Biometrika (2005), **92,** 1, *pp.* 105–118

Printed in Great Britain

Theory for penalised spline regression

BY PETER HALL

Centre for Mathematics and its Applications, Australian National University, Canberra, ACT0200, Australia

peter.hall@anu.edu.au

AND J. D. OPSOMER

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.

jopsomer@iastate.edu

SUMMARY

Penalised spline regression is a popular new approach to smoothing, but its theoretical properties are not yet well understood. In this paper, mean squared error expressions and consistency results are derived by using a white-noise model representation for the estimator. The effect of the penalty on the bias and variance of the estimator is discussed, both for general splines and for the case of polynomial splines. The penalised spline regression estimator is shown to achieve the optimal nonparametric convergence rate established by Stone (1982).

Some key words: Nonparametric regression; White noise model.

1. INTRODUCTION

Penalised spline regression is rapidly becoming a popular smoothing technique, because of its simplicity and its flexibility in handling a wide range of nonparametric and semiparametric modelling situations (Ruppert et al., 2003). While the concept of using a fixed spline basis combined with a penalty for model complexity is not new (Parker & Rice, 1985; O'Sullivan, 1986), Eilers & Marx (1996) introduced the penalised spline regression model most often used today and coined the term *P*-splines. One drawback of *P*-spline regression is that the theoretical investigation of its properties has lagged behind its application in practice; generally interpretable bias, variance and mean squared error expressions are still missing, and no consistency result nor rate of convergence is available. Such results would be useful for allowing formal comparisons between *P*-splines and other nonparametric regression methods such as kernel-based methods (Wand & Jones, 1995) and smoothing splines (Wahba, 1990), or for providing methodological guidance for constructing spline bases or selecting penalties.

The major difficulty in conducting a theoretical investigation of *P*-spline regression is that the method combines two different features that jointly determine its properties: the spline-based model itself and the selection of the penalty. In the absence of a penalty, penalised splines simplify to regression splines, and asymptotic results have been derived for this class of estimators (Huang & Stone, 2003). The bias of the regression spline

estimator has been shown to depend on the 'size' of the spline basis, and rates of convergence for specific classes of regression models and regression spline estimators are available. Smoothing spline regression can be viewed as a special case of penalised spline regression, in which the number of basis functions is equal to the sample size, and the penalty is constructed in such a way that the resulting estimator is a co-called 'natural spline'. Reviews of asymptotic results for smoothing splines can be found in Wahba (1990) and Eubank (1999).

When penalties are used in combination with a fixed basis, Wand (1999) derived an asymptotic approximation of *P*-splines mean squared error, and Aerts et al. (2002) extended this result to the generalised and multivariate regression contexts. However, those results are expressed only as functions of design matrix products, making it difficult to interpret them in general. Recently, Wand (2003) argued that the *P*-spline regression model can be treated as a mixed model, following an idea of Speed (1991) for smoothing splines. Under this formulation, some or all of the basis functions are treated as random effects and the estimation procedure is formally replaced by an empirical prediction. We will not discuss this approach further here.

Most other studies of the properties of *P*-splines have focused on computational aspects or have evaluated performance through simulation experiments (Eilers & Marx, 1996; Ruppert, 2002). The latter author showed that the properties of *P*-splines estimators are relatively insensitive to the choice of basis functions, as long as enough of them are used. The reported simulation experiments also clearly show that the bias due to the number of basis functions is negligible relative to the bias induced by the penalty.

In this paper, we will use a white-noise process representation of the penalised spline estimator to provide insights into its asymptotic properties. As will be explained in more detail below, this representation treats the data as being generated from a continuous random process, and the spline estimator as an integral over a continuously varying set of basis functions, subject to a penalty. In informal terms, this situation can be thought of as a limiting case in which the number of basis functions has been allowed to grow to infinity, so that the portion of the bias caused by the finite spline basis is ignored. This simplified setting will make it possible to derive exact bias and variance expressions, explicitly show how both terms depend on the penalty, and state consistency and rates of convergence results; see Brown & Low (1996) and Brown et al. (2002) for discussions of the white-noise process representation, and an account of its first-order equivalence to conventional, discrete nonparametric regression.

The main advantage of treating the data generating process and the spline estimator as continuous is that the complicating effect of the finite set of basis functions is removed. This enables us to explore the role of the penalty and its relationship with the sample size in ways that are not possible in the discrete-data, finite-basis setting. One of the interesting findings in this context is that asymptotic study of penalised spline regression using Taylor series expansion, as done in Wand (1999) for instance, should be undertaken with some care, because the first term in the expansion of the mean squared error is generally infinite. While we show this in the white-noise case, our finding raises issues for the finite-basis case as well if the number of basis functions is allowed to grow.

