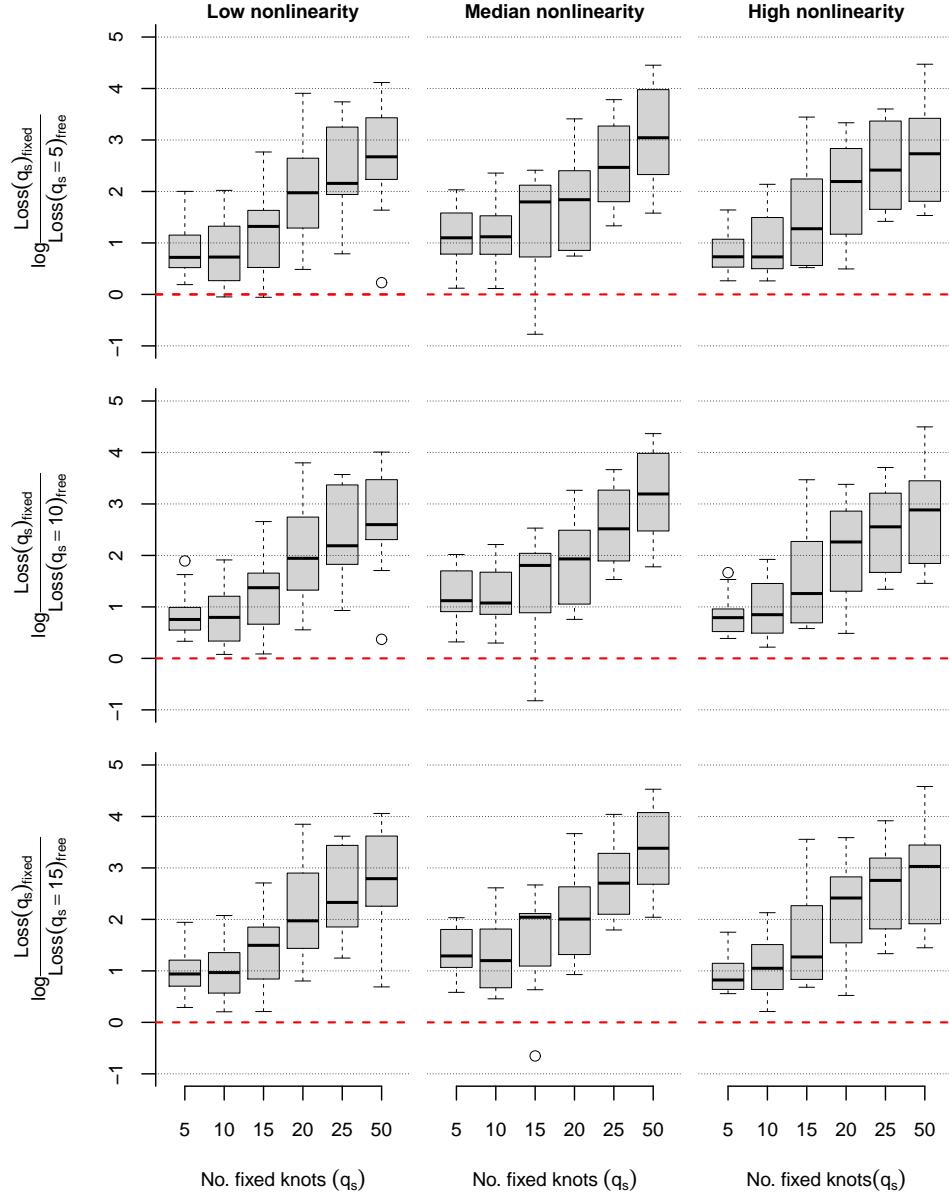


FIGURE S5. Boxplot of the log loss ratio comparing the performance of the fixed knots model with the free knots model for the DGP with $p = 2$ and $n = 1000$. The three columns of the figure correspond to different degrees of nonlinearity of the realized datasets, as measured by estimated DNL in (5) in the paper.

