Perfect simulation for Bayesian curve and surface fitting

C C Holmes B K Mallick

Abstract

This paper demonstrates how to apply the method of coupling from the past (CFTP) (Propp and Wilson, 1996) to the problem of sampling orthogonal predictors in the Bayes linear model. Using CFTP we construct a coupled Gibbs sampler that automatically determines its own run time to convergence. This removes the need to rely on approximate convergence diagnostics as in usual Markov chain Monte Carlo methods. We apply this approach to a number of curve and surface fitting examples using wavelets and Demmler-Reinsch smoothing splines where there are as many predictors as data points. The Demmler-Reinsch splines are derived from an orthogonalization of some standard smoothing models. Somewhat remarkably, this procedure actually adds to the transparency of the model. This proves especially beneficial in a Bayesian perspective as it facilitates the quantification of prior beliefs on the model space. A striking result in all of our examples is the rapid convergence of the Gibbs sampler. Possessing this knowledge allows us to repeatedly regenerate the sampler following known convergence and hence generate independent and identically distributed samples from the posterior model space.

Keywords: Exact sampling, perfect simulation, wavelets, splines, variable selection, Markov chain Monte Carlo, Demmler-Reinsch basis.

1 Introduction

In the analysis of the Bayes linear model one is often interested in accounting for model uncertainty surrounding the use of a single set of explanatory variables, $\boldsymbol{X}$, taken from a set of p potential predictors $\boldsymbol{X}_{\text{full}} = \{X_1, \ldots, X_p\}$. This problem has received considerable attention in the Bayesian literature where Markov chain Monte Carlo (MCMC) methods have become popular when p is large and an exact analysis is not feasible (George and McCulloch, 1993; Carlin and Chib, 1995; Raftery *et al* 1997).

A recognized problem with MCMC methods is knowing when to start collecting samples. That is, the Markov chain is started from some initial distribution ρ and is simulated for a number of time steps t to obtain a new distribution ρ^t. As t increases the process converges towards the steady-state distribution of the chain π which by construction is the density of interest. Hence, a natural question to ask is "how large should t be made before ρ^t is somehow close enough to π for us to reliably use the samples generated by the simulation".

In the general case we have no clear solution to this problem. However, recently Propp and Wilson (1996) described a simulation method known as coupling from the past (CFTP) whereby the algorithm itself can determine the burn-in time for $\rho^t = \pi$ exactly. In this paper we demonstrate the application of this method to the sampling of predictors where π relates to the posterior basis probability $p(\boldsymbol{X}|Y)$ of a number of popular curve fitting models including wavelets and Demmler-Reinsch splines. A necessary condition of our analysis is that the basis is orthogonal. At first sight this might appear to be a heavy restriction, however, we demonstrate that any smoothing model can be orthogonalized and, somewhat remarkably, when this process is undertaken using the Demmler-Reinsch procedure the orthogonalization actually leads to improved interpretation in the models. This has particular benefits for our Bayesian analysis where we must quantify prior beliefs on the model space. Chapter 3 of Eubank (1999) lists some other examples of orthogonal smoothing methods not mentioned here.

A striking result in the examples we consider is the rapid convergence of the Gibbs sampler. Knowing exactly when the sampler has converged allows us the luxury of repeatedly restarting the sampler to obtain independent and identically distributed (iid) samples from the posterior model space. This is a significant improvement over current procedures where the convergence time is estimated by heuristics. Without the knowedge of convergence necessitates the need for over cautious burn-in times which typically leads to a single MCMC run being attempted, producing correlated samples, with many post convergence samples being unknowingly discarded.

In section 2 we give a brief overview of the Bayesian model uncertainty problem. Section 3 introduces wavelets and Demmler-Reinsch orthogonal basis methods and Section 4 describes coupling from the past and how it can be applied to these models. A number of applications are given in Section 5.

All of the procedures in this paper including the construction of the Demmler-Reinsch splines and the CFTP model sampling algorithm are available as Matlab code from the first author.

2 Bayesian model uncertainty

In principle the Bayesian approach to model uncertainty is simple. All information about $\boldsymbol{X}$ is contained in the posterior probability distribution

$p(\boldsymbol{X}|Y)$, where Y denotes a $n \times 1$ matrix of response variables. For the Bayes linear model,

$$
\begin{aligned}
Y &= \boldsymbol{X}\boldsymbol{\beta} + \epsilon \qquad (1)\\
p(\boldsymbol{\beta}, \epsilon) &= \phi(\boldsymbol{\beta}, \epsilon)
\end{aligned}
$$

where ϕ is some joint prior density on the regression coefficients and noise term, denoted by $\boldsymbol{\beta}$ and ϵ respectively. The marginal probability of a particular collection of predictors, say $\boldsymbol{X}_i$, is from Bayes theorem,

$$
p(\boldsymbol{X}_i|Y) = \frac{p(\boldsymbol{X}_i)\int\int p(Y|\boldsymbol{X}_i, \boldsymbol{\beta}_i, \epsilon_i)\phi(\boldsymbol{\beta}_i, \epsilon_i)d\boldsymbol{\beta}_i d\epsilon_i}{\sum_{j=1}^{M} p(\boldsymbol{X}_j)\int\int p(Y|\boldsymbol{X}_j, \boldsymbol{\beta}_j, \epsilon_j)\phi(\boldsymbol{\beta}_j, \epsilon_j)d\boldsymbol{\beta}_j d\epsilon_j} \qquad (2)
$$

where M is the cardinality of the state space $\mathcal{S}_{\boldsymbol{X}}$ containing all possible combinations of predictors from $\boldsymbol{X}_{\text{full}}$, i.e. $\mathcal{S}_{\boldsymbol{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M\}$ and $M = 2^p - 1$, if we assume that at least one predictor is always included.

The computation of the marginal probability (2) is much simplified if the noise term is assumed to be normal with mean zero and unknown variance σ^2 and a conjugate prior is adopted for $p(\boldsymbol{\beta}_i, \sigma_i^2)$ which in the case of the linear model is the normal-inverse-gamma distribution, $p(\boldsymbol{\beta}_i, \sigma_i^2) = NIg(\boldsymbol{\beta}_i, \sigma_i^2|s_i, d_i, \boldsymbol{m}_i, \boldsymbol{V}_i)$ (O'Hagan, 1997) and the integral can then be expressed analytically as,

$$
\int\int p(Y|\boldsymbol{X}_i, \boldsymbol{\beta}_i, \sigma_i^2)p(\boldsymbol{\beta}_i, \sigma_i^2)d\boldsymbol{\beta}_i d\sigma_i^2 = \frac{|\boldsymbol{V}_i^*|^{1/2} s_i^{d/2}\Gamma(d_i^*/2)}{|\boldsymbol{V}_i|^{1/2}\pi^{n/2}\Gamma(d_i/2)}(s_i^*)^{-d_i^*/2} \qquad (3)
$$

where $|\cdot|$ refers to the determinant of a matrix and s_i^*, d_i^*, $\boldsymbol{V}_i^*$ refer to the parameters of the posterior distribution $p(\boldsymbol{\beta}_i, \sigma_i^2|Y, \boldsymbol{X}_i) = NIg(\boldsymbol{\beta}_i, \sigma_i^2|s_i^*, d_i^*, \boldsymbol{m}_i^*, \boldsymbol{V}_i^*)$

where,

$$\begin{aligned}
\boldsymbol{V}_i^* &= (\boldsymbol{X}_i'\boldsymbol{X}_i + \boldsymbol{V}_i^{-1})^{-1} \qquad (4) \\
\boldsymbol{m}_i^* &= \boldsymbol{V}_i^*(\boldsymbol{V}_i^{-1}\boldsymbol{m}_i + \boldsymbol{X}_i'Y) \\
s_i^* &= s_i + \boldsymbol{m}_i'\boldsymbol{V}_i^{-1}\boldsymbol{m}_i + Y'Y - (\boldsymbol{m}_i^*)'(\boldsymbol{V}_i^*)^{-1}\boldsymbol{m}_i^* \\
d_i^* &= d + n.
\end{aligned}$$

The calculation of the expression in (3) is of order $O(p_i^3)$ where $p_i = \dim(\boldsymbol{X}_i)$ and hence for many linear regression problems where the number of predictors is small the exact elicitation of the posterior probability density (2) is feasible.

Our interest here is in Bayesian nonparametric curve fitting models, such as smoothing splines, radial basis functions and wavelets, which are also of form (1), but where the columns of $\boldsymbol{X}$ relate to the output of a set of basis functions. In these models it is typical to have as many basis functions as there are data points, with $p = n$, and hence for most applications the calculation of (2) is too demanding.

One solution to this problem is to approximate $p(\boldsymbol{X}|Y)$ with samples drawn from $p(\boldsymbol{X}|Y)$. This can be achieved using Markov chain Monte Carlo (MCMC) methods whereby a Markov chain is simulated whose steady-state distribution π matches the density of interest, i.e. $\pi = p(\boldsymbol{X}|Y)$. A number of papers have utilized this approach (Smith and Kohn, 1996; Denison *et al* 1998; Holmes and Mallick, 1998; Clyde *et al*, 1998). However, a problem with all of these methods is that the convergence of the Markov chain cannot be assured. This paper aims to address this issue for the important and special case of orthogonal models.

As a further motivation for this study, we note that accounting for model uncertainty is especially important for curve fitting models where the selection of a salient base is known to be critical to the performance of the method. The choice of the basis set controls spatial adaption and flexibility in the final model.

3 Orthogonal smoothing methods

Many of the common smoothing methods for curve fitting are written in basis form,

$$\hat{y}_i = \sum_{j=1}^{p} beta_j b_j(x_i) \tag{5}$$

where $\hat{y}_i$ is the fitted value at x_i and $b_j(x)$ is some non-linear basis function of x. From here on we shall use the notation $\boldsymbol{B}$ to denote the design matrix of basis responses in order to highlight the fact that we are dealing with non-linear transformations of the original data,

$$\boldsymbol{B}_{\text{full}} = \begin{pmatrix} b_1(x_1) & \cdots & b_p(x_1) \\ b_1(x_2) & \cdots & b_p(x_2) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_p(x_n) \end{pmatrix}, \tag{6}$$

An important subclass of basis models are those where the base is orthogonal, i.e. $\boldsymbol{B}'_{\text{full}}\boldsymbol{B}_{\text{full}} = \text{diag}(d_1, \ldots, d_p)$. Chapter 3 in Eubank (1999) reviews some of these models. In this paper we shall concentrate on two particular forms for $b(x)$ namely as wavelets (Bruce and Gao, 1996) and Demmler-Reinsch splines (Nychka, 1999). We will first briefly overview wavelets before concentrating on the lesser known Demmler-Reinsch splines.

3.1 Wavelets

Wavelet methods have become increasingly popular in statistics in recent years. The basis functions relate to translations and dilations of some parent constructor basis with as many bases as there are data points so $p = n$. Fig. 1 shows the response of four so called "Least Asymmetric" LA(8) wavelet bases for a 64 point data set equally spaced on $(0, 1)$. The titles to the plots in Fig. 1 indicate the index of the basis. Higher indexed bases capture finer details in the data and the bases are orthonormal by design so that $\boldsymbol{B}'\boldsymbol{B} = I$. Bruce and Gao (1996) provide a detailed description of the wavelet model and we refer the interested reader to this reference for further information.

