Biometrika Advance Access published June 24, 2015

Biometrika (2015), *pp.* 1–15
Printed in Great Britain

doi: 10.1093/biomet/asv009

Efficient computation of smoothing splines via adaptive basis sampling

By PING MA

Department of Statistics, University of Georgia, 101 Cedar Street, Athens, Georgia 30602, U.S.A.

pingma@uga.edu

JIANHUA Z. HUANG and NAN ZHANG

Department of Statistics, Texas A&M University, 155 Ireland Street, College Station, Texas 77843, U.S.A.

jianhua@stat.tamu.edu nanzhang@stat.tamu.edu

Summary

Smoothing splines provide flexible nonparametric regression estimators. However, the high computational cost of smoothing splines for large datasets has hindered their wide application. In this article, we develop a new method, named adaptive basis sampling, for efficient computation of smoothing splines in super-large samples. Except for the univariate case where the Reinsch algorithm is applicable, a smoothing spline for a regression problem with sample size n can be expressed as a linear combination of n basis functions and its computational complexity is generally $O(n^3)$. We achieve a more scalable computation in the multivariate case by evaluating the smoothing spline using a smaller set of basis functions, obtained by an adaptive sampling scheme that uses values of the response variable. Our asymptotic analysis shows that smoothing splines computed via adaptive basis sampling converge to the true function at the same rate as full basis smoothing splines. Using simulation studies and a large-scale deep earth core-mantle boundary imaging study, we show that the proposed method outperforms a sampling method that does not use the values of response variables.

Some key words: Bayesian confidence interval; Core-mantle boundary; Nonparametric regression; Penalized least squares; Reproducing kernel Hilbert space; Sampling.

1. Introduction

Consider the nonparametric regression model

$$y_i = \eta(x_i) + \epsilon_i \qquad (i = 1, \ldots, n), \tag{1}$$

where y_i is the ith observation of the response variable, x_i is the ith observation of the predictor variable on the domain $\mathcal{X} \subset \mathbb{R}^d$ $(d \geqslant 1)$, η is the nonparametric function to be estimated, and the ϵ_is are independent and identically distributed random errors with mean zero and unknown constant variance σ^2. A widely used method for estimating the unknown function η in (1) is via minimization of the penalized least squares criterion

$$\mathrm{PLS}(\eta)=\frac{1}{n}\sum_{i=1}^{n}\{y_i-\eta(x_i)\}^2+\lambda J(\eta), \qquad (2)$$

where $J(\eta)$ is a quadratic functional quantifying the roughness of η. The first term on the right of (2) discourages lack of fit, and the second term penalizes the roughness of η. The penalty parameter λ controls the trade-off between the goodness-of-fit and smoothness of η. Multivariate penalty parameters can be introduced when estimating a multivariate function, but we focus on the single penalty case. See Wahba (1990), Gu (2013) and Wang (2011) for overviews of this method, including how to introduce multivariate penalty parameters.

The standard formulation of smoothing splines performs the minimization of (2) in a reproducing kernel Hilbert space $\mathcal{H}=\{\eta : J(\eta)<\infty\}$, where $J(\cdot)$ is a squared semi-norm. Let $\mathcal{N}_J=\{\eta : J(\eta)=0\}$ be the null space of $J(\eta)$ and assume that $\mathcal{N}_J$ is a finite-dimensional linear subspace of $\mathcal{H}$ with basis $\{\xi_i : i=1,\ldots,m\}$, where $m=\dim(\mathcal{N}_J)$. Denote by $\mathcal{H}_J$ the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$ such that $\mathcal{H}=\mathcal{N}_J\oplus\mathcal{H}_J$. Let P be the orthogonal projection operator from $\mathcal{H}$ onto $\mathcal{H}_J$. Then $J(\cdot)$ is a well-defined squared norm of $\mathcal{H}_J$ and for any $\eta\in\mathcal{H}$, $J(\eta)=J(P\eta)=\|P\eta\|^2_{\mathcal{H}_J}$. With this norm, $\mathcal{H}_J$ is also a reproducing kernel Hilbert space, and we denote its reproducing kernel by $R_J(\cdot,\cdot)$.

The reproducing kernel Hilbert space provides a very general framework for nonparametric regression where the penalty term $J(\eta)$ can be chosen to serve different purposes. For univariate function estimation on a compact interval $\mathcal{X}$, one can use

$$J(\eta)=\int_{\mathcal{X}}(\eta^{(m)})^2\,\mathrm{d}x.$$

In particular, $m=2$ corresponds to the commonly-used second derivative penalty and the minimizer of (2) is a natural cubic spline. For estimating a multivariate function on a compact domain $\mathcal{X}\subset\mathbb{R}^d\,(d>1)$, one can use the thin-plate spline penalty

$$J_{md}(\eta)=\int\cdots\int_{\mathcal{X}}\sum_{\nu_1+\cdots+\nu_d=m}\frac{m!}{\nu_1!\cdots\nu_d!}\left(\frac{\partial^m\eta}{\partial x_1^{\nu_1}\cdots\partial x_d^{\nu_d}}\right)^2\mathrm{d}x_1\cdots\mathrm{d}x_d \qquad (3)$$

where m is the order of derivatives and d is the number of predictor variables (Duchon, 1977). As a special case, when $m=2$ and $d=2$ we have

$$J_{22}(\eta)=\iint_{\mathcal{X}}\left(\frac{\partial^2\eta}{\partial x_1^2}\right)^2+\left(\frac{\partial^2\eta}{\partial x_1\partial x_2}\right)^2+\left(\frac{\partial^2\eta}{\partial x_2^2}\right)^2\mathrm{d}x_1\,\mathrm{d}x_2.$$

See Gu (2013) for details about defining the penalty term and corresponding reproducing kernel Hilbert space for modelling a multivariate regression function using smoothing spline analysis of variance models.

Univariate smoothing splines can be computed in $O(n)$ operations by applying the Reinsch (1967) algorithm. In general, as we shall see in the next section, the computational cost of finding the minimizer of (2) is in the order of $O(n^3)$ and thus is very expensive for big datasets. To lower the computational cost, over the past decades, there have been efforts to find sparse sets of basis functions to approximate the minimizer of (2). Luo & Wahba (1997) and Zhang et al. (2004) applied variable selection techniques, but it is not clear whether the resulting estimators share the good asymptotic properties of standard smoothing splines. Gu & Kim (2002) and

Kim & Gu (2004) developed a simple random sampling approach for basis function selection and established a coherent theory for the convergence of their approximated smoothing splines. To overcome the computational burden of smoothing splines, pseudosplines (Hastie, 1996) and penalized splines (Ruppert et al., 2003) have also been proposed. Both use a small number of fixed basis functions to approximate the smoothing splines; they are similar in spirit to Gu & Kim (2002) and Kim & Gu (2004) but differ in the construction of the basis functions.