2. DEFINITION OF PENALISED SPLINE ESTIMATOR

Suppose we are given n observations from the regression model

$$y_i = m(x_i) + \varepsilon_i, \tag{1}$$

where m is an unknown smooth function, the ε_i are independent random variables with mean zero and respective variances $v(x_i)$, and the design points x_i all come from a known compact interval $\mathscr{I}$. In order to estimate m, we might approximate it by a 'generalised spline',

$$m(x; \beta, K) = \sum_{k=0}^{K} \beta_k \phi_k(x). \tag{2}$$

Here, $\beta = (\beta_0, \ldots, \beta_K)^{\mathrm{T}}$ denotes the coefficient vector, and $\phi_0, \phi_1, \ldots$ represents the basis for a corresponding class of function on $\mathscr{I}$.

The function at (2) is generally fitted by least squares, with an additive penalty applied to all but the first $p+1$, say, of the components; that is, $\beta = \hat{\beta}$ is chosen to minimise

$$\frac{1}{n} \sum_{i=1}^{n} \{y_i - m(x_i; \beta, K)\}^2 + \lambda \sum_{k=1}^{K-p} \beta_{p+k}^2, \tag{3}$$

where selection of $\lambda \geqslant 0$ determines the degree of smoothness of the resulting fit.

In addition to this generalised spline setting, we will also derive results for the special case where the ϕ_k are the polynomial spline basis used in Ruppert et al. (2003) for example. In that case,

$$\phi_k(x) \equiv x^k \quad (0 \leqslant k \leqslant p), \quad \phi_k(x) \equiv (x - \kappa_{k-p})_+^p \quad (p+1 \leqslant k \leqslant K), \tag{4}$$

where $(x - \kappa_k)_+^p$ denotes $\max(0, x - \kappa_k)$ raised to the pth power, p represents the degree of the spline, and $\kappa_1 < \ldots < \kappa_{K-p}$ are fixed points, or knots, in $\mathscr{I}$. If K increases in such a manner that the knots become dense in $\mathscr{I}$, then the basis functions ϕ_k, for $k \geqslant p+1$, span the set of square-integrable functions on $\mathscr{I}$. The first $p+1$ components in (4) are included since then the spline fit can equivalently be represented using other basis functions, such as a pth-degree B-spline basis (Eilers & Marx, 1996) or the Demmler–Reinsch basis (Nychka & Cummins, 1996), which are more efficient from a computational standpoint. However, when K is sufficiently large the first $p+1$ components may be dropped without appreciably affecting performance, since the remaining components almost form a complete basis for square-integrable functions on $\mathscr{I}$; see § 4·4 for further discussion of this point.

In preparation for the white-noise representation we will adopt below, it is useful to write the piecewise polynomial basis functions as continuous functions of the knots, i.e.

$$\phi(x|\kappa) = (x - \kappa)_+^p, \tag{5}$$

where $\kappa \in \mathscr{I}$, so that $\phi_k(x) = \phi(x|\kappa_{k-p})$ for $k \geqslant p+1$. In this setting, write

$$x_i^* = (\phi_0(x_i), \ldots, \phi_p(x_i), \phi(x|\kappa_1), \ldots, \phi(x|\kappa_{K-p})),$$

where the notation x_i^* distinguishes this transformed covariate from its original form, x_i. Then (3) is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^* \beta)^2 + \lambda \sum_{k=1}^{K-p} \beta_{p+k}^2,$$

which is minimised for β equal to

$$\hat{\beta} = (X^{\mathrm{T}} X + n\lambda D)^{-1} X^{\mathrm{T}} Y, \tag{6}$$

where X represents the matrix with rows x_i^* for $i = 1, \ldots, n$, Y denotes the column vector of response variables y_i, and D is a diagonal matrix with $p+1$ zeros on the diagonal followed by K ones. The penalty term in (6) is written as $n\lambda D$ rather than the more customary λD, so that a constant amount of smoothing under increasing sample size implies a constant value for the parameter λ. The resulting estimator of m at x is

$$\hat{m}(x) = \sum_{k=0}^{K} \hat{\beta}_k \phi_k(x), \tag{7}$$

where $\hat{\beta}_k$ denotes the kth component of $\hat{\beta}$.

