Bayesian approach to Hybrid splines nonparametric regression

Ronaldo Dias

Universidade Estadual de Campinas.

Departamento de Estatística. São Paulo, Brasil

E-mail address: dias@ime.unicamp.br

Dani Gamerman

Universidade Federal do Rio de Janeiro

Abstract

One of the procedures in nonparametric regression is to estimate the regression curve by using a combination of basis functions and smoothing techniques. A Bayesian approach is considered to estimate the number of knots, the smoothing parameter and the knots' positions. The method to obtain the estimate of the regression curve is the reversible jump MCMC (Green, 1995).

1 Introduction

Since the pioneer work of Craven and Wahba (1979) using splines for nonparametric regression several methods were suggested. Recently, Luo and Wahba (1997) and Dias (1999) proposed more adaptive methods to obtain good estimates of the regression curves. The H-splines method introduced by Dias (1998) in the case of nonparametric density estimation, Luo and Wahba (1997) and Dias (1999) in the context of nonparametric regression, combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively. Under the Bayesian scheme, Denison, Mallick and Smith (1998) suggest selecting regression models by using reversible jump MCMC.

The set up for splines nonparametric regression relies on the fact that a unknown function g of one or more variables and a set of measurements are given such that:

$$y_i = \mathcal{L}_i g + \varepsilon_i$$

where $\mathcal{L}_1, \ldots, \mathcal{L}_n$ are linear functionals defined on some linear space $\mathcal{H}$ containing g, and $\varepsilon_1, \ldots, \varepsilon_n$ are measurement errors usually assumed to be independently identically normal distributed with mean zero and unknown variance σ^2. Typically, the $\mathcal{L}_i$ will be point evaluation of the function g.

Straight forward least square fitting is often appropriate but it produces a function which is not sufficiently smooth for some data fitting problems. In such cases, it may be better to look for a function which minimizes a criterion that involves a combination of goodness of fit and an appropriate measure of smoothness. Such criterion is the well known penalized least square problem defined as the following: Finding the minimizer of the penalized least square equation which is,

$$A_\lambda(g) = \sum_{i=1}^{n} (y_i - \mathcal{L}_i g)^2 + \lambda J(g), \tag{1.1}$$

where $J(g)$ is the penalty term usually taken as $\int (g'')^2$ with $g \in \mathcal{W}_2^2 = \{g : g'$ abs. continous and $\int (g'')^2 < \infty\}$ and λ is the smoothing parameter which controls the trade off between fidelity to the data and smoothness.

It is well known that the minimizer $\hat{g}$ is necessarily a natural cubic spline with knots at t_i (see, for example, Silverman and Green (1994), Wahba (1981) and Craven and Wahba (1979)). Note that the roughness penalty $\int_a^b (g''(t))^2 dt$ has the property of reducing the problem of choosing g from an infinite-dimensional class of functions to a finite class of functions since $\hat{g}$ can be written as linear combination of basis functions. Although this fact might lead someone to think that the nonparametric regression problem becomes a parametric problem, one notices that the number of parameters can be as large as the number of observations, and there may be difficulties in interpreting a curve or surface g. Moreover, if the number of observations is large, the system of linear equations for exact solution is too expensive to solve.

In the smoothing techniques the number of basis functions is chosen to be as large as the number of observations and then let the choice of the smoothing parameter to control the flexibility of the fitting (Bates and Wahba, 1982). The H-splines method (Luo and Wahba (1997), Dias (1998) and Dias (1999)) combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively. By taking the point evaluation functionals $\mathcal{L}_i g = g(x_i)$ the equation (1.1) becomes,

$$L_\lambda(g) = ||\mathbf{y} - \mathbf{g}||^2 + \lambda \int (g'')^2, \tag{1.2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{g} = (g(x_1), \ldots, g(x_n))^T$.