In this paper, extending the simple random sampling approach of Gu & Kim (2002) and Kim & Gu (2004), we develop an adaptive basis sampling method for approximating smoothing splines. Its novelty is that we select the basis functions according to the slicing along the range of the response variable. These adaptively selected basis functions form a reduced model space, called the effective model space. We compute the approximated smoothing spline estimator on the reduced space to achieve efficient computation. This adaptive sampling strategy differs from all existing methods based on sampling basis functions on the direction of the predictors. It can recover fine details of the response surface better than the simple random sampling scheme.

We develop an asymptotic theory on the rate of convergence of our approximated smoothing spline estimator. This theory is nonstandard because of the response-dependent sampling scheme, and yields conditions on the dimension of the effective model space to warrant the same convergence rate as the regular smoothing spline estimators. Such conditions provide useful practical guidelines for the sample size of the adaptive sampling.

2. Smoothing splines and computational issues

We first state the so-called representer theorem (e.g., Wahba, 1990), which declares that although the original penalized least squares problem for smoothing splines is formulated in the infinite-dimensional function space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$, the solution lies in a finite-dimensional space. Recall that $\mathcal{H}$ has the tensor-sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, $\{\xi_i\}_{i=1}^m$ spans the null space $\mathcal{N}_J$ of the quadratic functional J, and $R_J(\cdot, \cdot)$ is the reproducing kernel of $\mathcal{H}_J$.

Theorem 1. *There exist vectors $d = (d_1, \ldots, d_m)^{\mathrm{T}} \in \mathbb{R}^m$ and $c = (c_1, \ldots, c_n)^{\mathrm{T}} \in \mathbb{R}^n$ such that the minimizer of* (2) *over $\mathcal{H}$ can be represented as*

$$\eta(x) = \sum_{k=1}^{m} d_k \xi_k(x) + \sum_{i=1}^{n} c_i R_J(x_i, x), \qquad x \in \mathcal{X}. \tag{4}$$

Theorem 1 implies that we need search for the minimizer of (2) only over the collection of functions of form (4), so the problem reduces to finding the coefficient vectors d and c that satisfy a system of linear equations. Let $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ be the vector of observed values of the predictor variable, and $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the vector of corresponding observations of the response variable. Let $\eta = \{\eta(x_1), \ldots, \eta(x_n)\}^{\mathrm{T}}$ denote the n evaluations of $\eta(\cdot)$ at x, S denote the $n \times m$ matrix with the (i, j)th entry $\xi_j(x_i)$, and R denote the $n \times n$ matrix with the (i, j)th entry $R_J(x_i, x_j)$. Then the decomposition (4) applied to x yields the system of equations

$$\eta = Sd + Rc,$$

and thus the first term on the right-hand side of (2) becomes

$$n^{-1}(y - Sd - Rc)^{\mathrm{T}}(y - Sd - Rc). \tag{5}$$

On the other hand, for any function η with the expansion (1), the penalty function $J(\eta)$ in (2) can also be written in a matrix form using the reproducing property of $R_J(\cdot,\cdot)$, i.e.,

$$\langle R_J(x_i,\cdot), R_J(x_j,\cdot)\rangle_{\mathcal{H}_J} = R_J(x_i,x_j).$$

Recall that $P:\mathcal{H}\to\mathcal{H}_J$ is a projection operator. For any η as in (4), $P\eta=\sum_{i=1}^n c_i R_J(x_i,\cdot)$. Hence

$$\begin{aligned} J(\eta) = \|P\eta\|^2_{\mathcal{H}_J} &= \left\langle \sum_{i=1}^n c_i R_J(x_i,x), \sum_{i=1}^n c_i R_J(x_i,x) \right\rangle_{\mathcal{H}_J} \\ &= \sum_{i=1}^n\sum_{j=1}^n c_i R_J(x_i,x_j)c_j = c^{\mathrm{T}} R\, c. \end{aligned} \tag{6}$$

Combining (5) and (6), we see that the penalized least squares criterion (2) is reduced to

$$\textsc{pls}(\eta) = \frac{1}{n}(y - Sd - Rc)^{\mathrm{T}}(y - Sd - Rc) + \lambda\, c^{\mathrm{T}} R\, c. \tag{7}$$

Since $\textsc{pls}(\eta)$ is a quadratic form in both d and c, its minimizer has a closed-form expression. Differentiating (7) with respect to d and c and setting the derivatives to zero, we obtain the linear system of equations

$$\begin{pmatrix} S^{\mathrm{T}}S & S^{\mathrm{T}}R \\ R^{\mathrm{T}}S & R^{\mathrm{T}}R + n\lambda R \end{pmatrix}\begin{pmatrix} d \\ c\end{pmatrix} = \begin{pmatrix} S^{\mathrm{T}}y \\ R^{\mathrm{T}}y\end{pmatrix}.$$

To solve this system, of size $m+n$, the computational cost is generally of the order $O(n^3)$, which can be prohibitive when the sample size n is large. From Theorem 1, the number of basis functions used to represent the solution is $m+n$, which grows with n. While the m basis functions for $\mathcal{N}_J$ are needed, it may not be necessary to use all n basis functions for $\mathcal{H}_J$. If a smaller number of basis functions can provide a good approximation of the smoothing spline solution, then a computationally efficient algorithm can be developed to handle cases with large sample size. We discuss two sampling approaches for selecting basis functions in the next section.

3. Sampling of basis functions

3·1. *Uniform sampling of basis functions*

We first review an approach to selecting basis functions by randomly sampling the observations of the predictor variable and discuss its limitations, and then present our new sampling approach that involves the response variable.

From the representer theorem, each of the n basis functions for representing the function in $\mathcal{H}_J$ is uniquely associated with an observed value of the predictor variable. Thus a natural idea for selecting the basis functions is through randomly sampling the observed values of the predictor variable. Specifically, we draw a random sample of size n^* from the observed predictor values $\{x_i\}_{i=1}^n$, denoted as $x^*=(x_1^*,\ldots,x_{n^*}^*)^{\mathrm{T}}$, and use the corresponding basis functions, $\{R_J(x_i^*,x)\}_{i=1}^{n^*}$, to represent functions in $\mathcal{H}_J$. We then solve the penalized least squares problem in the effective model space $\mathcal{H}_E=\mathcal{N}_J\oplus\mathrm{span}\{R_J(x_i^*,x), i=1,\ldots,n^*\}$. When n^* is much smaller than n, the computational cost can be significantly reduced.