3. Estimator in white-noise representation

As noted in § 1, it is difficult to evaluate the properties of the P-splines regression estimator (6) theoretically. Therefore, we will reformulate the penalised regression estimation problem in a form that is more amenable to investigation. The white-noise version of the model (1) is

$$y_t = m(t) + e_t,$$

where t varies in the continuum and the noise e_t can be interpreted as $n^{-\frac{1}{2}} v(t)^{\frac{1}{2}} DW(t)$, where $v(t)$ is the error variance when the design point equals t, and $DW(t)$ can be figuratively represented as a 'derivative' of standard Brownian motion, W, in the sense that $DW(t)dt = dW(t)$. We refer to Brown & Low (1996) and Brown et al. (2002) for discussions of this representation, and for a proof of the asymptotic equivalence of nonparametric regression and white-noise models.

In the white-noise case, assuming that the basis functions may be expressed as in (5), and permitting K to increase so that we can interpret the series in (2) and (3) as integrals, we rewrite the penalised least squares criterion (3) as

$$I(\beta) \equiv \int_{\mathscr{I}} \left\{ y_t - \sum_{k=0}^{p} \beta_k \phi_k(t) - \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds \right\}^2 f(t)dt + \lambda \int_{\mathscr{I}} \beta(t)^2 dt, \tag{8}$$

where β now denotes $\beta_0, \ldots, \beta_p$ as well as the function $\beta(.)$. Also, f is the density of the distribution of the x_i's, and $\rho(t)$ equals the asymptotic value of the proportion of knots κ_k which are distributed in a neighbourhood of $t \in \mathscr{I}$, as K increases. In particular, $\rho \geqslant 0$. Likewise, $\beta(t)$ is the limit, as $K \to \infty$, of the average value of β_{k+p} for indices k such that κ_k is close to t. If the knots were equally spaced on $\mathscr{I}$ then ρ would be identically constant, whereas if the knots were approximations to quantiles of the density f then ρ would be proportional to f. In smoothing splines, the knots correspond to the locations of the observations, so that a white-noise representation of smoothing spline regression would be more closely related to the $\rho \sim f$ case.

It is possible to work with the full model, in which $\beta_0, \ldots, \beta_p$, as well as the function $\beta(.)$, are estimated by minimising $I(\beta)$. However, the effect of estimating $\beta_0, \ldots, \beta_p$ has negligibly small influence on both bias and variance of the final estimator of m. Indeed, the basis functions $\phi(t|s)$ that we shall consider are complete in the class of square-integrable functions on $\mathscr{I}$, and so do not need to be augmented, for example by the set of polynomials of degree p in the polynomial spline case. This is shown formally in the Appendix. Hence we may take $\beta_0, \ldots, \beta_p$ to be arbitrary. In particular they can be assumed

'known', or taken to be identically zero, in which case they are effectively omitted. Therefore we shall omit the parameters $\beta_0, \ldots, \beta_p$ from the objective function $I(\beta)$ in (8) as well, and instead minimise it over the function $\beta(.)$ only. In what follows, 'β' will denote the continuous function $\beta(.)$ unless otherwise specified.

We make some preliminary assumptions that will be used in what follows. First, assume the following.

Assumption 1. The design density f and knot density ρ are bounded away from zero and ∞ on $\mathscr{I}$.

For any set of spline basis functions ϕ, we define a functional operator ψ by letting

$$\psi(u, v) = \int_{\mathscr{I}} \phi(t|u)\phi(t|v)f(t)dt \tag{9}$$

and taking the operator to be the functional which maps the square-integrable function α to $\psi\alpha$, defined by $(\psi\alpha)(u) = \int_{\mathscr{I}} \psi(u, v)\alpha(v)dv$. Here and below, we use the same symbol for both the operator and its 'kernel' in (9).

Assumption 2. The basis functions $\phi(t|s)$ satisfy

$$\sup_{t \in \mathscr{I}} \int_{\mathscr{I}} \phi(t|s)^2 ds < \infty. \tag{10}$$

It follows from this assumption, the definition of ψ, and the assumption that f is bounded, that $\iint_{\mathscr{I}^2} \psi^2 < \infty$. This implies that $\psi\alpha$ is square-integrable whenever α is square-integrable.

As proven in the Appendix, in order for β to be uniquely determined as the minimiser of $I(\beta)$ when $\lambda = 0$ and Assumptions 1 and 2 hold, it is necessary and sufficient that the following be true.

Assumption 3. The operator ψ is nonsingular.

The operator is readily seen to be positive semidefinite; that is,

$$\int_{\mathscr{I}} \psi(u, v)\alpha(u)\alpha(v)dudv \geqslant 0$$

for each function α. In order for ψ to be nonsingular, as required by Assumption 3, this inequality must be strict whenever α does not vanish almost everywhere on $\mathscr{I}$; see § 4·4 for discussion of this issue in the polynomial spline case.