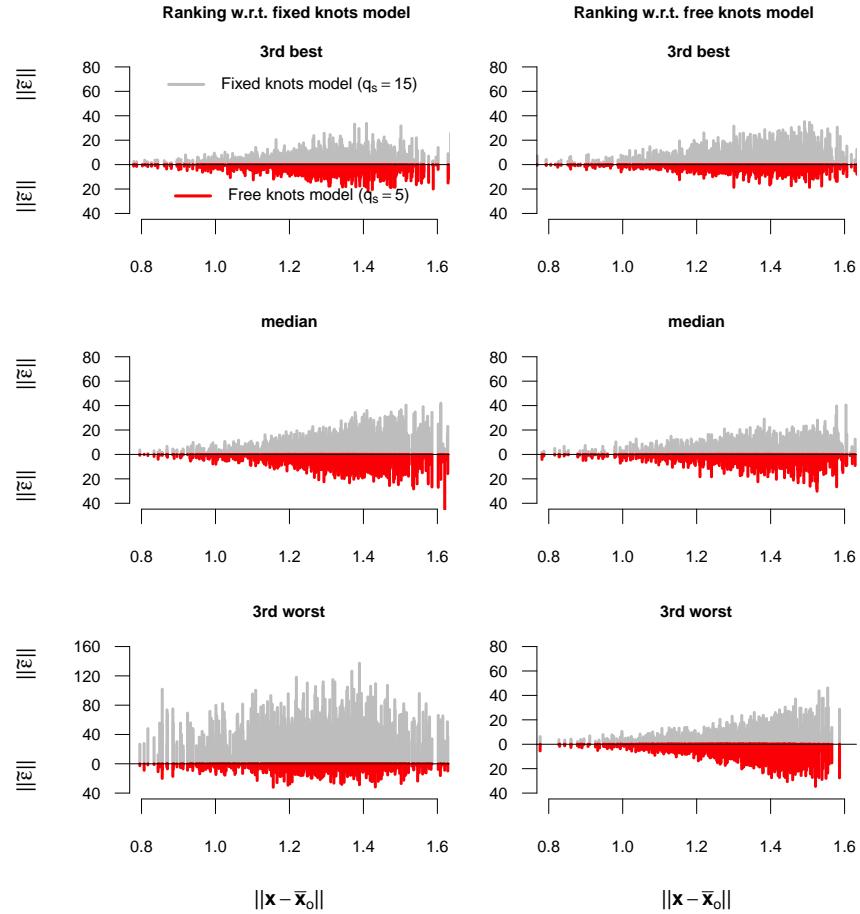


FIGURE S6. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 1000$. The lines in each subplot are the normed residuals from the model with 15 fixed surface knots (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.

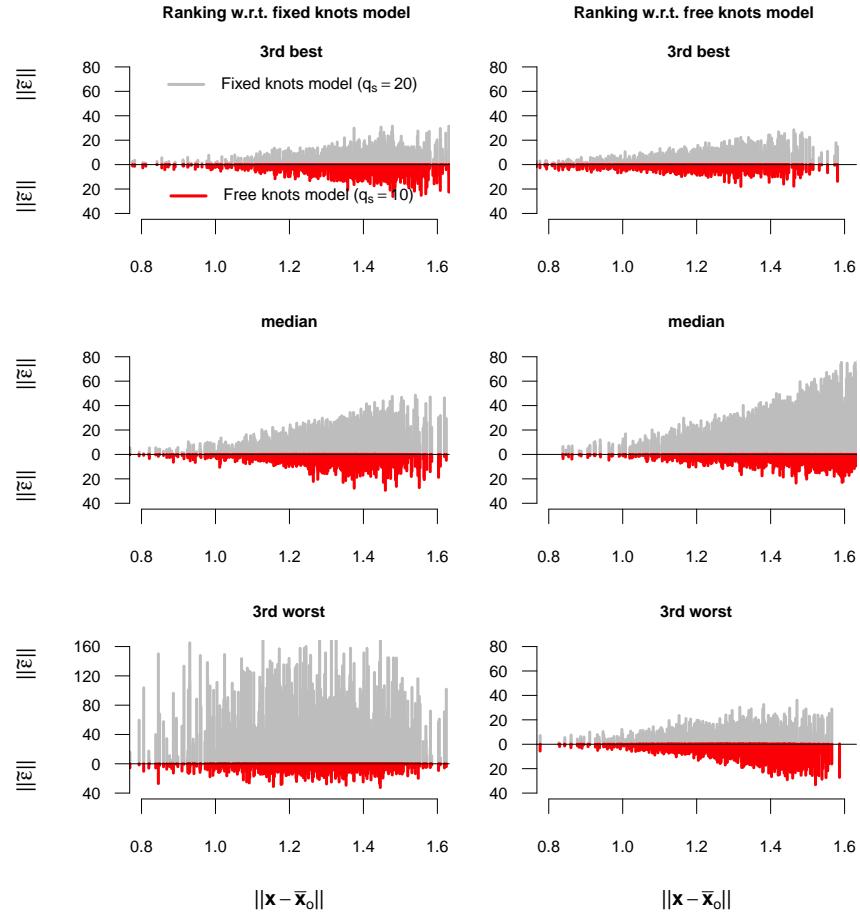


FIGURE S7. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 1000$. The lines in each subplot are the normed residuals from the model with 20 fixed surface knots (vertical bars above the zero line), and the model with 10 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.

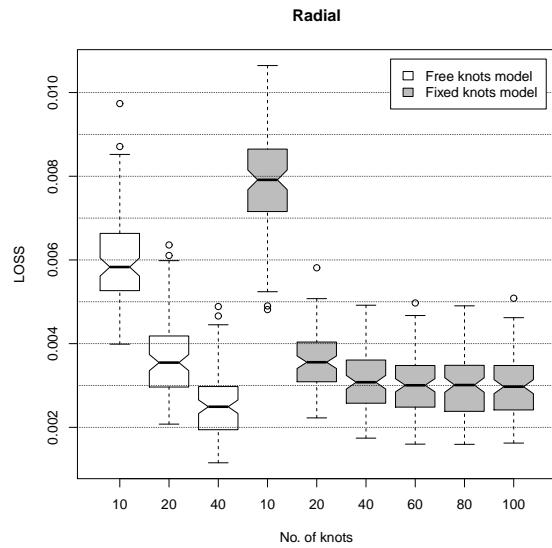


FIGURE S8. Boxplots of the loss in the simulations for the radial mean function.