3.2 Demmler-Reinsch splines

The Demmler-Reinsch (DR) basis (Demmler and Reinsch, 1975) provides a canonical representation for a number of smoothing methods including the 1d cubic spline (Eubank, 1999). The 1d DR base is derived from an orthogonalization of the standard cubic spline model and does not rely on equally spaced data as for the wavelet method.

Typically, orthogonalising non-orthogonal models for the purpose of selecting bases causes severe problems in terms of interpretation. However, for the smoothing splines the orthogonalization process actually makes the resulting model more transparent. This has been noted by Eubank (1999), Hastie (1996) and Nychka (1999) amongst others. To derive the Demmler-Reinsch representation we first write the cubic spline model in smoothing matrix form, following Green and Silverman (Section 2.3, 1994),

$$\begin{aligned} \hat{Y} &= (I + \alpha \boldsymbol{K})^{-1} Y \\ &= \boldsymbol{S} Y \end{aligned} \tag{7}$$

where $\boldsymbol{S}$ is an $n \times n$ smoothing matrix that maps the responses Y to their predicted values $\hat{Y}$ under the model. In (7) the variable α is a smoothing parameter that controls the trade off between fidelity to the data and roughness of the fitted curve.

The Demmler-Reinsch representation is formed from the eigendecomposition of $\boldsymbol{S}$. Since $\boldsymbol{S}$ is square symmetric, it has a decomposition

$$\boldsymbol{S} = \boldsymbol{B}\boldsymbol{\theta}\boldsymbol{B}' \tag{8}$$

where the columns of $\boldsymbol{B}$ form an orthonormal Demmler-Reinsch base and $\boldsymbol{\theta}$ is diagonal with eigenvalues $\theta_i \in (0,1]$ and decreasing in i. Clearly there are as many basis functions as there are data points.

Fig. 2 shows the response and index of four Demmler-Reinsch bases using the same data as in the wavelet example. Demmler and Reinsch (1975) showed that for $j \geq 3$, $b_j(x)$ has $j-1$ zero crossings. Hence, in a similar fashion to wavelets, the higher indexed bases capture finer scale details in the data, though now the bases are not localized. The cubic spline can now be written in Demmler-Reinsch form as,

$$\hat{y}_i = \sum_{j=1}^{p} \tilde{\beta}_j b_j(x_i) \tag{9}$$

where b_j is the jth DR basis and $\tilde{\beta}$ are generalized ridge type estimators

$$\tilde{\beta}_j \quad = \quad \theta_j \hat{\beta}_j \tag{10}$$

where $\hat{\beta}_j = \sum_{i=1}^{n} b_j(x_i) y_i$ is the least squares (LS) estimate of β_j. The estimator (10) shrinks the LS estimate by an amount determined by the eigenvalue θ_j associated with the jth basis. The extent of the shrinkage is determined by the smoothing parameter α. Fig. 3 shows the values of θ_j for two settings of the smoothing parameter $\alpha = 10^{-5}$ and $\alpha = 10^{-6}$. The use of the DR representation clearly highlights how smoothing is achieved under the cubic spline model. That is, by damping down the higher order DR bases with the extent of the tapering controlled by the smoothing parameter α.

Further interpretation is achieved by noting that the eigenvalues θ_j are derived from the smoothing parameter α and an eigendecomposition of $\boldsymbol{K}$ in (7),

$$\theta_j = \frac{1}{1 + \alpha d_j} \tag{11}$$

where d_j is the eigenvalue associated with the jth eigenvector of $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$ where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_p)$, which is independent of α. The eigenvalues $(d_1, \ldots, d_p)$ for the data set in this example are plotted in Fig. 4

The standard cubic spline is generalized to higher dimensions by the class of radial basis functions (Powell 1987, 1992). These also have a Demmler-Reinsch representation that aids in their interpretation (Nychka 1999). We mention here the thin-plate spline model for two dimensional spatial smoothing though further extensions are straightforward.

The conventional thin-plate spline model is written in basis form as

$$\begin{aligned} \hat{y}_i &= \sum_{t=1}^{q} \delta_t \eta_t(x_i) + \sum_{j=1}^{k} \tau_j g_j(x_i) \qquad (12) \\ g_j(x_i) &= ||x_i - x_j||^2 \log(||x_i - x_j||) \end{aligned}$$

where $\eta(\cdot)$ are low order polynomial functions and $||\cdot||$ denotes the Euclidean distance metric. In matrix form,

$$\hat{Y} = \boldsymbol{T}\boldsymbol{\delta} + \boldsymbol{M}\boldsymbol{\tau} \tag{13}$$

where $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$ are regression coefficients, $\boldsymbol{T} = \{\eta_j(x_i)\} = \{1, X_1, X_2\}$ denotes the $n \times 3$ matrix of linear terms and $\boldsymbol{M} = \{g_j(x_i)\}$ denotes the $n \times n$ matrix of basis function responses. A constraint $\boldsymbol{T}'\boldsymbol{\tau} = 0$ is imposed to ensure identifiability and linear boundary effects. Similar to the cubic spline, the thin-plate spline trades of roughness of the surface, in terms of some integrated squared derivative, against fidelity to the data. This trade off is controlled by the smoothing parameter α acting on a roughness penalty which is simply $\boldsymbol{\tau}'\boldsymbol{M}\boldsymbol{\tau}$. The non-Bayesian estimate of $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ then follows,

$$\begin{aligned} \hat{\boldsymbol{\delta}} &= (\boldsymbol{T}'(\boldsymbol{M} + \alpha I)^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'(\boldsymbol{M} + \alpha)^{-1}Y \\ \hat{\boldsymbol{\tau}} &= (\boldsymbol{M} + \alpha I)^{-1}(Y - \boldsymbol{T}\boldsymbol{\delta}). \end{aligned} \tag{14}$$

To construct the Demmler-Reinsch spline we intially re-parameterize the model into its 'natural' base, where there are as many basis functions as there are data points, before orthogonalising the model with respect to the new basis and roughness penalty.

First, let $\boldsymbol{F}$ denote the orthonormal matrix derived from the QR decomposition of $\boldsymbol{T}$, i.e. $\boldsymbol{T} = \boldsymbol{F}\boldsymbol{R}$, where $\boldsymbol{R}$ is upper triangular. Now, $\boldsymbol{F}$ can be partitioned into $\boldsymbol{F} = [\boldsymbol{F}_1|\boldsymbol{F}_2]$ where the columns of $\boldsymbol{F}_1$ span the column space of $\boldsymbol{T}$ and $\boldsymbol{F}_2$ has columns that span the space orthogonal to the columns of $\boldsymbol{T}$. The model (13) can then be re-parameterized enforcing the necessary linear constraints,

$$\begin{aligned} \hat{Y} &= \boldsymbol{G}\boldsymbol{\omega} \\ \boldsymbol{G} &= [\boldsymbol{T}|\boldsymbol{M}\boldsymbol{F}_2] \end{aligned} \tag{15}$$

where $\boldsymbol{G}$ is the $n \times n$ matrix of 'natural' basis and $\boldsymbol{\omega}$ are the new regression coefficients. The roughness penalty for this new basis space is $\boldsymbol{\omega}'\boldsymbol{H}\boldsymbol{\omega}$ where

$$\boldsymbol{H} = \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{F}_2'\boldsymbol{M}\boldsymbol{F}_2 \end{pmatrix}, \tag{16}$$

and the zero elements indicate no penalty is placed on the linear terms. The Demmler-Reinsch base is then derived from an orthogonalization of this 'natural' thin-plate spline model where the orthogonalization diagonalises both the basis set and the roughness penalty. In order to achieve this, let $\boldsymbol{A}$ denote the inverse symmetric square root of $(\boldsymbol{G}'\boldsymbol{G})$, i.e. $(\boldsymbol{A}'\boldsymbol{A}) = (\boldsymbol{G}'\boldsymbol{G})^{-1}$ and let $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'$ be the singular value decomposition of $\boldsymbol{A}\boldsymbol{H}\boldsymbol{A}$. Then the Demmler-Reinsch basis $\boldsymbol{B}$ is constructed by setting $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{U}$ with,

$$\boldsymbol{B} = \boldsymbol{G}\boldsymbol{P}, \tag{17}$$

and it is straightforward to verify that $\boldsymbol{P}'(\boldsymbol{G}'\boldsymbol{G})\boldsymbol{P} = I$ and $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{D}$. The Demmler-Reinsch representation is an orthonormal base that spans the same subspace as the original model.

In a similar fashion to the cubic spline, the 2d thin-plate spline DR basis has a natural ordering, using the eigenvalues in $\boldsymbol{D}$, with higher indexed bases capturing finer detail in the surface. Fig. 5 illustrates several of the 2d Demmler-Reinsch bases on a uniform grid on $[0,1]_2$ and Fig. 6 shows the corresponding contour shading plots. The non-Bayesian thin-plate spline with smoothing parameter α is then written as

$$\begin{aligned} \hat{y}_i &= \sum_{j=1}^{p} \tilde{\beta}_j b_j(x_i) \\ \tilde{\beta}_j &= \frac{\hat{\beta}_j}{1 + \alpha d_j} \end{aligned} \tag{18}$$

where b_j relates to the jth column of $\boldsymbol{B}$, d_j is the jth diagonal element of $\boldsymbol{D}$ and $\hat{\beta}_j$ is the LS estimate.

The improved transparency gained from adopting the DR form has important implications in a Bayesian analysis as it aids in the quantification of beliefs about the prior probability of a basis function being included in the model. These issues are discussed in section 4. It is also worth noting that the Demmler-Reinsch form holds added benefits from a computational viewpoint, which is also important for non-Bayesian spline smoothing. The matrix inversions in the DR method only need to be performed once and for non-Bayesians the selection of α, say by, generalised cross-validation is then rapid to compute (Nychka, 1999).

3.3 Bayesian Analysis

The Wavelet and DR models are orthonormal which results in further simplification of the marginal likelihood (3). To aid in the analysis that follows we introduce a basis indicator vector $\boldsymbol{\gamma}_i = \{\gamma_{i1}, \ldots, \gamma_{ip}\}$ associated with the ith basis configuration $\boldsymbol{B}_i \in \mathcal{S}_{\boldsymbol{B}} = \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_M\}$. where $M = 2^n$ and $\gamma_{ij} = 1$ if the jth basis function from $\boldsymbol{B}_{\text{full}}$ is included in $\boldsymbol{B}_i$ and $\gamma_{ij} = 0$ otherwise. We will use the notation p_i to denote the number of basis functions in $\boldsymbol{B}_i$ and hence $p_i = \sum_{j=1}^{p} \gamma_{ij} = \dim(\boldsymbol{B}_i)$.

We will adopt the following prior for the within model parameters throughout,

$$p(\boldsymbol{\beta}_i, \sigma_i^2) = NIg(\beta_i, \sigma_i^2 | s, d, 0, vI_{p_i}), \tag{19}$$

where I_a denotes the $a \times a$ identity matrix and s, d and v are common to all models in the state space $\mathcal{S}_{\boldsymbol{B}}$.