Assume that $g \approx g_K = \sum_{i=1}^K \theta_i B_i = X\theta$ so that $g_K \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the space of natural cubic splines (NCS) spanned by the basis functions $\{B_i\}_{i=1}^K$ and X is a $n \times K$ matrix with entries $X_{ij} = B_i(x_j)$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n$. Then, the numerical problem is to find a vector $\theta = (\theta_1, \ldots, \theta_K)^T$ that minimizes,

$$L^*_\lambda(\theta) = ||\mathbf{y} - X\theta||_2^2 + \lambda\theta^T \Omega\theta,$$

where Ω is $K \times K$ matrix with entries $\Omega_{ij} = \int B_i''(t) B_j''(t) dt$. Standard calculations (de Boor, 1978) provide θ as a solution of the following linear system $(X^T X + \lambda\Omega)\theta_\lambda = X^T y$. Note that the linear system now involves $K \times K$ matrices instead of using $n \times n$ matrices which is the case of smoothing splines. Both K and λ controls the trade off between smoothness and fidelity to the data. In particular, when $\lambda = 0$ we have the regression spline case, where K is the parameter that controls the flexibility of the fitting. To exemplify the action of K on the estimated curve, let us consider an example by simulation with $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \varepsilon$ with $\varepsilon \sim N(0, .05)$.

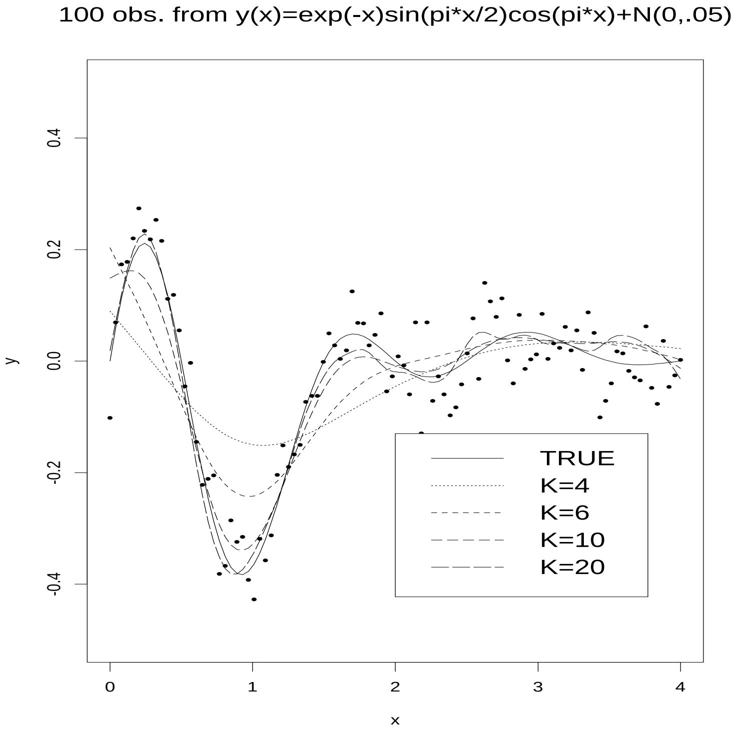

Figure 1 shows the effect of varying the number of basis functions on the estimation of the true curve. Note that the number of basis functions is the same as the number of knots since it is assumed that we are dealing with natural cubic splines space. Observe that small values of K make smoother the estimate and hence oversmoothing may occur. Large values of K may cause undersmoothing.

2 Bayesian approach

Suppose we have the following regression model,

$$y_i = g(x_i) + \varepsilon_i \quad i = 1, \dots, n.$$

where ε_i's are uncorrelated with a $N(0, \sigma^2)$. Moreover, assume that the parametric form of the regression curve g is unknown. Then the likelihood of g given the observations $\mathbf{y}$ is,

$$l_{\mathbf{y}}(g) \propto (\sigma^{-2})^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{g}||^2\}. \tag{2.1}$$