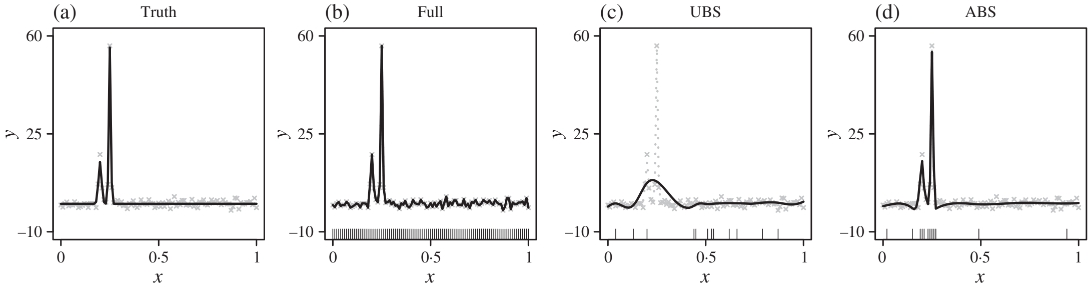

Fig. 1. Toy regression function with two close peaks: (a) true signal (solid line) and 100 observations (grey crosses); (b) smoothing spline fit (solid line) with full basis; (c) smoothing spline fit (solid line) with 12 uniformly sampled basis functions (UBS); (d) smoothing spline fit (solid line) with 12 adaptively sampled basis functions (ABS). In (b)–(d), short vertical lines at the bottom mark the data points corresponding to the selected basis functions; observations are indicated by grey crosses; the true signal is shown as dotted grey lines.

Gu & Kim (2002) and Kim & Gu (2004) proved that this uniform sampling scheme has some nice theoretical properties. Under some reasonable conditions, the smoothing spline estimator computed under this scheme can achieve the same asymptotic convergence rate as the full basis smoothing spline estimator that uses all the basis functions indicated in the representer theorem.

When the number of sampled basis functions increases, the estimator from the uniform sampling strategy will approach the smoothing spline estimator and reveal the underlying true function. However, if constrained by computational resources, one may not sample enough basis functions to achieve a satisfactory result. Figure 1 illustrates this with a toy example. The underlying true function is the density function of a two-component mixture of normal distributions. Panel (c) shows the smoothing spline fit using 12 uniformly sampled basis functions, which does not reveal the two peaks of the mixture components because uniform sampling does not select the basis function corresponding to the point with the largest y-value. Unless the number of basis functions is greatly increased, there is little chance that the estimator can capture this peak.

3·2. *Adaptive sampling of basis functions*

We propose a new sampling scheme to select basis functions which makes use of the observed values of the response variable. This scheme may sample more basis functions in regions where the response function has big changes and sample fewer basis functions where the response surface is relatively flat. We call this new scheme adaptive basis sampling.

Like the uniform sampling scheme discussed in § 3·1, adaptive sampling also samples the basis functions from the collection $\{R_J(x_i, \cdot) : i = 1, \ldots, n\}$ as indicated in the representer theorem. The difference is the way the sampling is performed. In adaptive basis sampling, we first group the x_is according to the corresponding value of the response variable, and then draw random samples within each group. The detailed procedure is given below.

Step 1. Divide the range of the responses $\{y_i\}_{i=1}^{n}$ into K disjoint intervals, $S_1, \ldots, S_K$. Let $|S_k|$ denote the number of observations in S_k.

Step 2. For each S_k, consider the collection of all pairs (x_i, y_i) where $y_i \in S_k$, and draw a random sample of size n_k from this collection. Denote the sampled predictor values by $x^{*(k)} = (x_1^{*(k)}, \ldots, x_{n_k}^{*(k)})$.

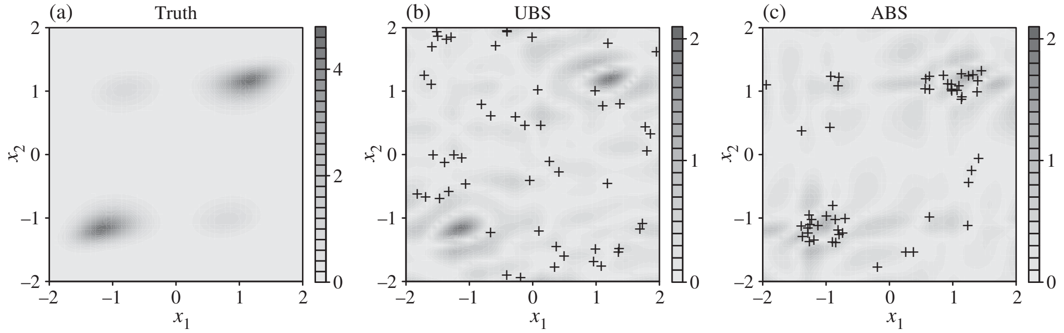

Fig. 2. Bivariate nonparanormal copula density function. (a): contour plot of true function; (b)–(c): contour plots of absolute values of fitting residuals by smoothing splines based on uniform basis sampling (UBS) and adaptive basis sampling (ABS). The sampled basis functions are marked by $+$.

Step 3. Combine $x^{*(1)}, \dots, x^{*(K)}$ together to form a set of sampled predictor values $\{x_1^*, \dots, x_{n^*}^*\}$. This set has size $n^* = \sum_{k=1}^{K} n_k$.

Step 4. Form the effective model space

$$\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_j^*, \cdot),\ j = 1, \dots, n^*\}.$$

Minimize the penalized least squares criterion (2) over this effective model space.

The first step of the adaptive basis sampling procedure groups together observations with similar response values. It is the same operation as binning when constructing histograms and slicing in sliced inverse regression (Li, 1991; Cook, 1998). Each set $\{(x_i, y_i) : y_i \in S_k\}$ is referred to as a slice of the data. We expect this adaptive sampling scheme to select more effective basis functions than uniform sampling.

Figure 1(d) displays the smoothing spline fit from the adaptive sampling scheme with 12 basis functions. The fit reveals the two peaks of the mixture components well, since basis functions corresponding to the peak points are sampled.

To further illustrate how adaptive basis sampling works and compare it with uniform basis sampling, we considered a two-dimensional example for which the response surface is a bivariate nonparanormal copula density function; see § 5 for its analytical form. Figure 2(a) depicts the contour plot of the true function, showing four peaks: two are significantly higher than the others. Contour plots of absolute values of residuals after smoothing spline fitting, presented in Figs. 2(b)–(c), indicate that the estimated two big peaks from adaptive basis sampling are closer to the truth than from uniform basis sampling. That the adaptive basis sampling smoothing spline yields a better estimate can be explained by the distribution of sampled basis functions, also shown in Figs. 2(b) and (c): the basis functions sampled by uniform basis sampling are spread over the whole domain while those sampled by adaptive basis sampling are mainly distributed around the four peaks, especially the two significant ones.