The marginal likelihood (3) for a particular basis set is then,

$$p(Y|\boldsymbol{B}_i) = C(1+v)^{-p_i}(s + Y'Y - \sum_{j=1}^{p} \gamma_{ij} SS_j)^{-\frac{d+n}{2}} \tag{20}$$

where C is a constant that does not depend on $\boldsymbol{B}_i$ and

$$\begin{aligned} SS_j &= (1+v^{-1})^{-1}\hat{\beta}_j^2 \\ &= (1+v^{-1})^{-1}\left(\sum_{i=1}^{n} b_j(x_i)y_i\right)^2, \end{aligned} \tag{21}$$

where, as before, $\hat{\beta}_j$ denotes the LS estimate.

4 Coupling Markov chains from the past

The paper by Propp and Wilson (1996) on coupling from the past has generated considerable interest in the statistics community. For our purposes, CFTP can be viewed as a protocol for MCMC sampling such that the algorithm itself determines the necessary 'burn-in' time to equilibrium. This is an important contribution to MCMC methods as we are often left with little or no indication of how close we are to the steady-state distribution of the chain. For a review of various convergence diagnostics see Brooks and Roberts (1999).

We are interested in MCMC methods for sampling the basis set in the curve fitting models described in Section 3. The indicator vector $\boldsymbol{\gamma}$ provides a convenient vehicle for constructing a MCMC algorithm to draw samples from $p(\boldsymbol{B}|Y)$. To be precise, we can construct a cyclic Gibbs sampler (Gelfand and Smith, 1990) that iterates through the elements of the indicator vector $\boldsymbol{\gamma}$ drawing each element from its posterior distribution conditional on the values of the other elements in the vector. The jth element of $\boldsymbol{\gamma}$ determines whether the j basis b_j is included in the model and matching the prior densities $p(\boldsymbol{\gamma}) = p(\boldsymbol{B})$ gives $p(\boldsymbol{B}|Y) = p(\boldsymbol{\gamma}|Y)$. Hence the samples of $\boldsymbol{\gamma}$ equate to samples from the posterior model space.

The standard procedure is to run the Gibbs sampler for a number of time steps until some heuristic measure of convergence is met and then start collecting samples, assuming convergence. CFTP can provide an exact indicator of when convergence has occurred. To illustrate how CFTP can be applied to the basis sampling problem we present a simple example.

Consider the problem of sampling models from a three variable linear regression problem, so that $p = 3$ and there are $M = 2^3 = 8$ possible models or states of $\boldsymbol{\gamma}$ with state space $\mathcal{S}_{\boldsymbol{\gamma}} = \{(0,0,0), (0,0,1), \ldots (1,1,1)\}$. Suppose we simulate a separate Gibbs sampler starting from each of the 8 possible states $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_8\}$. Moreover, suppose conceptually that the samplers are started at some time $t = -N$ in the past and run up to time $t = 0$, where one time step denotes one complete sweep through the p elements of each indicator vector.

At each time step and for each element the Gibbs sampler requires one draw of a random variable which we denote by R, where $R \sim U(0,1)$. Following the draw the value assigned to the element γ_{ij} by the Gibbs sampler is a deterministic function, say ϕ, of R and the cumulative density $F_{\gamma_{ij}}$, where $F_{\gamma_{ij}}$ is conditioned on all the elements of $\boldsymbol{\gamma}_i$ other than γ_{ij}. Fig. 7 illustrates the deterministic nature of the Gibbs sampler conditional on R. Clearly the same value for R will result in the same value being assigned to γ_{ij}.

We now assume that the 8 Gibbs samplers in our example are "coupled" in that at each time step and for each element, all 8 Gibbs samplers will update using a single draw of $R \sim U(0,1)$. That is, at each time step the value of the jth element of γ_{ij} is determined by $\gamma_{ij} = \phi(F_{\gamma_{ij}}, R)$ for $i = 1, \ldots, 8$. Note that a new draw of R is made for each element.

At some point in the simulation two chains may come to coalesce, meaning that they come to occupy the same state, i.e $\gamma_i = \gamma_k$ for some $i \neq k$. When this happens the two chains will remain together from then on, as they are using the same draws of R, and clearly the cumulative densities $F_{\gamma_{ij}}$ are equivalent. Moreover, if at time $t = 0$ we find that all the chains have coalesced to a single state then that single state is an exact sample from $p(\boldsymbol{\gamma}|Y)$. This can be seen by considering a single Markov chain simulated from time $t = -\infty$. At time $t = -N$ the chain must be in one of the 8 possible states, though we don't know which one. However all possible states at time $t = -N$ led to the single state we obtained at time $t = 0$. Therefore the chain has completely forgotten its starting distribution and forms a single sample from the steady-state distribution of the Markov chain which by construction is $p(\boldsymbol{\gamma}|Y)$.

If we find at $t = 0$ that the chains have not all coalesced then we just need to go back further, to say $t = -2N$, and try again. However, we must use the same sequence of random draws of R as before from $t = -N$ to $t = 0$. For computer simulations this can be achieved by resetting the seed in the pseudo-random number generator when $t = -N$. This process of going back further and reusing the random variables continues until we find that all chains have coalesced. In this way the algorithm determines the necessary run time to convergence.

If this was all that could be said about the procedure then not much is gained since we needed to run the sampler simultaneously from every state in $\mathcal{S}_{\boldsymbol{\gamma}}$. However, Propp and Wilson (1996) showed that the computational burden can be reduced to that of simulating just two chains if the following sufficient conditions are met.

1. There exists a partial ordering $\leq$ on the state space $\mathcal{S}_{\boldsymbol{\gamma}}$, and a unique

minimum state $\hat{0} \in \mathcal{S}_{\boldsymbol{\gamma}}$ and maximum state $\hat{1} \in \mathcal{S}_{\boldsymbol{\gamma}}$ such that for all $\boldsymbol{\gamma} \in \mathcal{S}_{\boldsymbol{\gamma}}$, $\hat{0} \leq \boldsymbol{\gamma} \leq \hat{1}$, and;

2. Under this partial ordering the Gibbs sampler is monotonic such that $\boldsymbol{\gamma}_i \leq \boldsymbol{\gamma}_k$ implies $\phi(F_{\gamma_{ij}}, \cdot) \leq \phi(F_{\gamma_{kj}}, \cdot)$

In this case we need only simulate two coupled chains starting from the upper and lower states. If these two chains have coalesced at time $t = 0$ then, from conditions (1) and (2) above, every other chain must have coalesced as well. This dramatic reduction in the computation makes CFTP possible even for large state spaces.

4.1 Algorithm for Perfect Sampling

We now define a partial ordering on the orthogonal model space with $\boldsymbol{\gamma}_i \leq \boldsymbol{\gamma}_k$ iff $\gamma_{ij} \leq \gamma_{kj}$ for $j = 1, \ldots, p$ and hence there is a unique lower state $\hat{0} = \{0_1, \ldots, 0_p\}$, no bases in the model, and unique upper state $\hat{1} = \{1_1, \ldots, 1_p\}$, all bases in the model.

If we assume independent priors on the elements of $\boldsymbol{\gamma}$,

$$p(\boldsymbol{\gamma}_i) = \prod_{i=1}^{p} p(\gamma_{ij}). \tag{22}$$

then it is straightforward to see from the form of (20) that the Gibbs sampler is monotonic under this partial ordering.

Conditions (1) and (2) from Section 4 above are met and the following algorithm determines the run time to convergence drawing a perfect sample from $p(\boldsymbol{B}|Y)$ at the end.

- Initialize $\boldsymbol{\gamma}_{\hat{1}} = \{1_1, \ldots, 1_p\}$ and $\boldsymbol{\gamma}_{\hat{0}} = \{0_1, \ldots, 0_p\}$.
- Initialize a dynamic matrix R_{store} to hold random draws of R.

- Set $N = 2$.
- Repeat
  - For $t = -N$ to 0
    - For $j = 1$ to p do Gibbs sampler:
      - If been here before use stored R value: $R = R_{\text{store}}(1-t, j)$
      - Else draw $R \sim U(0,1)$ and store value: $R_{\text{store}}(1-t, j) = R$.
      - Update: $\gamma_{\hat{0}j} = \phi(F_{\gamma_{\hat{0}j}}, R)$
      - Update: $\gamma_{\hat{1}j} = \phi(F_{\gamma_{\hat{1}j}}, R)$
    - End For j loop
  - End For t loop
  - $N = 2 \times N$
- While $\gamma_{\hat{1}} \neq \gamma_{\hat{0}}$.
- Return $\gamma_{\hat{1}}$ as a perfect sample
- End.

For model sampling, there are two possibilities at this point. Either a single converged Gibbs sampler can be continued from this point on or we can take just one sample and restart the algorithm thus obtaining another exact sample and in this manner build up a collection of *iid* samples from the model space. Clearly the latter procedure is more attractive though its application will depend on computational considerations and typical times to coalesce. In the next section we demonstrate that the Gibbs sampler shows remarkably fast convergence times for the orthogonal smoothing models we consider and hence *iid* sampling is possible.

5 Examples

This section presents applications of the CFTP methodology on a number of curve and surface fitting examples. In all of the simulations we set the prior parameters in (19) to be $v = 1000$, $s = 1$, $d = 1$. Priors on the basis space $p(\boldsymbol{B})$ are discussed below.

5.1 Wavelet denoising of Heavisine data set

We first analyze a wavelet decomposition for the Heavisine function proposed by Donoho and Johnstone (1995). Following the arguments developed in Holmes and Mallick (1999) and Holmes and Denison (1999) we choose to adopt a model prior that penalizes the degrees of freedom so that

$$p(\gamma_i) \propto \exp\left(-\frac{c}{1+v^{-1}}\sum_{j=1}^{p}\gamma_{ij}\right) \tag{23}$$

where the setting of the constant c can be aided by the calibration with classical model choice criteria. For example, $c = 1$ calibrates the posterior model probabilities with Mallows C_p statistic (Mallows, 1978), or $c = \log n$ calibrates to the Risk Inflation Criteria (Foster and George, 1995). See also George and Foster (1997). We note that this choice of model complexity prior is not critical to the CFTP sampling method.

As in Donoho and Johnstone (1995) we constructed a data set of 1024 evenly spaced points by corrupting the true Heavisine function with $N(0,1)$ noise which gives a signal to noise ratio of 7. The true function and the corrupted data are shown in the top graph of Fig. 8. We then drew 1000 perfect samples of $\boldsymbol{\gamma}$ using the algorithm described in Section 4.1. with two settings of the prior parameter, $c = 1$ and $c = \log(n)$.

Fig. 9 shows the difference between the upper and lower chains after each time step in the Gibbs sampler for 500 separate runs. Clearly the sampler converges remarkably quickly. The speed of convergence makes it computationally feasible to collect a single sample following known convergence and then restart the algorithm to generate another independent perfect sample. In this way we can generate *iid* samples from the posterior model space.

We generated 1000 posterior samples for the two settings of the prior parameter. An estimate of the posterior probability of inclusion for a basis function can be given as the average number of times the basis function appears in the posterior samples. This is shown in Fig. 10 for $c = 1$ and $c = \log(b)$ alongside the prior probability of inclusion given by the dashed lines (the prior for $c = \log(1024)$ is not visible as $p(\gamma_{ij} = 1 | c = \log(1024)) \approx 0.001$). Note that even thought the prior probability of inclusion is very low the posterior probability of the broad scale, low indexed, bases is high. The posterior mean approximation computed as the average curve from the 1000 samples is shown as the solid curves in Fig. 11, for the two prior settings, alongside the true curve (dotted). The C_p prior ($c = 1$) appears to let some fine scale noise through.