The Bayesian justification of penalized maximum likelihhod is to place a prior density proportional to $\exp\{-\frac{1}{2}\int(g'')^2\}$ over the space of all smooth functions. One may see that with this prior the optimization problems 1.2 and 2.1 are equivalents, (see details in Silverman and Green (1994) and Kimeldorf and Wahba (1970)). However, a infinite dimensional case has a paradox alluded to Wahba (1983). To avoid the paradoxes and difficulties involved in the infinite dimensional case, Silverman (1985) proposed a finite dimensional Bayesian formulation. For this, let $g_K = \sum_{i=1}^{K} \theta_i B_i^t = X\theta_K$ with a knot sequence t placed at order statistics. Thus the penalized likelihood is,

$$l_p(K, t, \sigma^2) \propto (\sigma^{-2})^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \langle\theta_K, X\rangle||^2\} \times \exp\{-\frac{\lambda}{2}(\theta_K^T \Omega \theta_K)\} \tag{2.2}$$

Let $p(\phi)$ be a prior density such that $\phi = (K, \lambda, \sigma^2)$. Hence the posterior density is,

$$\pi(\phi) \propto l_{\mathbf{y}}(\phi)p(\phi), \tag{2.3}$$

where $l_{\mathbf{y}} \propto (\sigma^{-2})^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y}-\langle\theta_K, X\rangle||^2\}$ and $p(\phi) \propto \exp\{-\frac{\lambda}{2}(\theta_K^T \Omega \theta_K)\}$. Then inference is carried out assuming that g can be approximated by g_K which is in a sequence of subspaces with dimension K. The overall parameter space Φ can be written as countable union of subspaces $\Phi = \cup_{k=0}^{\infty}\Phi_k$ where Φ_k is a subspace $\mathbb{R}^{m_K}$.

A complete Bayesian approach would assigned prior distribution to the coefficients of the expansion but this leads to a serious computational difficulties pointed out by Denison et al. (1998) where a comparative study was developed and showed that the least square estimates for the vector θ leads to a non-significant deterioration in performance for overall curve estimation. Moreover, for a given K the interior knots are place at order statistics. This is because particular interest are paid in problems where $y(x) = \int K(x,z)f(z)dz + \varepsilon_i$ which is well known to be a ill-posed problem (see details in Wahba (1982)). Hence, any change in the knots positions could cause considerable change in the function y.

Samples from $\pi(\phi)$ will be taken by using reversible jump MCMC (Green, 1995). For this, a prior density for ϕ must be specified. Define

$$p(\phi) = p(K, \lambda)p(\sigma^2) = p(\lambda|K)p(K)p(\sigma^2),$$

where $\sigma^{-2} \sim Gamma(a, b)$,

$$p(K) = \frac{\exp\{-\alpha\}\alpha^K/K!}{1 - \exp\{-\alpha\}(1 + q_{max})},$$

where $q_{max} = \sum_{j=K_{max+1}}^{\infty} \alpha^j/j!$, and

$$p(\lambda|K) = \psi(K)\exp\{-\psi(K)\lambda\},$$

where ψ is monotone function of K.

In order to sample from the posterior $\pi(\phi)$ we have to consider a variation of dimensionality of this problem. Hence we must design move types between subspaces Φ_k. For this, let the transitions be,

(a) movement of the smoothing parameter λ,

(b) addition of a basis function and

(c) deletion of a basis function.

Note that only the steps (b) and (c) change the dimension of the model. At each transition, an independent random choice is made between each of the three move types. These have probabilities η_K for step (a), b_K for step (b) and d_K for step (c). Naturally, $b_K + d_K + \eta_K = 1$ for all $K = 1, \ldots, K_{max}$ with $b_0 = 1$ $d_0 = 0$ and $\eta_0 = 0$. These probabilities were chosen so that

$$b_K = c \min\{1, p(K+1)/p(K)\}, \quad d_{K+1} = c \min\{1, p(K)/p(K+1)\}$$

and so, $\eta_K = 1 - (b_K + d_K)$. In the simulation we set $c = 0.4$, but other values are valid. Following the notation of Green (1995) we define,

$$\alpha = \min\{1, \text{likelihood} \times \text{prior ratio} \times \text{proposal ratio}\}.$$

The move (a) is on moving from λ to λ' and the accept probability is given by

$$\alpha(\lambda, \lambda') = \min\Big\{1, \frac{p(\lambda'|K)}{p(\lambda|K)}\Big\}.$$

The step (b) changes the dimension of the parameter space. That is, $K \to K+1$ and the accept probability is

$$\alpha(K, K{+}1) = \min\Big\{1, \exp\{-(1/2\sigma^2)||y{-}X\theta_{K+1}||{-}||y{-}X\theta_K||\}\frac{p(K+1)}{p(K)}\frac{d_K}{b_K}\frac{K}{K+1}\Big\}.$$

The acceptance probability for the corresponding deletion of a basis function has the same form with the appropriate changes and the ratio terms inverted.