In § 4, we show that the adaptive sampling scheme can achieve the asymptotic rate of convergence of the original smoothing spline estimator, although a much smaller set of basis functions is employed. The theoretical results of Gu & Kim (2002) and Kim & Gu (2004) for uniform sampling cannot be applied to adaptive sampling, because values of the response variable are used in selecting the basis functions.

3·3. *Efficient computation*

We now present the details of the computational algorithm when adaptive basis sampling is used to compute the smoothing spline estimator. Recall that the selected data points are denoted by $x^* = (x_1^*, \dots, x_{n^*}^*)^{\mathrm{T}}$. Under adaptive basis sampling, the minimizer of (2) is approximated by

$$\eta_A(x) = \sum_{k=1}^{m} d_k \xi_k(x) + \sum_{j=1}^{n^*} c_j R_J(x_j^*, x).$$

We let S denote the $n \times m$ matrix with (i, j)th entry $\xi_j(x_i)$. Let R_* be a $n \times n^*$ matrix with the (i, j)th entry $R_J(x_i, x_j^*)$ and R_{**} be a $n^* \times n^*$ matrix with the (i, j)th entry $R_J(x_i^*, x_j^*)$. If we rearrange the original data by putting the selected data points x^* at the front, R_* is just the left part of R while R_{**} is the top-left corner of R. The evaluations of η_A at locations x, $\eta_A = \{\eta_A(x_1), \dots, \eta_A(x_n)\}^{\mathrm{T}}$, satisfy

$$\eta_A = S d_A + R_* c_A,$$

where $d_A = (d_1, \dots, d_m)^{\mathrm{T}}$ and $c_A = (c_1, \dots, c_{n^*})^{\mathrm{T}}$.

Similar to (7), we have

$$\text{PLS}(\eta_A) = \frac{1}{n}(y - Sd_A - R_* c_A)^{\mathrm{T}}(y - Sd_A - R_* c_A) + \lambda\, c_A^{\mathrm{T}} R_{**}\, c_A,$$

whose minimizer $(\hat{d}_A, \hat{c}_A)$ satisfies the linear system of equations

$$\begin{pmatrix} S^{\mathrm{T}} S & S^{\mathrm{T}} R_* \\ R_*^{\mathrm{T}} S & R_*^{\mathrm{T}} R_* + n\lambda R_{**} \end{pmatrix} \begin{pmatrix} d_A \\ c_A \end{pmatrix} = \begin{pmatrix} S^{\mathrm{T}} y \\ R_*^{\mathrm{T}} y \end{pmatrix}. \tag{8}$$

System (8) can be solved using a method described in Golub & Van Loan (1989). First, a pivoted Cholesky decomposition is performed such that the first matrix on the left-hand side of (8) equals $G^{\mathrm{T}} G$, where G is an upper triangular matrix. Then, forward and backward substitutions are used to solve the system of equations to obtain the estimated coefficients. However, care should be taken when R_* is singular, i.e., the bottom diagonal elements of G are zeros. Kim & Gu (2004) suggested replacing those zeros by an appropriate small value δ and proceeding as if R_* is of full rank.

A standard method for data-driven choice of the penalty parameter λ is to minimize the generalized crossvalidation criterion (Craven & Wahba, 1979). To give a formal definition of this, note that the fitted values $\hat{y} = \{\hat{\eta}_A(x_1), \dots, \hat{\eta}_A(x_n)\}^{\mathrm{T}}$ can be obtained from the estimated coefficients as $\hat{y} = S\hat{d}_A + R_* \hat{c}_A$. In light of (8), $\hat{y} = A(\lambda) y$, where $A(\lambda)$ is the smoothing matrix

$$A(\lambda) = (S, R_*) \begin{pmatrix} S^{\mathrm{T}} S & S^{\mathrm{T}} R_* \\ R_*^{\mathrm{T}} S & R_*^{\mathrm{T}} R_* + n\lambda R_{**} \end{pmatrix}^{+} \begin{pmatrix} S^{\mathrm{T}} \\ R_*^{\mathrm{T}} \end{pmatrix},$$

and C^+ denotes the Moore–Penrose inverse of C. The criterion is defined as

$$\text{GCV}(\lambda) = \frac{n^{-1} y^{\mathrm{T}} \{I - A(\lambda)\}^2 y}{[n^{-1} \mathrm{tr}\{I - A(\lambda)\}]^2}, \tag{9}$$

and we minimize it as a function of the penalty parameter λ (Tenorio et al., 2011), using standard nonlinear optimization algorithms. We use the modified Newton algorithm developed by Dennis & Schnabel (1996).

Now we calculate the computational complexity, using the fact that $m \ll n^* \ll n$ to simplify the expressions. The construction of the linear system (8) is of the order $O(nn^{*2})$, the Cholesky decomposition takes $O(n^{*3})$ flops, the subsequent forward and backward substitutions take $O(n^{*2})$ flops respectively, and the evaluation of (9) requires the calculation of $\text{tr}\{A(\lambda)\}$, which takes $O(nn^{*2})$ flops. The overall computational cost is of the order $O(nn^{*2})$. The efficient computational scheme can also be used to compute Bayesian confidence intervals (Wahba, 1983); see the Supplementary Material for details.

4. Convergence rates for function estimation

4·1. *Regularity conditions*

We first introduce an inner product associated with the marginal density $f_X(\cdot)$ of the predictor variable X. For any g_1 and g_2 in $\mathcal{L}_2(\mathcal{X})$, define

$$V(g_1, g_2) = \langle g_1, g_2 \rangle = \int_{\mathcal{X}} g_1(x) g_2(x) f_X(x)\, \mathrm{d}x.$$

The norm induced by this inner product is a weighted version of the $\mathcal{L}_2$-norm and the weighting function is the marginal density of the predictor. We define the mean squared error of the estimator $\hat{\eta}_A$ in estimating the regression function η as the quadratic functional

$$V(\hat{\eta}_A - \eta) = \|\hat{\eta}_A - \eta\|^2 = \langle \hat{\eta}_A - \eta, \hat{\eta}_A - \eta \rangle = \int_{\mathcal{X}} \{\hat{\eta}_A(x) - \eta(x)\}^2 f_X(x)\, \mathrm{d}x.$$

This is a common measure in studying statistical properties of smoothing splines (e.g., Gu & Qiu, 1994).