The flat priors used in this example are perhaps too naive as they fail to take into account our prior knowledge about the form of the basis functions. A more intuitive prior might place decreasing mass on the higher indexed bases. This is considered in the next example for the Demmler-Reinsch basis, though similar priors can be adopted for the wavelet problem (see for example, Clyde *et al*, 1998).

5.2 Demmler-Reinsch smoothing of unequally spaced data

We now consider a scatter plot smoothing problem with irregularly spaced data. Data was generated by sampling a linear plus Gaussian function and then adding normal noise with variance 0.2. The data is shown in Fig. 12 along with the true generating curve.

A question arises as to the appropriate prior for the Demmler-Reinsch base $p(\boldsymbol{B})$ or equivalently $p(\boldsymbol{\gamma})$. Clearly, knowledge of both the form of the bases shown in Fig. 2 and of the damping nature of the standard cubic spline fit has altered our prior beliefs about the probability of a basis being included in the model. It therefore seems appropriate to place decreasing mass on increasing index. This is one advantage from a Bayesian perspective of adopting the Demmler-Reinsch representation as the basis functions are much more interpretable.

Here is one such distribution, though by no means the only one that could be thought of. We make use of the eigendecomposition of $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$ in (7) and propose independent priors on the bases as,

$$p(b_j|\alpha) = 0.5 \times \frac{1}{1 + \alpha d_j} \tag{24}$$

where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_p)$. That is, we use the damping curves shown in Fig. 4 to guide our prior probabilities for inclusion of a basis. The scaling factor 0.5 in (24) quantifies prior indifference to inclusion of the first few bases and then progressively penalizes higher order bases. Fig. 13 shows the log prior of inclusion for various settings of α on this curve fitting problem. The exact setting of α can be aided by examining curves drawn from the prior. In Fig. 14 we illustrate the expected cubic spline curves taken by averaging 1,000 curves drawn from the prior for the settings shown in Fig 13, namely $\alpha = \{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}\}$. Naturally, the higher the value of α the smoother the prior curve is as there is lower probability on higher ordered bases entering into the model.

The CFTP algorithm was then used to assess convergence of the Gibbs sampler. As for the wavelet example, the sampler converges remarkably quickly. In all of our simulations convergence was achieved in under 8 sweeps of the sampler. A plot of the difference, during simulation time, between the upper and lower chains for 500 separate runs of the algorithm is shown in Fig. 15.

Knowing the speed allows us to regenerate the chain after convergence and take a single sample at a time. In this manner we draw 1,000 samples from the posterior model space using the four settings of the prior parameter shown in Fig. 13. We then use the samples to summarize the posterior probability of a basis being included in the model by averaging the number of times it appeared in the posterior samples. This is plotted in Fig. 16 alongside the prior which is plotted as a dashed line. The graph is truncated to the first 50 bases as none of the samples contained basis greater than this number. Two things are apparent: First, the posterior inclusion of a basis is non-monotonic e.g. the posterior probability of inclusion for the 5th basis is more common than the 4th basis under all of the priors considered; Second, the data can overwhelm the prior for some of the lower level basis where the likelihood is large, e.g. the 7th and 9th bases for the top two graphs in Fig. 16 have small prior mass but large posterior probability.

Finally we can see the effect of α by plotting the posterior expected curves using the data in Fig. 17 for the four settings. This is plotted in Fig. 17 where the solid line refers to $\alpha = 10^{-7}$; dashed $\alpha = 10^{-5}$; dotdash $\alpha = 10^{-3}$, and; dot $\alpha = 10^{-1}$. Even though the prior curves shown in Fig. 14 differ widely the posterior curves are similar. This demonstrates, for this example, that the analysis is reasonably insensitive to a large variation in the prior parameter $\alpha \in (10^{-1}, 10^{-7})$.

5.3 Surface smoothing

We generated a data set in 2d using the following function

$$f(x_1, x_2) = 1.9\{1.35 + e^{x_1} \sin[13(x_1 - 0.6)^2] e^{-x_2} \sin(7x_2)\} \tag{25}$$

which is taken from Hwang *et al* (1994). A data set was then constructed by drawing 225 points randomly on the unit square and the response, y, was calculated as $y = f(x_{1i}, x_{2i}) + \epsilon_i$ where $f()$ is the true function and ϵ_i is Gaussian noise drawn from $N(0, 0.25^2)$. The data set and true function are shown in Fig. 18.

We constructed the Demmler-Reinsch base and as before we used the eigenvalues to help determine the prior probability of a basis inclusion. The prior for 4 settings of α is shown in Fig. 19 and we can generate from the prior in order to help in the calibration. The expected surface taken by averaging 100 samples from the prior for the four settings is shown in Fig. 20.

The perfect simulation algorithm was run and 1,000 *iid* samples for each setting of α in Fig. 19 were drawn. All of the simulations coalesced in under 8 sweeps. The posterior probability of inclusion taken from the samples is shown in Fig. 21, alongside the prior shown as the dashed lines. In Fig. 22 we show the mean posterior surfaces under each of the priors. They are all very similar and close to the true function with average squared error to the truth, calculated on a 50×50 grid, as 0.058, 0.049, 0.033 and 0.031 for the prior settings $\alpha = \{10^1, 10^{-1}, 10^{-3}, 10^{-5}\}$ respectively. The 95% credible surfaces for $\alpha = 10^{-5}$, alongside the residual surface is shown in Fig. 23.

5.4 Ozone data set

We complete the examples section with a real data set. The ozone data found in S-plus records the average levels of ozone at 20 stations in the Chicago urban area for the summer of 1987. The data is plotted in Fig. 24.

We constructed the Demmler-Reinsch basis and as in the other examples we used the eigenvalues to determine the prior probability of inclusion. Fig. 25 shows the probability of inclusion for a number of settings of α and Fig. 26 illustrates expected surfaces drawn from the prior using the settings $\alpha = \{10^5, 10^3, 10^1, 10^{-1}\}$.

We then drew 1,000 *iid* samples from the posterior model space using the four settings for α. Once more the Gibbs sampler typically converged within 4 sweeps and at most took 8 iterations. The posterior probability of inclusion using these prior settings is shown in Fig. 27 where the prior probability is given by the dashed line. The posterior mean surfaces are shown in Fig. 28 where we observe that the prior setting $\alpha = \{10^5, 10^3\}$ produces nearly linear surfaces. In Fig. 29 we illustrate the posterior mean surface as well as a contour plot for the model with $\alpha = 10^1$. Finally, the 95% credible surfaces with $\alpha = 10^1$ are shown in Fig. 30.

6 Discussion

We have demonstrated a Gibbs sampler for sampling orthogonal predictors in the Bayes linear model. The method uses coupling from the past to determine its own run time to convergence and this was applied to curve and surface fitting problems using wavelets and Demmler-Reinsch bases. A comprehensive bibliography of perfect simulation theory and methods is maintained by David Wilson at `http://dimacs.rutgers.edu/~dbwilson/exact`

In the examples we considered the Gibbs sampler converged remarkably quickly. Knowledge of this has a number of advantages. First, knowing the speed of convergence affords us the luxury of taking a single sample from each run and then restarting the chain thus obtaining *iid* samples from the posterior model space. Second, the extra computation involved in running 2 chains up to coalescence is more than compensated by the fact that we need not use over cautious burn-in times that come with samplers that do not provide time to convergence. For example, in the Demmler-Reinsch curve fitting problem in Section 4.1 the calculation of around 8,000 sweeps of the Gibbs sampler (in total) produced 1,000 perfect samples.

The Demmler-Reinsch representation was derived from an orthogonalization of the standard smoothing spline models. However, it was shown that this orthogonalization process actually aids in the interpretation of the model by illustrating exactly how smoothing is acheived. The Demmler-Reinsch parameterisation has additional benefits from a Bayesian perspective as it facilitates the quantification of prior beliefs on the probability of a basis function being included in the model.

Finally we note that the Demmler-Reinsch representation is not restricted to radial basis functions. In general it can be applied to any linear smoother where the fitted surface $\hat{Y}$ can be written in smoothing form $\hat{Y} = \boldsymbol{S}Y$.

References

Brooks, S. P. and Roberts, G. O. (1999). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing.* To Appear.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain

Monte Carlo. *J. Royal Statist. Soc. B.* **57**, pp. 473-484.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, pp. 391-402.

Demmler, A. and Reinsch, C. (1975) Oscillation matrices and spline smoothing. *Numer. Math.*, **24**, pp. 375-382.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaption by wavelet shrinkage. *Biometrika*, **81**, pp. 425-55.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998) Automatic Bayesian curve fitting. *J. Royal Statist. Soc. B.* **60**, pp. 333-350.

Eubank, R. L. (1999) *Nonparametric regression and spline smoothing.* 2nd Ed. Marcel Decker.

Foster, D and George, E.I. (1994) The risk inflation criterion for multiple regression. *Annals of Statistics* 22:1947-1975.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association.* **85**, pp. 398-409.

George, E.I. and Foster, D. P. (1997). Calibration and Empirical Bayes Selection. Univ of Texas at Austin, Tech. Report.

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

Green, P. J. (1995) Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, pp. 711-732.

Green, P. J. and Silverman B. (1994). *Nonparametric regression and generalized linear models.* Chapman and Hall.

Hastie, T. (1996) Pseudosplines. *J. Royal Statist. Soc. B.* **58**, pp. 379-396.

Holmes, C. C. and Mallick, B. K. (1999). Bayesian wavelet networks for nonparametric regression. To appear. *IEEE. Trans. Neural Nets.*.

Holmes, C. C. and Mallick, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation.* **10**, pp. 1217 - 1233.

Holmes, C. C. and Denison, D. G. T. (1999). Bayesian wavelet analysis with a model complexity prior. To appear. *Bayesian Statistics 6.*.

Hwang, J-N., Lay, S-R., Maechler, M., Martin, D. and Schimert, J. (1994). Regression modeling in back-propagation and projection pursuit learning. *IEEE Trans. Neural Networks.* **5**, 342-353.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 4, pp. 661-675.

Nychka, D. (1999). Spatial process estimates as smoothers. In *Smoothing and Regression.* Eds. Schimek, M. G. Wiley, New York.

Powell, M. J. D. (1987) Radial basis functions for multivariate interpolation: a review. In *Algorithms of Approximation*, Mason, J. C. and Cox, M. G. eds. pp. 143-167. Oxford: Clarendon Press.

Powell, M. J. D. (1992). The theory of radial basis function approximation in 1990. In *Advances in Numerical Analysis*, Light, W. eds. pp. 105-210. Oxford University Press.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms.* **9**, pp. 223 - 252.

Raftery, A. E., Madigan, D. M. and Hoeting, J. (1997). Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association.* **92**, pp. 179-191.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics.* **75**, pp. 317-344.