Algorithm 2.1

```
    1. Initialize by setting the hyperparameters for the prior $p(\phi)$ .
    2. Generate $u$ from a uniform $[0, 1]$.
    3. Go to the move type
  i) if $u \leq b_K$ then go to the addition step (b);
  ii) else if $b_K \leq u \leq b_K + d_K$ then go to deletion step (c);
  iii) else go to step (a)
```

3 Simulation

In this section we present several typical examples of this method. In all examples a vague but proper prior was used for σ^{-2}, namely, $Gamma(10^{-3}, 10^{3})$.

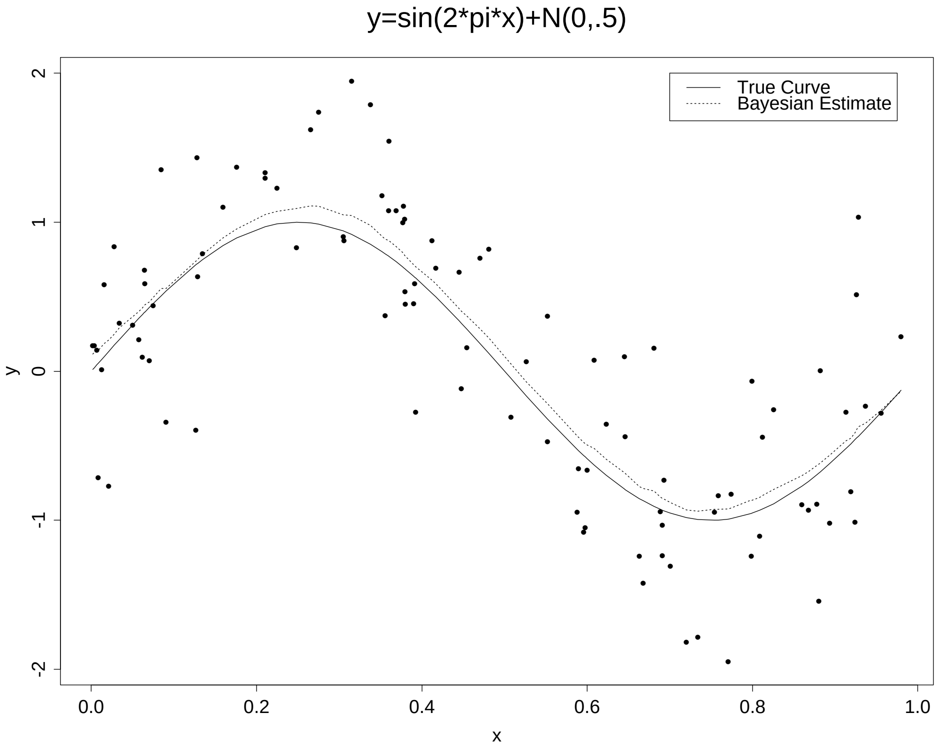

Figure 3.1: One hundred observations with hyperparameters $\alpha = 6$ and $\psi(K) = K^{-2}$

Among several other examples, the next two were taken from Denison et al. (1998) and a comparison with smoothing splines estimate is shown.

Visually, one may observe that this procedure produces a very good estimates even with such vague priors.