In the literature, the convergence rate of smoothing splines is usually characterized by an eigenanalysis of the penalty functional J with respect to the quadratic functional V. We now state two commonly-used technical conditions (Gu, 2013). A quadratic functional B is said to be completely continuous with respect to another quadratic functional A, if for any $\epsilon > 0$, there exists a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_1(\eta) = \cdots = L_k(\eta) = 0$ implies that $B(\eta) \leqslant \epsilon A(\eta)$; see Weinberger (1974, § 3.3).

Condition 1. The functional V is completely continuous with respect to J.

By Theorem 3.1 of Weinberger (1974), Condition 1 implies that V and J can be simultaneously diagonalized; see, e.g., Silverman (1982) and Gu (2013, § 9.1). More precisely, there exist a sequence of eigenfunctions $\phi_\nu \in \mathcal{H}$ and the associated nonnegative sequence of eigenvalues $\rho_\nu \uparrow \infty$ such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$ where $\delta_{\nu\mu}$ is the Kronecker delta. Furthermore, any function f satisfying $J(f) < \infty$ can be expressed as a Fourier series expansion $f = \sum_\nu f_\nu \phi_\nu$, where $f_\nu = V(f, \phi_\nu)$.

Condition 2. For some $r > 1$ and $\beta > 0$, $\rho_\nu > \beta \nu^r$ for sufficiently large ν.

This condition on the growth rate of the eigenvalues is essentially a requirement on the smoothness of $\eta \in \mathcal{H}$. For one-dimensional cubic spline smoothing on a compact interval $\mathcal{X}$ with $J(\eta) = \int_{\mathcal{X}} \{\eta''\}^2$, Conditions 1 and 2 are satisfied with $r = 4$ when $V(\eta)$ is equivalent to the standard L_2 norm (Utreras, 1981). For thin-plate splines on a bounded domain of $\mathcal{X} \in \mathbb{R}^d$ with the penalty (3), Conditions 1 and 2 are satisfied with $r = 2m/d$. For tensor-product smoothing splines with penalty $J(\eta) = \sum_{\beta=1}^{s} \theta_\beta^{-1} \| P_\beta \eta \|_{\mathcal{H}^\beta}^2$, one can prove that Condition 1 holds using the argument in Example 9.2 of Gu (2013), and Condition 2 holds with $r = 4 - \epsilon$, where $\epsilon > 0$ (Wahba, 1990).

Condition 3. For a constant $C < \infty$, $\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)\} \leqslant C$ for all ν and μ.

Since ϕ_ν is an orthonormal system relative to $V(\cdot, \cdot)$, i.e.,

$$V(\phi_\nu, \phi_\mu) = \int_{\mathcal{X}} \phi_\nu(x)\phi_\mu(x) f_X(x)\, \mathrm{d}x = \delta_{\nu\mu},$$

we have that

$$\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)\} = \int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x) f_X(x)\, \mathrm{d}x - \delta_{\nu\mu}.$$

Thus Condition 3 is equivalent to the requirement that $\int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x) f_X(x)\, \mathrm{d}x$ is uniformly bounded for all ν and μ.

4·2. *Convergence rates*

This section presents our main results on convergence rates. All proofs are given in the Supplementary Material.

In our adaptive sampling scheme, the search for the smoothing spline estimator is restricted to the effective model space $\mathcal{H}_E$. We first establish a lemma that justifies the use of the effective model space. Let $\mathcal{H} \ominus \mathcal{H}_E$ denote the orthogonal complement of $\mathcal{H}_E$ in the reproducing kernel Hilbert space $\mathcal{H}$.

Lemma 1. *As $\lambda \to 0$ and $n^*\lambda^{2/r} \to \infty$, if the function h is not in the effective model space, i.e., $h \in \mathcal{H} \ominus \mathcal{H}_E$, we have $V(h) = o_p\{\lambda J(h)\}$.*

This result suggests that compared to $\lambda J(h)$, $V(h)$ is negligible when h is orthogonal to $\mathcal{H}_E$, and implies that the space orthogonal to the effective model space $\mathcal{H}_E$ is effectively suppressed by the penalty $\lambda J(\eta)$. Hence, we can capture the essential features of the true function η_0 by restricting the estimator to the effective model space $\mathcal{H}_E$.

For completeness, we state below a standard result for the convergence rate of smoothing splines (e.g., Theorem 9.17 of Gu, 2013).

Theorem 2. *If $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ for some $p \in [1, 2]$, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, then $(V + \lambda J)(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p)$.*

We now present our main result on the convergence rate of the smoothing spline estimator based on the proposed adaptive basis sampling scheme.

Theorem 3. *If $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ for some $p \in [1, 2]$, as $\lambda \to 0$ and $n^*\lambda^{2/r} \to \infty$, then $(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p)$. In particular, when $\lambda \asymp n^{-r/(pr+1)}$, the estimator achieves the optimal convergence rate,*

$$(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p\{n^{-pr/(pr+1)}\}.$$

This theorem states that, under regularity conditions, the convergence rate of the smoothing spline estimator using an adaptively sampled basis equals that of the smoothing spline estimator using the full basis indicated by the representer theorem. The parameter p in the condition yields a faster rate of convergence for certain functions: for the roughest η satisfying $J(\eta) < \infty$, we have $p = 1$, whereas for the smoothest η, we have $p = 2$; see Wahba (1985) for details.

Note that $J(\eta_0) = \sum_i \rho_i V(\eta_0, \phi_i)^2$. When $J(\eta_0) < \infty$, the condition in Theorem 3 holds with $p = 1$, and the convergence rate is $O_p(n^{-r/(r+1)})$. When η_0 is in the Sobolev space $W^{m,2}$ on a bounded domain in $\mathbb{R}^d$, we have $r = 2m/d$ and Theorem 3 yields the convergence rate $n^{-2m/(2m+d)}$, which is the optimal rate of convergence (Stone, 1982). For the case $d = 1$, Claeskens et al. (2009) and Wang et al. (2011) showed that penalized splines can also achieve the optimal rate of convergence.

Theorem 3 helps determine the dimension of the effective model space $\mathcal{H}_E$. With $\lambda \asymp n^{-r/(pr+1)}$, Lemma 1 and Theorem 3 require that $n^*\lambda^{2/r} \to \infty$, which suggests that a suitable choice of n^* should satisfy $n^* \asymp n^{2/(pr+1)+\delta}$, where δ is an arbitrary small positive number. For univariate cubic smoothing splines with the penalty $J(\eta) = \int_0^1 (\eta'')^2$, $r = 4$ and $\lambda \asymp n^{-4/(4p+1)}$, a suitable choice of the dimension of the effective model space is $n^* = n^{2/(4p+1)+\delta}$, which lies in the interval $(O(n^{2/9+\delta}), O(n^{2/5+\delta}))$ for p taking values in $[1, 2]$. For tensor-product splines, $r = 4 - \epsilon$, where $\epsilon > 0$, a suitable choice of the dimension of effective model space is $n^* = n^{2/(4p+1)+\delta}$, which is roughly in interval $(O(n^{2/9+\delta}), O(n^{2/5+\delta}))$. In our simulation study and real data analysis, we take the dimension of the effective model space n^* to be between $5n^{2/9}$ and $20n^{2/9}$.