Figures

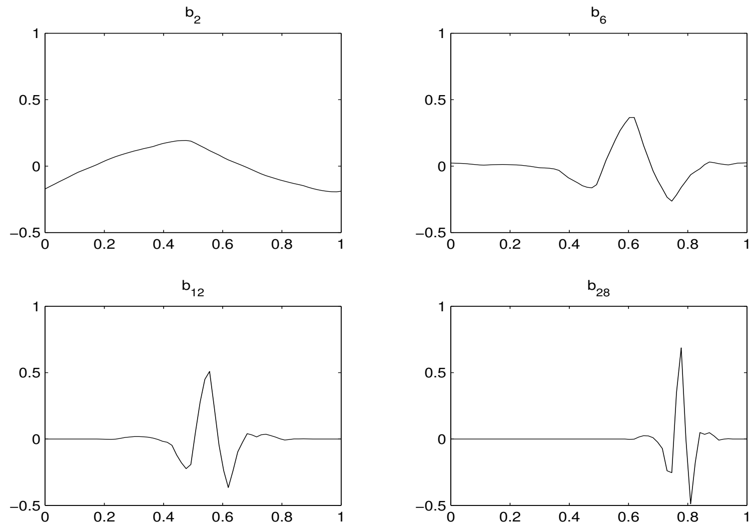

Figure 1: LA(8) bases for 64 point data set on $(0, 1)$

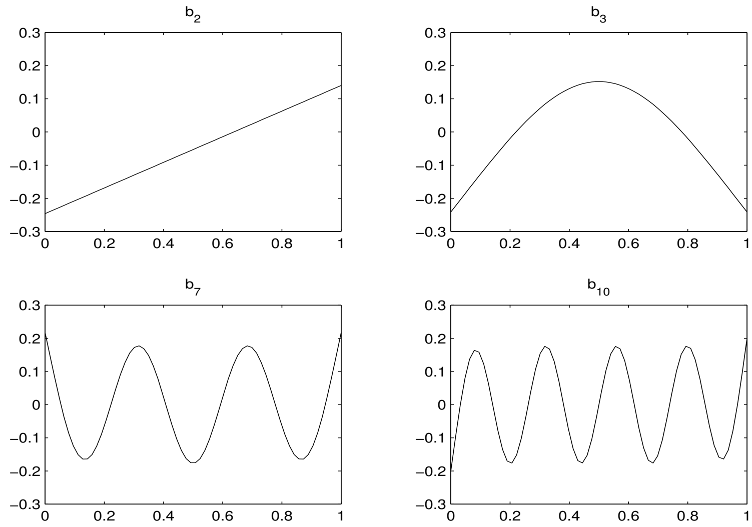

Figure 2: Demmler-Reinsch bases for 64 point data set on $(0, 1)$

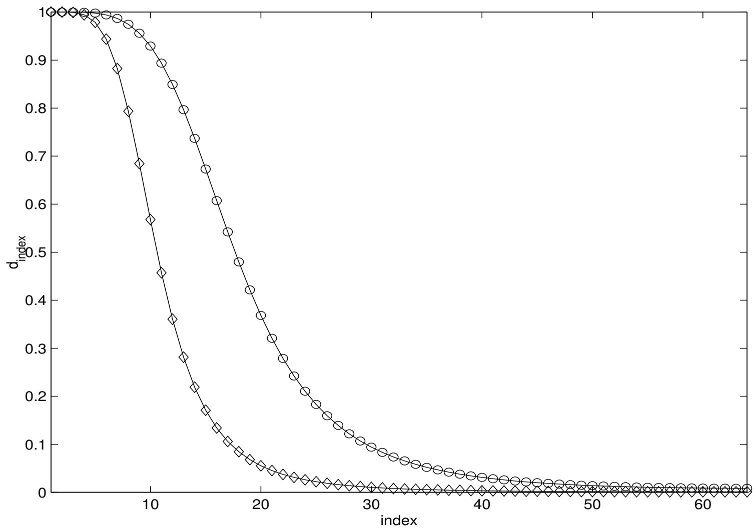

Figure 3: Eigenvalues from the eigendecomposition of the cubic spline smoothing matrix for $\alpha = 10^{-5}$ diamonds and $\alpha = 10^{-6}$ circles. The eigenvalues d_j determine the extent of the damping for the corresponding Demmler-Reinsch basis.

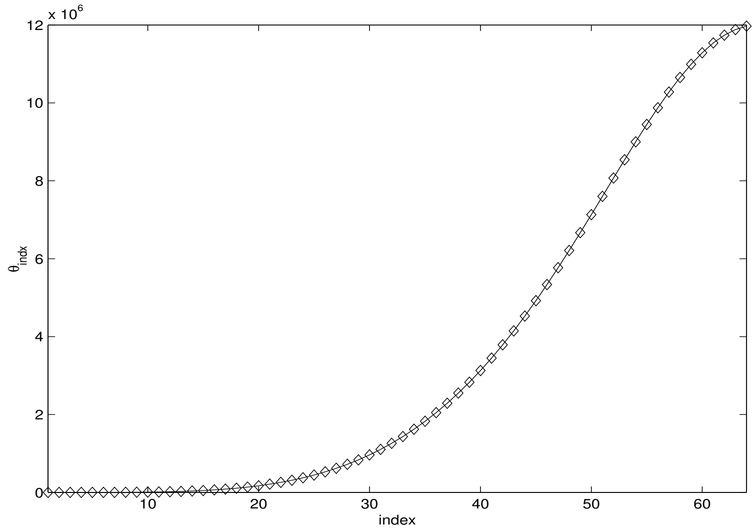

Figure 4: Eigenvalues from the eigendecomposition of $\boldsymbol{K}$ where $\boldsymbol{K}$ is from the cubic spline smoothing matrix $\hat{Y} = (I + \alpha \boldsymbol{K})Y$.

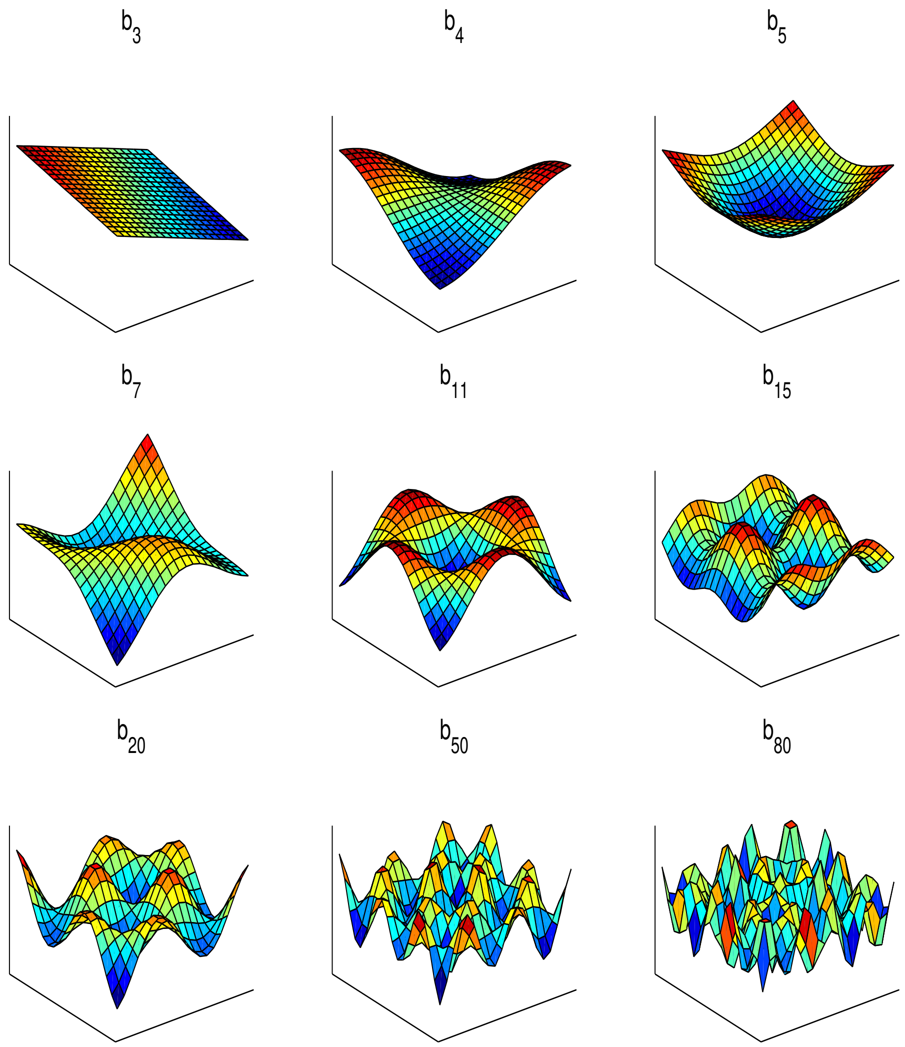

Figure 5: 2d Demmler-Reinsch bases for data on $(0, 1)$.

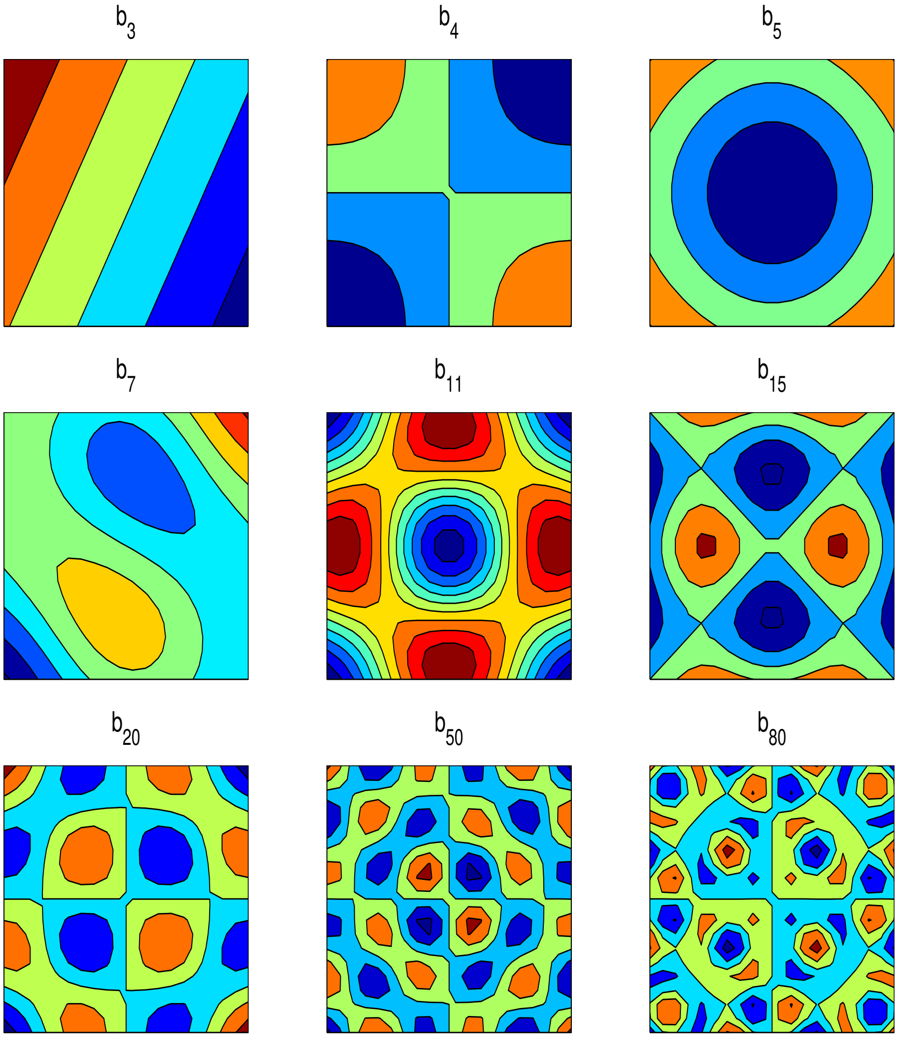

Figure 6: Contour plots for the 2d Demmler-Reinsch bases shown in Fig. 5.