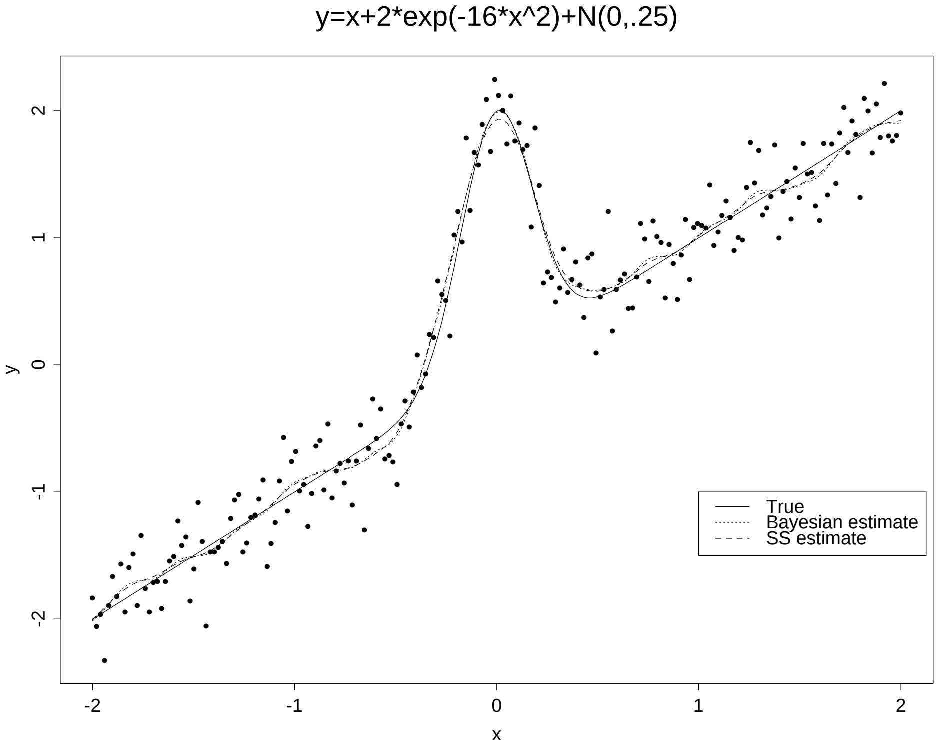

Figure 3.2: Two hundred observations with hyperparameters $\alpha = 8$ and $\psi(K) = K^{-2}$

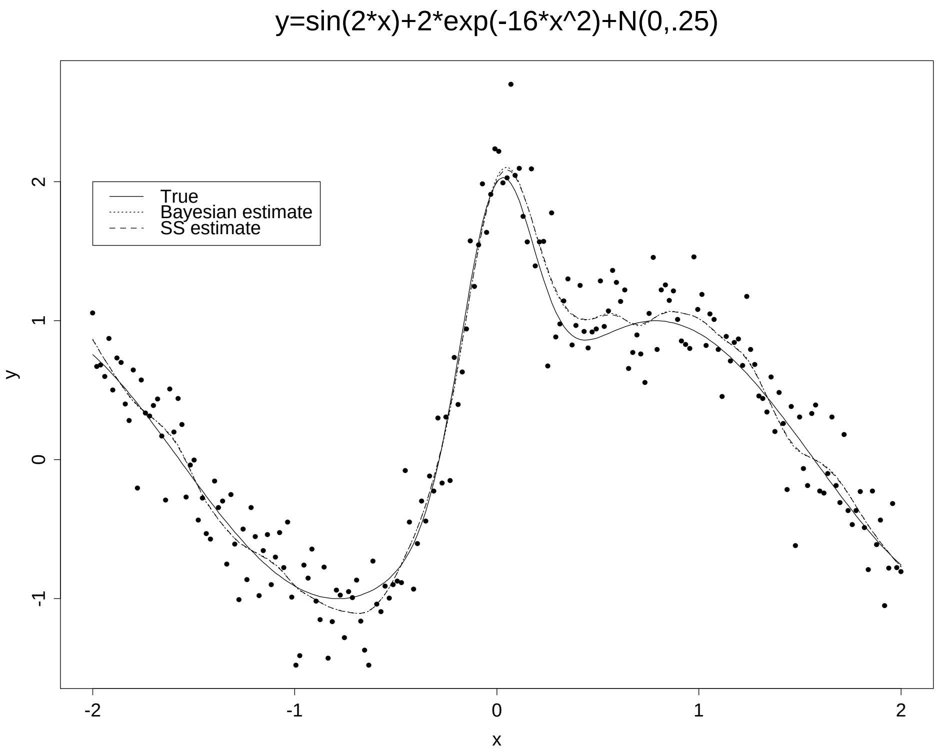

Figure 3.3: Two hundred observations with hyperparameters $\alpha = 20$ and $\psi(K) = K^{-4}$