5. Simulation results

Using simulated multivariate regression functions, we compared the smoothing spline estimators based on adaptive basis sampling and uniform basis sampling in terms of estimation accuracy and computational time. We also compared adaptive basis sampling with fast bivariate P-splines, an efficient algorithm for bivariate spline smoothing (Xiao et al., 2013).

Some of our simulation set-ups involve the joint probability density of a p-dimensional nonparanormal distribution (Liu et al., 2009)

$$\eta_{\text{copula}}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\{f(x) - \mu\}^{\mathrm{T}}\Sigma^{-1}\{f(x) - \mu\}\right] \prod_{j=1}^{p} |f_j'(x_j)|, \qquad (10)$$

where $\mu = 0$, Σ has ones as diagonal entries, 0·5 as off-diagonal elements, and

$$f_j(x) = \alpha_j \operatorname{sign}(x)\,|x|^{\alpha_j} \qquad (j = 1, \ldots, p).$$

This is essentially a probability density function for a Gaussian copula model.

We generated data according to model (1) where the predictor variable x was randomly generated from the uniform distribution over the domain of interest. The signal-to-noise ratio, defined as $\text{var}\{\eta(X)\}/\sigma^2$, was set to three levels: 10, 2, 0·4. For each simulation set-up, samples of $n = 1600$ were generated. We considered four regression function settings:

1. a bivariate blocks function, $\eta_{\text{blocks}}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = \text{blocks}(x_{\langle 1\rangle})$, where blocks$(\cdot)$ is the univariate blocks function used in (Donoho & Johnstone, 1994). It has frequent and irregular abrupt changes in one direction and stays constant in the other. The domain of interest is the unit square;
2. a bivariate copula function, given in (10), with $p = 2$, $\alpha_1 = 2$, $\alpha_2 = 3$. The domain of interest is $[-2, 2]^2$;

3. a 4-d additive function, $\eta(x) = \eta_{\text{blocks}}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) + \eta_{\text{copula}}(x_{\langle 3\rangle}, x_{\langle 4\rangle})$, where η_{blocks} and η_{copula} are as in set-ups 1 and 2;
4. a 6-d copula function, the function given in (10), with $p = 6$ and $\alpha_j = 0{\cdot}1$ for all j. The domain of interest is $[-1, 1]^6$.

For all four settings, we computed the smoothing spline estimator using the full basis, and using the bases chosen by adaptive basis sampling and uniform basis sampling. For adaptive basis sampling, the number of slices was chosen based on the Scott (1992) method and, based on the asymptotic results, the dimension of the effective model space was set to be $10n^{2/9}$, so $n^* = 52$ basis functions were sampled. For a fair comparison, the same number of basis functions was used for uniform basis sampling. A thin-plate penalty was used and the penalty parameter λ was selected by minimizing the generalized crossvalidation criterion. For cases with dimension higher than two, we assumed a smoothing spline analysis of variance model with second-order interactions to deal with the curse of dimensionality. For the two bivariate set-ups, we also applied fast bivariate P-splines (Xiao et al., 2013), for which the number of interior knots for each predictor variable was set to be 11, yielding 121 interior knots in total.

To assess the estimation accuracy, we calculated the mean squared error for an estimator, which is defined as $n^{-1}\sum_{i=1}^{n}\{\hat{\eta}(x_i) - \eta(x_i)\}^2$. Figure 3 presents boxplots of the mean squared errors based on 100 runs for each set-up under three signal-to-noise ratios. For all set-ups, adaptive basis sampling provides more accurate smoothing spline estimation than uniform basis sampling. Both methods yield higher mean squared errors than the full basis smoothing spline, but this is the price paid for efficient computation with large datasets. When the signal-to-noise ratio decreases, the mean squared error for all methods gets larger and the differences among the methods diminish.

Under the two bivariate settings, adaptive basis sampling performs as well as the fast bivariate P-splines of Xiao et al. (2013) for the bivariate copula function and significantly outperforms it for the bivariate blocks function. The bivariate blocks test function is an extension of the univariate blocks function commonly used to illustrate univariate spatial adaptive smoothers (Donoho & Johnstone, 1994). However, our proposed method is not designed to achieve spatial adaptivity, which requires location-varying penalty parameters, an idea extensively studied for univariate smoothing splines (Pintore et al., 2006; Liu & Guo, 2010; Wang et al., 2013).

Table 1 summarizes the CPU times of all methods based on 100 runs using an Intel Xeon 2·90 GHz processor with 64 GB of DDR3 RAM. The computing time for the full basis smoothing spline estimator is tens or hundreds times more than that for the basis sampling methods, and for the bivariate cases, the fast bivariate P-spline is the fastest in computation.

6. Real data example

At a depth of 2890 km in the Earth, the core-mantle boundary separates turbulent flow of liquid metals in the outer core from slowly convecting, highly viscous mantle silicates. The core-mantle boundary marks the most dramatic change in dynamic processes and material properties in our planet, and accurate images of the structure at or near it over large regions are important for our understanding of the geodynamical processes and the thermo-chemical structure of the mantle and mantle-core system.

To accurately image the core-mantle boundary region, Wang et al. (2006) and Ma et al. (2007) developed a generalized Radon transform to construct raw point images, and applied the smoothing spline method to the raw images. In particular, they extracted seismic waves reflected at core-mantle boundary regions from the public data management centre of the Incorporated Research Institutions for Seismology. The seismic waves extracted were generated by around