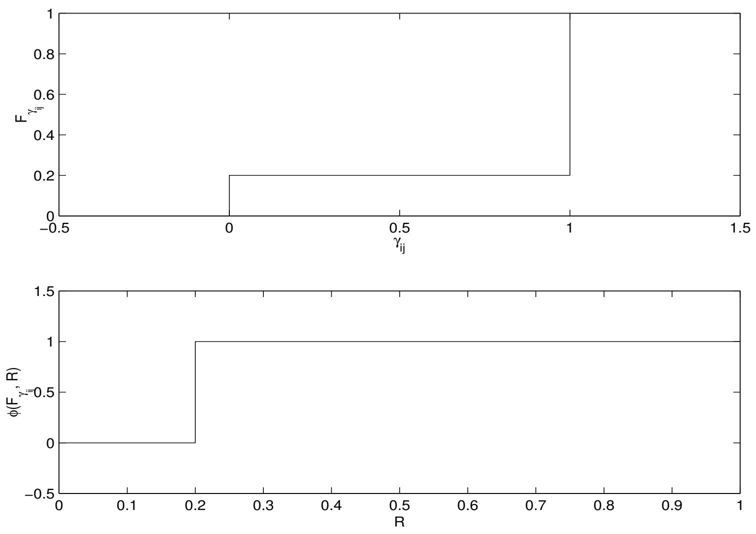

Figure 7: The top graph shows an example of a cumulative density function for the indicator element γ_{ij} which can take values $\{0,1\}$. The bottom graph shows the output from the Gibbs sampler for the example in the top graph. The output of the sampler is a deterministic function ϕ of $F_{\gamma_{ij}}$ and the random input $R \sim U(0,1)$. Clearly the same value for R results in the same output of ϕ.

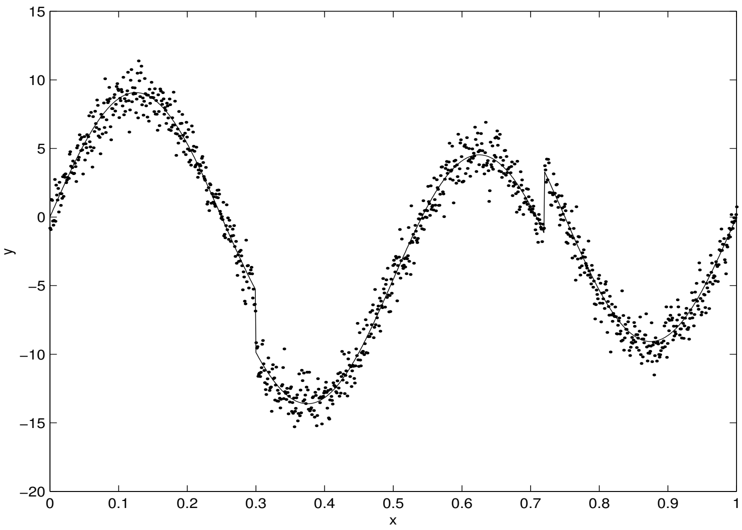

Figure 8: Heavisine data

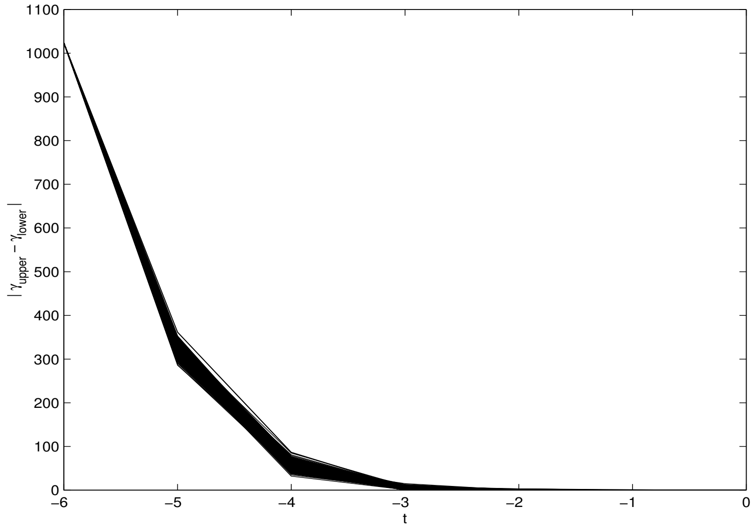

Figure 9: Coalescence times for 500 simulations of the wavelet model fitted to the Heavisine data. The graph shows the absolute difference in the upper and lower chains as a function of time in iterations of the Gibbs sampler.

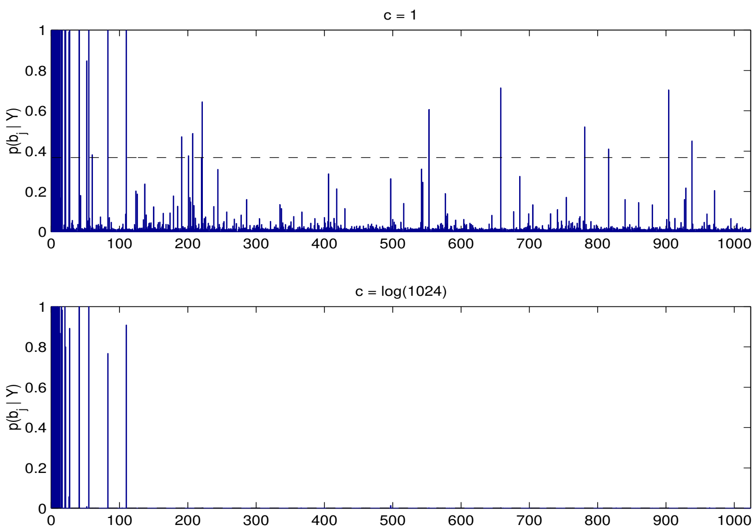

Figure 10: Posterior probability of inclusion for the wavelet basis using two prior settings, indicated by the dashed lines.

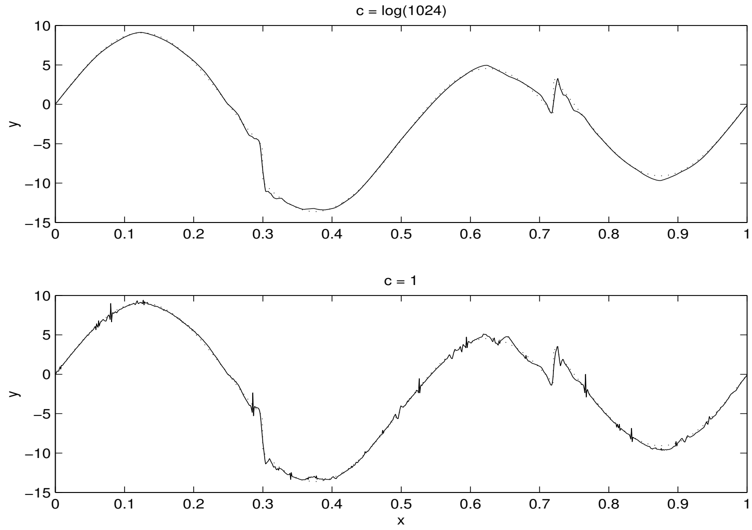

Figure 11: Posterior mean curves generated using 1,000 perfect samples for two settings of the prior.

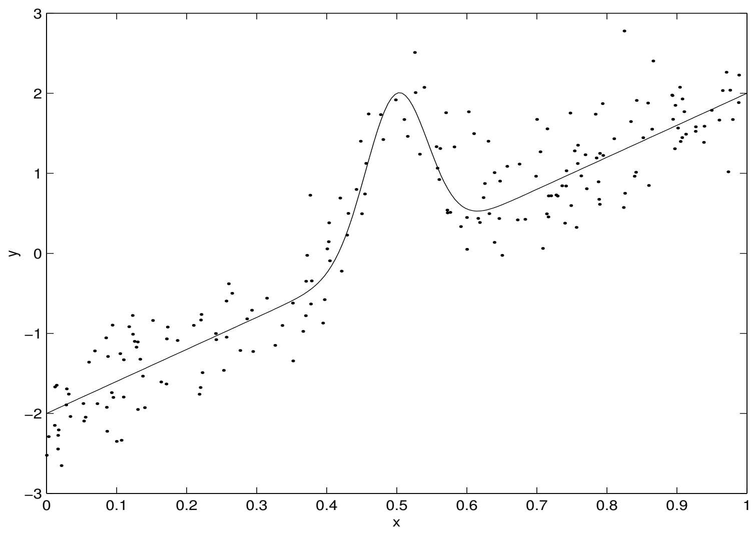

Figure 12: Curve fitting data set and truth.

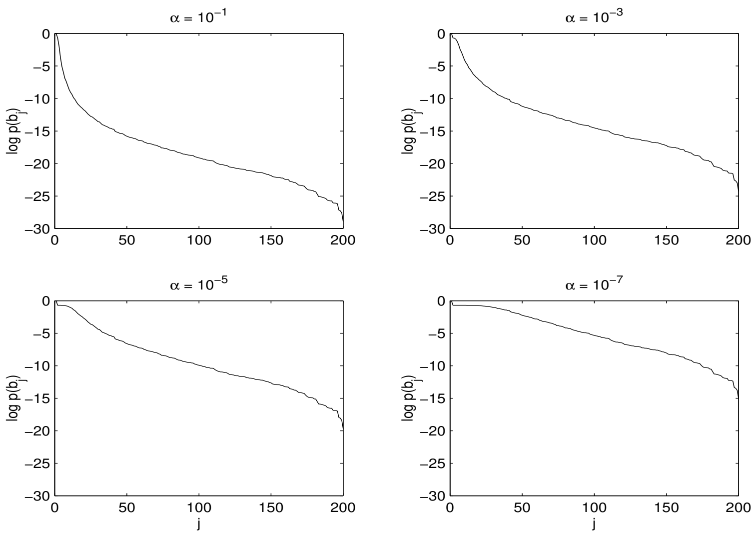

Figure 13: Prior probability of inclusion using formula (23) for four settings of α.

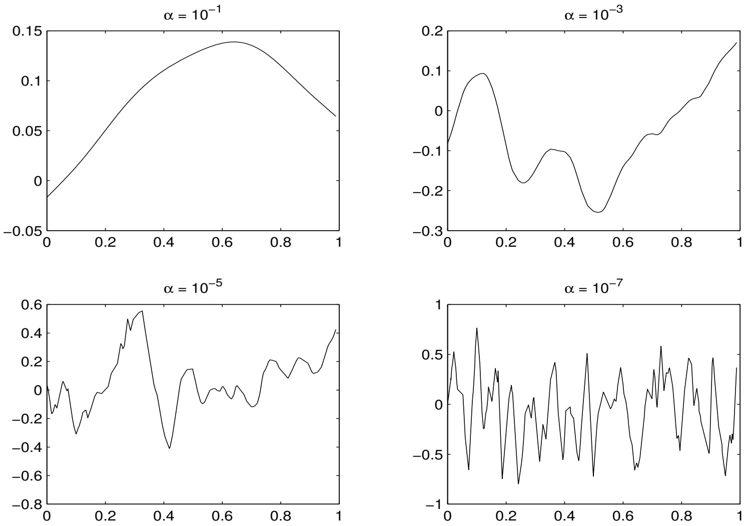

Figure 14: Expected curves using 1000 samples drawn from the priors in Fig. 6.