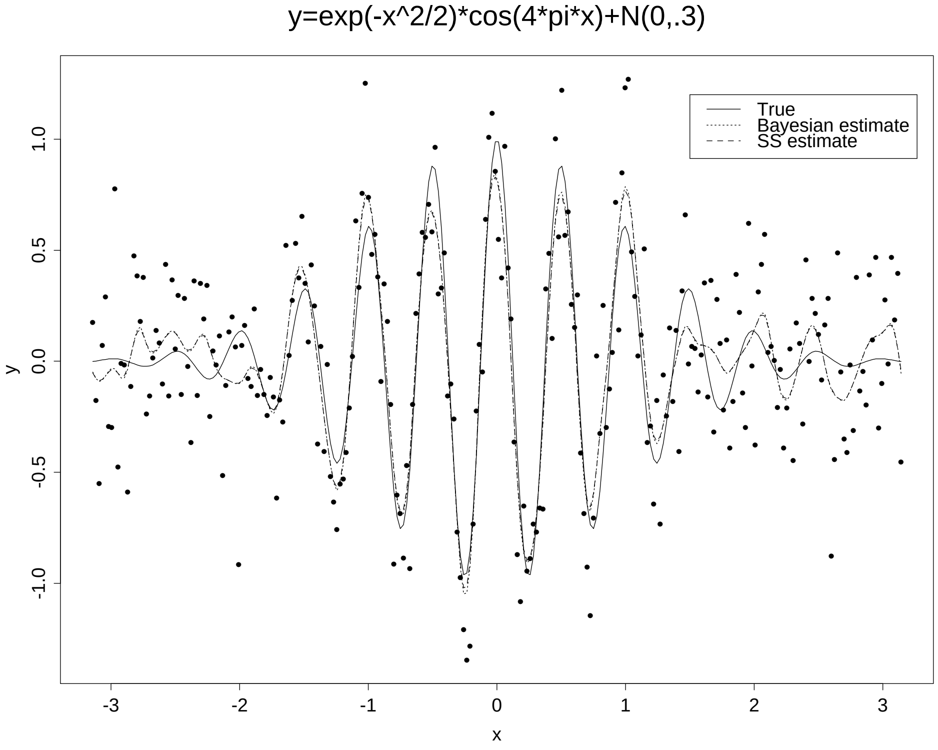

Figure 3.4: Two hundred observations with hyperparameters $\alpha = 40$ and $\psi(K) = K^{-4}$

ACKNOWLEDGEMENT

The firt author would like to thank Prof. Michael Newton and Prof. Mary Lindstrom for all suggestion made during the visit of this author at department of Statistic of Univeristy Wisconsin, Madison, FAPESP grant 95/4996-3, FAEP 259/98 and CNPq grant 450560/98.

References

Bates, D. and Wahba, G. (1982). *Computational Methods for Generalized Cross-Validation with large data sets*, Academic Press, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic bayesian curve fitting, *Journal of the Royal Statistical Society B* **60**: 363–377.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (1999). Sequential adaptive non parametric regression via h-splines, *Communications in Statistics: Computations and Simulations* **28**: issue 2.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination, *Biometrika* **82**: 711–732.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics* **41**: 495–502.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association* **92**: 107–116.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting, *Journal of the Royal Statistical Society, Series B, Methodological* **47**: 1–21.

Silverman, B. W. and Green, P. J. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall (London).

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Ann. of Statistics* **9**: 146–156.

Wahba, G. (1982). Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine, *in* S. S. Gupta and J. O. Berger (eds), *Statistical Decision Theory and Related Topics III, in two volumes*, Vol. 2, pp. 383–418.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline, *JRSS-B, Methodological* **45**: 133–150.