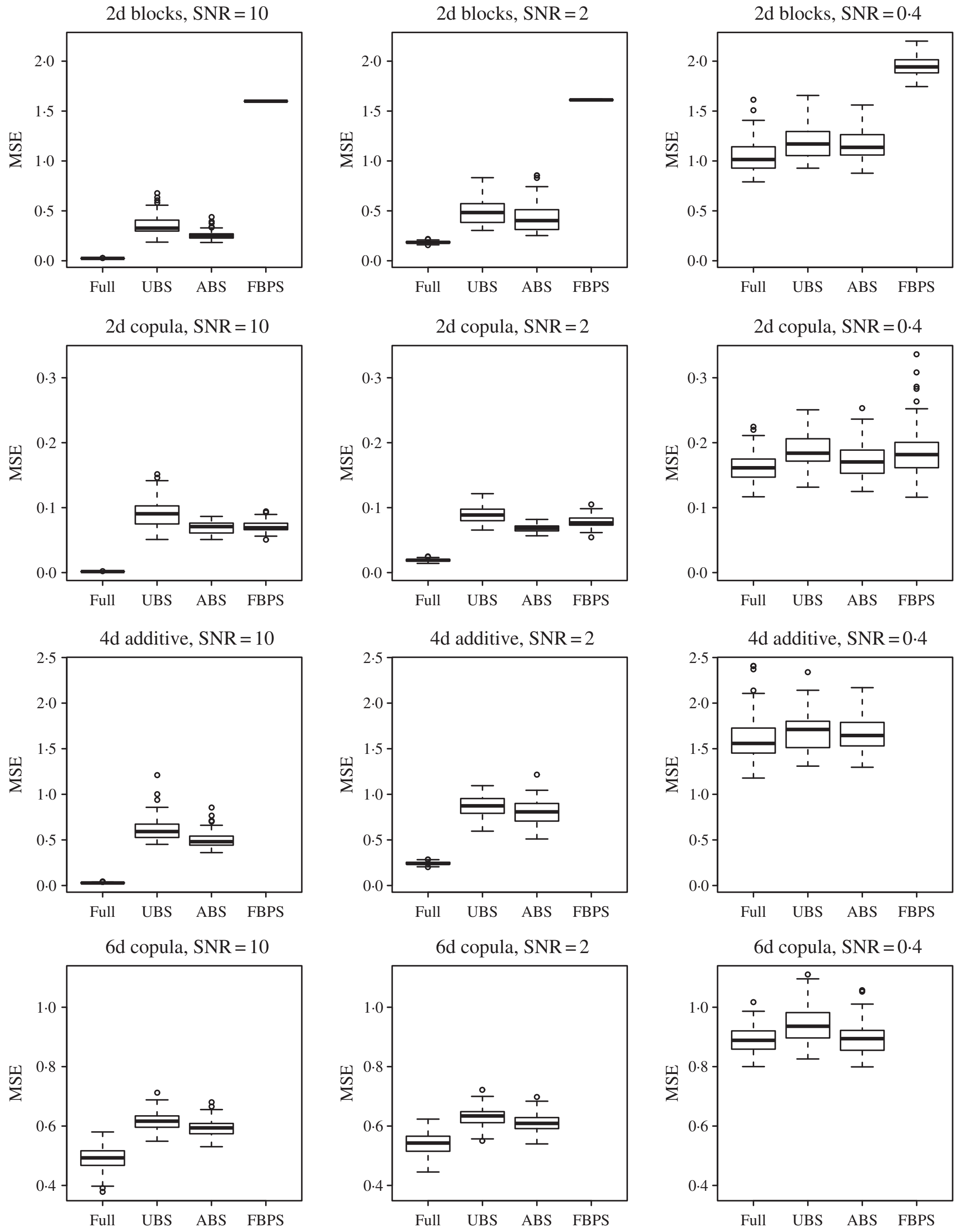

Fig. 3. Boxplots of the mean squared errors for four multivariate test functions under three signal-to-noise ratios, SNR, (10, 2, 0·4), based on 100 simulation runs. Full, UBS and ABS stand for smoothing spline estimators with full basis, uniform basis sampling and adaptive basis sampling. FBPS is fast bivariate P-splines.

1300 earthquakes with magnitude mb >5·2 that occurred between 1988 and 2002, and were recorded at one or more of a total of nearly 1200 stations in central America. Along a 2500 km strip, they then constructed point images of core-mantle boundary regions using a generalized Radon transform. They constructed 163 713 point images at various depths and locations of the

Table 1. *Means and standard errors (in parentheses) of computational time, in seconds, for four multivariate cases, based on* 100 *simulation runs*

True function	SNR	Full basis	UBS	ABS	FBPS
2d blocks	10	399 (12)	5·20 (0·12)	5·14 (0·10)	1·38 (0·03)
	2	408 (9)	7·16 (0·35)	6·40 (0·23)	1·41 (0·02)
	0·4	361 (7)	5·00 (0·17)	4·99 (0·17)	1·51 (0·02)
2d copula	10	260 (3)	6·56 (0·20)	6·41 (0·21)	1·63 (0·03)
	2	301 (6)	6·86 (0·18)	6·71 (0·33)	1·59 (0·03)
	0·4	317 (8)	4·69 (0·16)	4·79 (0·14)	1·58 (0·03)
4d blocks + copula	10	1247 (26)	15·16 (0·60)	13·84 (0·59)	–
	2	1222 (25)	16·62 (0·96)	15·54 (0·76)	–
	0·4	1135 (19)	13·16 (0·66)	13·27 (0·60)	–
6d copula	10	9336 (223)	162·88 (7·27)	145·14 (7·32)	–
	2	9572 (283)	179·12 (7·52)	181·27 (6·60)	–
	0·4	7639 (161)	143·10 (6·80)	135·01 (6·81)	–

SNR, signal-to-noise ratio; UBS, uniform basis sampling; ABS, adaptive basis sampling; FBPS, fast bivariate P-splines.

strip. At each depth and location, the point images constructed contain many noisy replicates resulting from different reflection angles of the seismic waves, so further statistical analysis is necessary to estimate the true image. In order to be computationally feasible, they estimated the true image using smoothing splines at each location and interpolated the estimated images from all locations to get the three-dimensional image. The image shows peaks of very different magnitudes at several unexpected locations (van der Hilst et al., 2007).

In this section, we apply a smoothing spline with adaptive basis sampling directly to all point images to estimate the three-dimensional image. We let y_{ij} denote the point image at the ith distance, $x_{\langle 1\rangle}$, and the jth depth, $x_{\langle 2\rangle}$. We consider the following model for the point images

$$y_{ij} = \eta(x_{\langle 1\rangle i}, x_{\langle 2\rangle j}) + \epsilon_{ij}.$$

Since the sample size is $n = 163\,713$, the regular tensor-product smoothing spline is computationally prohibitive. Instead, we apply our cubic tensor-product smoothing spline with adaptive basis sampling to the dataset with $K = 10$ slices and let the dimension of the effective model space be $n^* = 155$. Define $k_1(u) = u - 0{\cdot}5$,

$$k_2(x) = \frac{1}{2}\left\{k_1^2(x) - \frac{1}{12}\right\}, \qquad k_4(x) = \frac{1}{24}\left\{k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\right\},$$

and $R(u_1, u_2) = k_2(u_1)k_2(u_2) - k_4(|u_1 - u_2|)$. The cubic tensor-product smoothing spline estimator with adaptive basis sampling has the form

$$\eta(x) = \sum_{\nu=1}^{4} d_\nu \phi_\nu(x) + \sum_{j=1}^{n^*} c_j R_J(x_j^*, x),$$