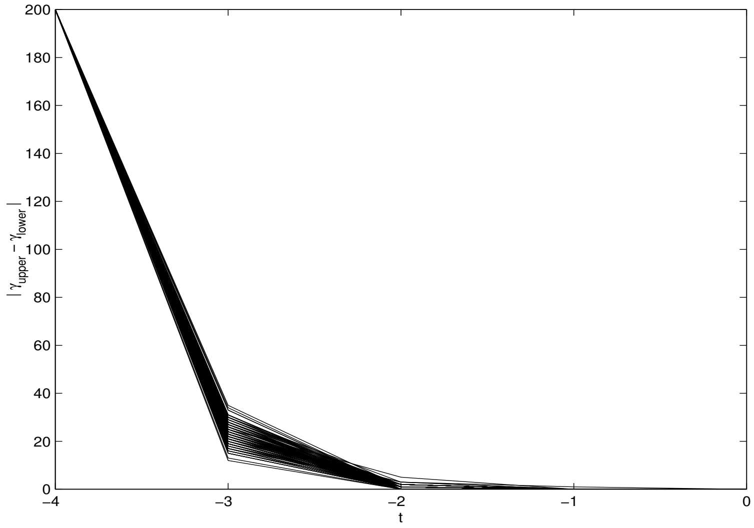

Figure 15: Coalescence times for 500 simulations for the curve fitting example. The plot shows the absolute difference in terms of bases in the model between the upper and lower chains as a function of time for the Gibbs sampler.

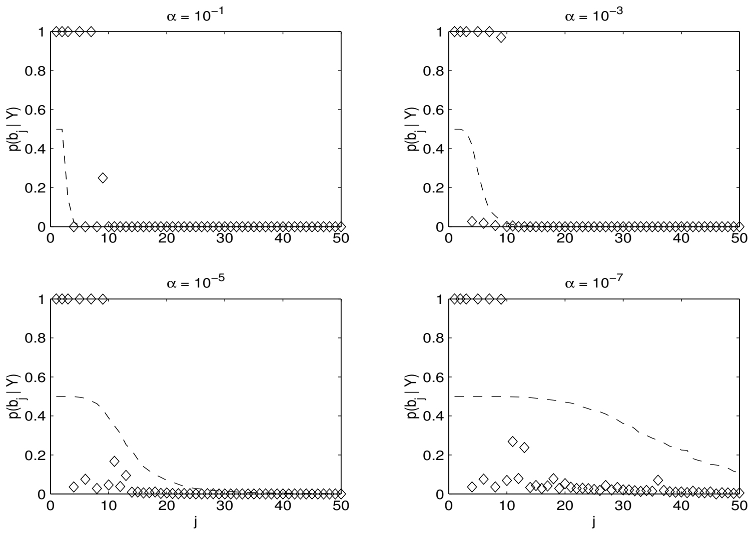

Figure 16: Posterior probability of inclusion taken from 1,000 perfect samples for the four settings of the prior α in Fig. 13. The prior probability of inclusion is shown as dashed lines in each graph.

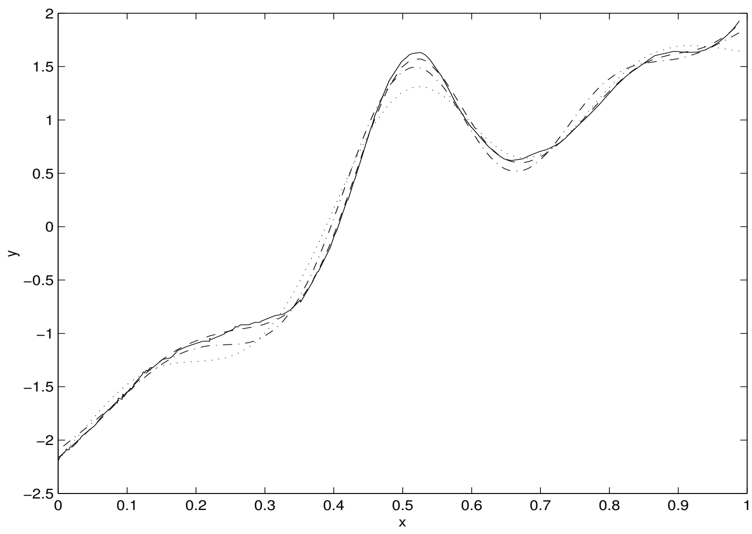

Figure 17: Posterior expectations of the curve using 1,000 *iid* samples for the four prior settings: $\alpha = 10e^{-1}$ (dotted); $\alpha = 10e^{-3}$ (dashdot); $\alpha = 10^{-5}$ (dashed); $\alpha = 10^{-7}$ (solid).

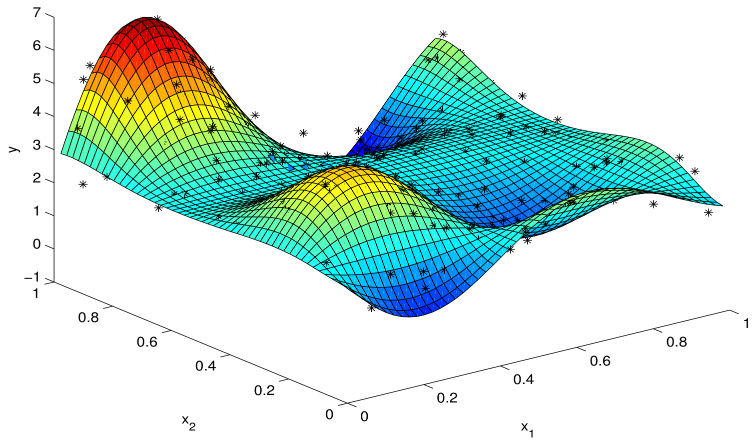

Figure 18: Surface and data points for example 4.3.

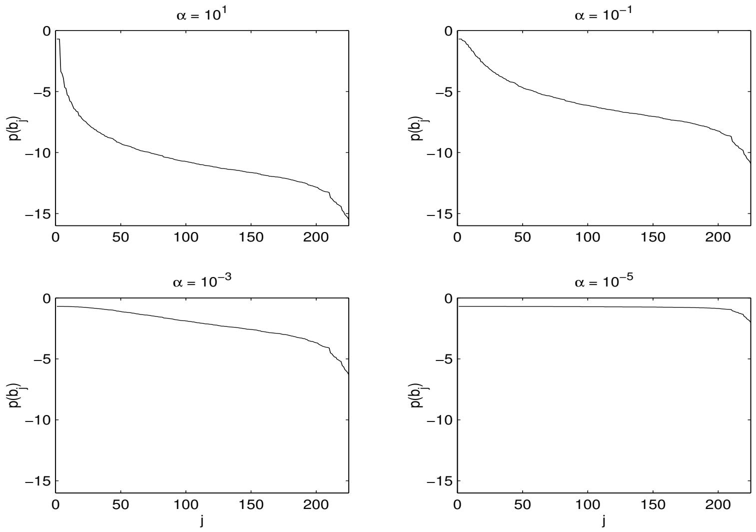

Figure 19: Prior probability of inclusion for the Demmler-Reinsch basis using four settings of the prior paramter α in equation (23).

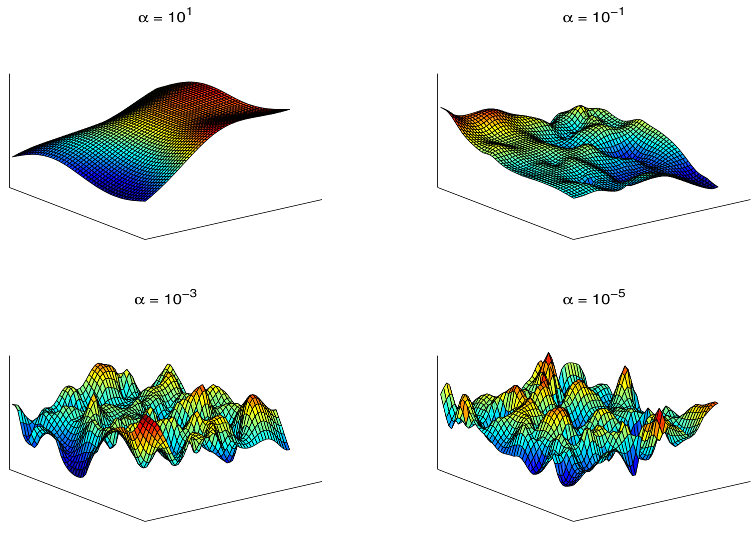

Figure 20: Expected surfaces drawn from the priors shown in Fig. 19.

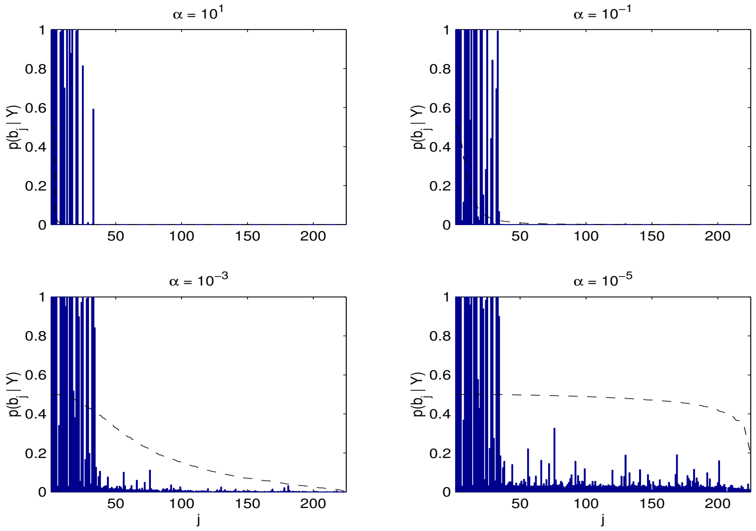

Figure 21: Posterior probability, using 1,000 samples, of basis inclusion for the surface smooth with prior probabiliy shown as dashed lines.

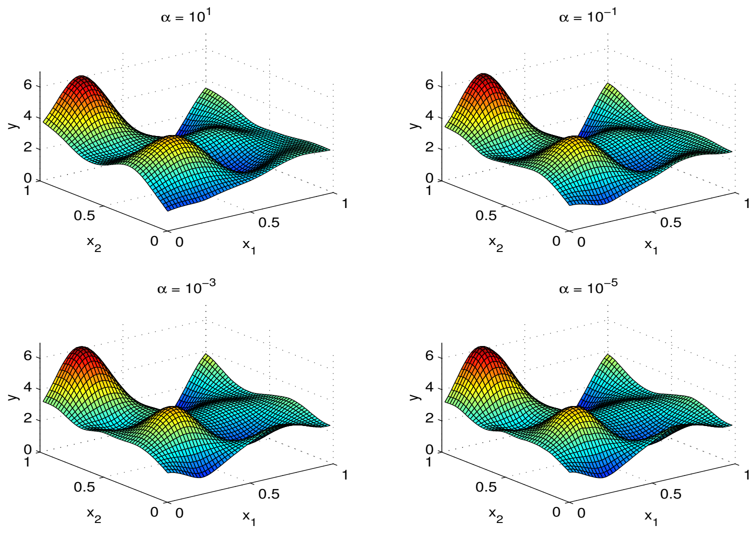

Figure 22: Mean posterior surfaces under different prior settings.