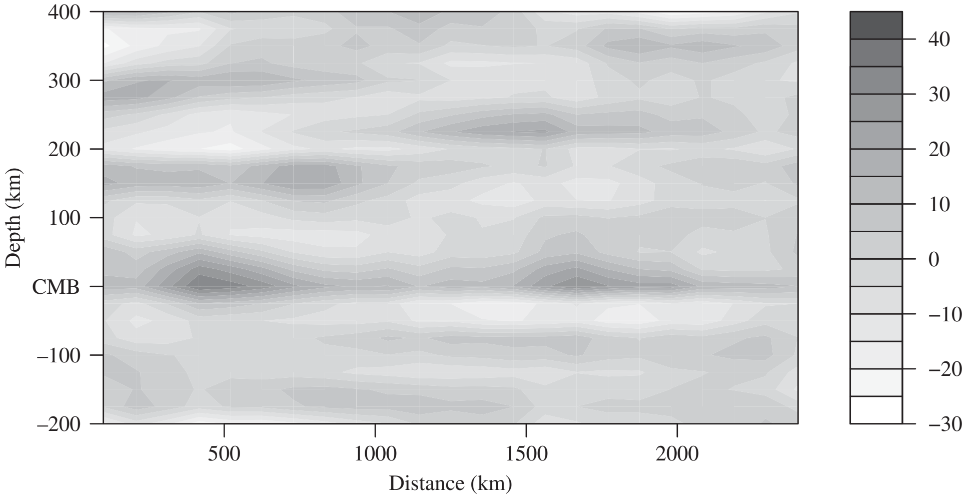

Fig. 4. The estimated image of core-mantle boundary (CMB) region structure using smoothing spline with adaptive basis sampling.

where $\phi_1(x) = 1$, $\phi_2(x) = k_1(x_{\langle 1\rangle})$, $\phi_3(x) = k_1(x_{\langle 2\rangle})$, $\phi_4(x) = k_1(x_{\langle 1\rangle})k_1(x_{\langle 2\rangle})$ and

$$\begin{aligned}R_J(x, y) = \theta R(x_{\langle 1\rangle}, y_{\langle 1\rangle}) + \theta_2 R(x_{\langle 2\rangle}, y_{\langle 2\rangle})\\ + \theta_3 R(x_{\langle 1\rangle}, y_{\langle 1\rangle})k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + \theta_4 R(x_{\langle 2\rangle}, y_{\langle 2\rangle})k_1(x_{\langle 1\rangle})k_1(y_{\langle 1\rangle})\\ + \theta_5 R(x_{\langle 1\rangle}, y_{\langle 1\rangle})R(x_{\langle 2\rangle}, y_{\langle 2\rangle}).\end{aligned}$$

The contour plot of the estimated image is provided in Fig. 4. There, we set the depth of the core-mantle boundary, 2890 km, as coordinate zero for depth. We can clearly see a peak at depth zero at all distances, which reveals that the core-mantle boundary is a major boundary. It is interesting that we see two disconnected peaks in the depth around 200 km above the core-mantle boundary: one is below and the other is above. We also calculated 95% Bayesian confidence intervals and found them to indicate that these peaks are significantly nonzero. These structures are likely to be the so-called D'' region, and have also been detected using nonparametric mixed-effect models developed in van der Hilst et al. (2007).

Acknowledgement

The first author thanks Chong Gu for many helpful discussions. Ma's work was partially supported by the National Science Foundation and the U.S. Department of Energy. Huang's work was partially supported by the National Science Foundation and King Abdullah University of Science and Technology.

Supplementary material

Supplementary material available at *Biometrika* online includes the detailed theoretical results.

References

Claeskens, G., Krivobokova, T. & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 529–44.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

Dennis, J. E. & Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, Eds. W. Schempp & K. Zeller. Berlin: Springer-Verlag, pp. 85–100.

Golub, G. & Van Loan, C. (1989). *Matrix Computations*. Baltimore, MD: The Johns Hopkins University Press, 2nd ed.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. New York: Springer, 2nd ed.

Gu, C. & Kim, Y.-J. (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist.* **30**, 619–28.

Gu, C. & Qiu, C. (1994). Penalized likelihood regression: A simple asymptotic analysis. *Statist. Sinica* **4**, 297–304.

Hastie, T. J. (1996). Pseudosplines. *J. R. Statist. Soc.* B **58**, 379–96.

Kim, Y.-J. & Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *J. R. Statist. Soc.* B **66**, 337–56.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–27.

Liu, H., Lafferty, J. & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295–28.

Liu, Z. & Guo, W. (2010). Data driven adaptive spline smoothing. *Statist. Sinica* **20**, 1143–63.

Luo, Z. & Wahba, G. (1997). Hybrid adaptive splines. *J. Am. Statist. Assoc.* **92**, 107–16.

Ma, P., Wang, P., Tenorio, L., de Hoop, M. V. & van der Hilst, R. D. (2007). Imaging of structure at and near the core-mantle boundary using a generalized Radon transform: 2. Statistical inference of singularities. *J. Geophys. Res.* **112**, B08303.

Pintore, A., Speckman, P. & Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93**, 113–25.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177–83.

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

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795–810.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–53.

Tenorio, L., Andersson, F., de Hoop, M. & Ma, P. (2011). Data analysis tools for uncertainty quantification of inverse problems. *Inverse Problems* **27**, 045001.

Utreras, F. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comp.* **2**, 349–62.

van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S. H., Ma, P. & Tenorio, L. (2007). Seismo-stratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* **315**, 1813–17.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B **45**, 133–50.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–1402.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wang, P., de Hoop, M. V., van der Hilst, R. D., Ma, P. & Tenorio, L. (2006). Imaging of structure at and near the core mantle boundary using a generalized radon transform: 1. Construction of image gathers. *J. Geophys. Res.* **111**, B12304.

Wang, X., Shen, J. & Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electron. J. Statist.* **5**, 1–17.

Wang, X., Du, P. & Shen, J. (2013). Smoothing splines with varying smoothing parameter. *Biometrika* **100**, 955–70.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Boca Raton: Chapman and Hall.

Weinberger, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. Philadelphia: SIAM.

Xiao, L., Li, Y. & Ruppert, D. (2013). Fast bivariate P-splines: The sandwich smoother. *J. R. Statist. Soc.* B **75**, 577–99.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *J. Am. Statist. Assoc.* **99**, 659–72.

[*Received December* 2013. *Revised February* 2015]