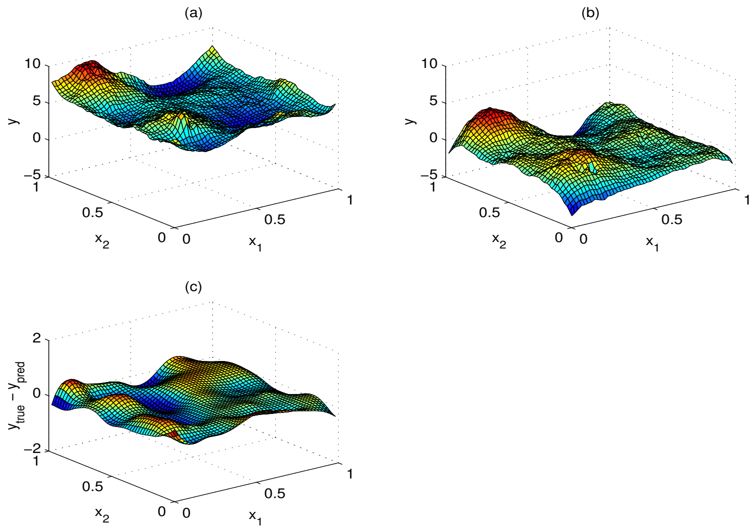

Figure 23: Plot (a) shows the 2.5% upper credible surface and plot (b) the 2.5% lower credible surface for prior setting $\alpha = 10^{-5}$. Plot (c) shows the residual surface of the true function and mean posterior curve for this prior.

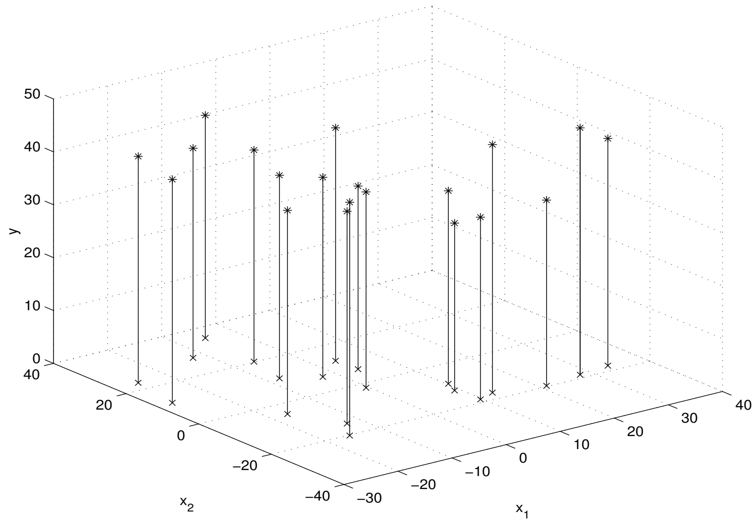

Figure 24: Ozone data.

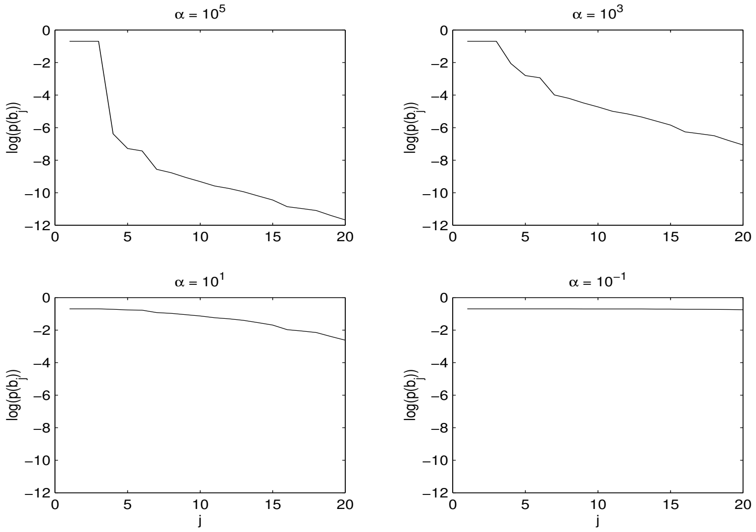

Figure 25: Prior probability of inclusion of the Demmler-Reinsch basis for the ozone data, for various settings of α.

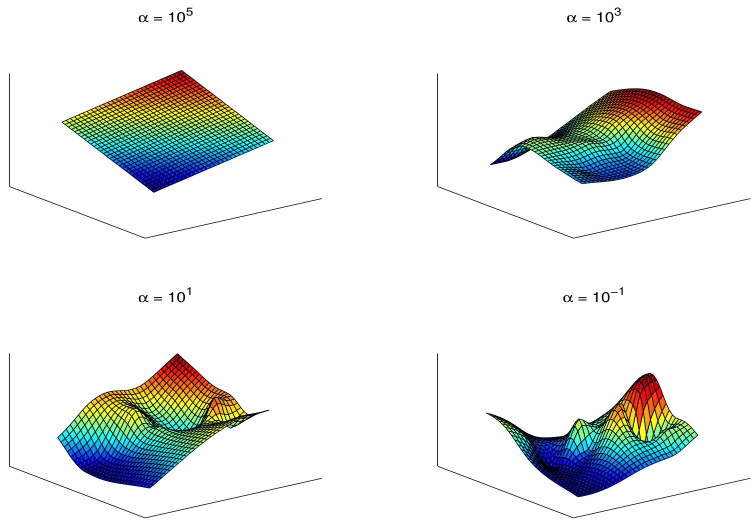

Figure 26: Surfaces drawn from the prior in Fig. 13.

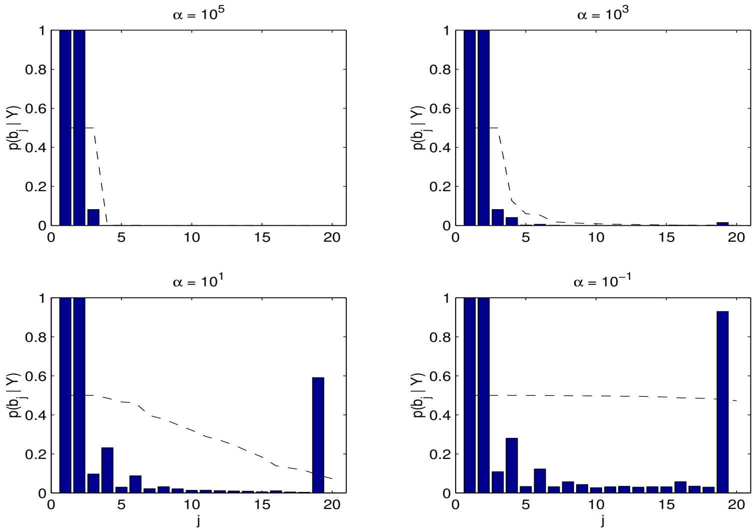

Figure 27: Posterior probability of inclusion for the Demmler-Reinsch basis for the ozone data, for various settings of α using 1000 perfect samples.

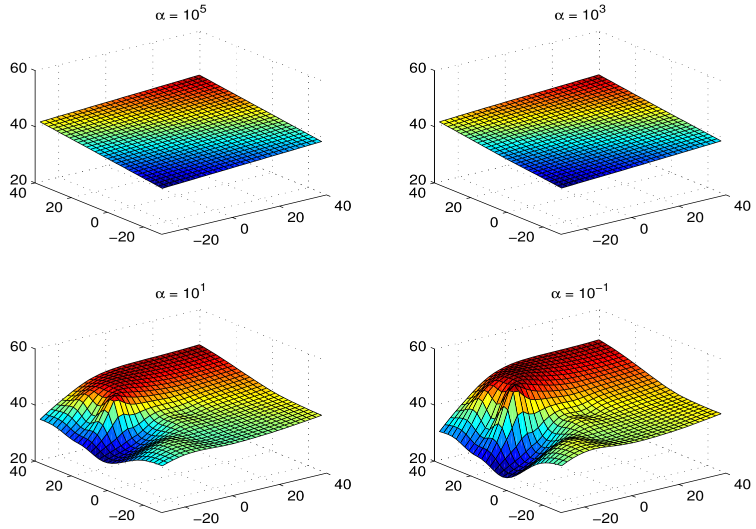

Figure 28: Mean posterior surfaces generated by averaging 1,000 perfect samples with prior settings as in Fig. 12.

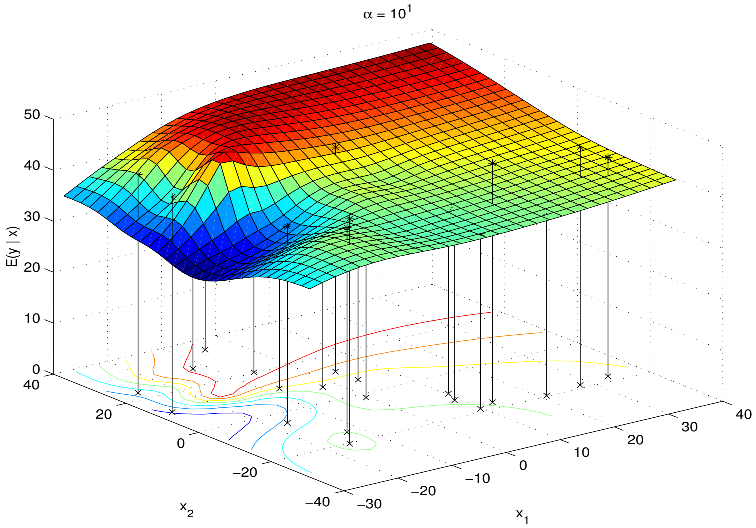

Figure 29: Posterior mean surface and contour plot using the prior setting $\alpha = 10^1$

.

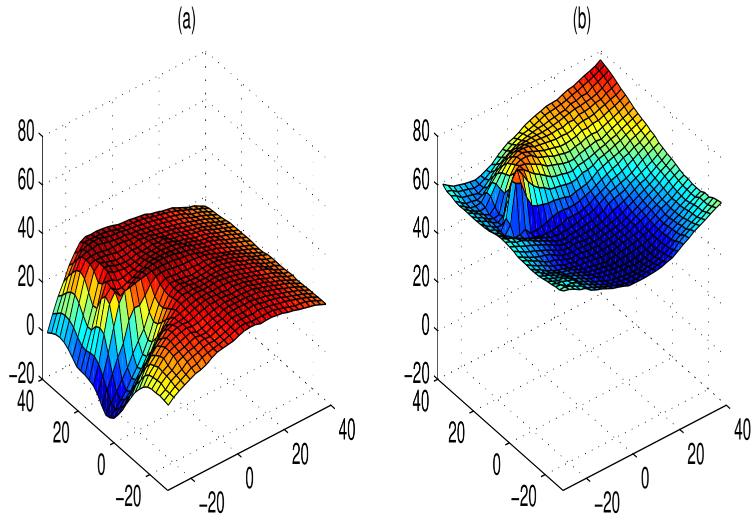

Figure 30: Plots (a) and (b) show the lower and upper 2.5% credible surfaces for the ozone data using $\alpha = 10^1$.