Using a calculus of variations argument, and recalling that we assume $\beta_0, \ldots, \beta_p$ to be known, we can see that the function $\beta = \hat{\beta}$ which minimises $I(\beta)$, defined at (8), must be the solution of the equation

$$\int_{\mathscr{I}} \left\{ y_t - \sum_{k=0}^{p} \beta_k \phi_k(t) - \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds \right\} \phi(t|s_0)f(t)dt = \lambda\beta(s_0)/\rho(s_0), \tag{11}$$

which has to hold for each $s_0 \in \mathscr{I}$. Except for the fact that we have taken $\beta_0, \ldots, \beta_p$ to be known, this is the analogue, in the white-noise model, of the normal equations leading to (6).

The estimator, $\hat{m}$, of m is given by

$$\hat{m}(t) = \sum_{k=0}^{p} \beta_k \phi_k(t) + \int_{\mathscr{I}} \hat{\beta}(s)\rho(s)\phi(t|s)ds, \tag{12}$$

where $\hat{\beta}$ is the function solving (11). This formula is the white-noise analogue of (7), assuming that $\beta_1, \ldots, \beta_p$ are known.

4. Formulae for bias and variance in white-noise model

4·1. *Systematic and stochastic errors for generalised splines*

First we shall derive equations that define the systematic and stochastic errors of the function estimator $\hat{\beta}$. We shall assume throughout § 4 that Assumptions 1–3 hold. Put $m^*(t) = m(t) - \sum_{0 \leqslant k \leqslant p} \beta_k \phi_k(t)$, $a(t) = -\beta^0(t)/\rho(t)$ and $A(t) = \int_{\mathscr{I}} e_s \phi(s|t) f(s) ds$, and define the functions β^0, δ and Δ to be the solutions of the respective equations

$$m^*(t) = \int_{\mathscr{I}} \beta^0(s)\rho(s)\phi(t|s)ds \quad (\text{for all } t \in \mathscr{I}), \tag{13}$$

$$\int_{\mathscr{I}} \delta(s)\rho(s)\psi(s, t)ds + \lambda\delta(t)\rho(t)^{-1} = a(t) \quad (\text{for all } t \in \mathscr{I}), \tag{14}$$

$$\int_{\mathscr{I}} \Delta(s)\rho(s)\psi(s, t)ds + \lambda\Delta(t)\rho(t)^{-1} = A(t) \quad (\text{for all } t \in \mathscr{I}), \tag{15}$$

where ψ is given by (9). We state an additional assumption here.

Assumption 4. Each of β^0, δ and Δ is well defined and square-integrable. In particular, $\int_{\mathscr{I}} E(\Delta^2) < \infty$ and

$$\int_{\mathscr{I}} (\beta^0)^2 < \infty. \tag{16}$$

In view of Assumption 3, if there exist square-integrable solutions to (13)–(15) then they are unique. Explicit formulae for β^0, δ and Δ will be given in § 4·2, but it is clear from (13) that β^0 is the function that exactly 'maps' the spline functions ϕ to the unknown function m^* and hence m. It is shown in the Appendix that

$$\hat{\beta} = \beta^0 + \lambda\delta + \Delta, \tag{17}$$

where $\hat{\beta}$ denotes the solution of (11). Together, (12), (13) and (17) imply that $\hat{m}$ is given by

$$\hat{m}(t) = m(t) + \int_{\mathscr{I}} \{\lambda\delta(s) + \Delta(s)\}\rho(s)\phi(t|s)ds. \tag{18}$$

As will be made explicit in § 4·2, the first integral, $\lambda \int_{\mathscr{I}} \delta(s)\rho(s)\phi(t|s)ds$, corresponds to the systematic error, or bias, of $\hat{m}(t)$, and the second integral, $\int_{\mathscr{I}} \Delta(s)\rho(s)\phi(t|s)ds$, to the stochastic error.

To connect the white-noise model with the discrete model discussed in § 2, note that β^0 is the limit, as $n \to \infty$ and $\lambda \to 0$, of the estimator $\hat{\beta}$ defined in § 2. To be specific, the limit of the kth component of $\hat{\beta}$, as k and K increase, and as $n \to \infty$, in such a manner that $k/K \to t$, is given by $\beta^0(t)$.

4·2. *Explicit formulae for systematic and stochastic errors*

Section 4·1 provided equations implicitly defining the systematic and stochastic error components of the *P*-spline estimator. Using spectral functional decomposition arguments, we can derive explicit expressions for both components that will be more useful when discussing the mean squared error properties of the estimator. We only consider the case in which ρ is a positive constant, corresponding to equally spaced knots on $\mathscr{I}$; other cases appear to be substantially more complex. If we denote the constant simply by ρ, it is shown in the Appendix that

$$\beta^0 = \rho^{-1} \sum_{j=1}^{\infty} \frac{\int_{\mathscr{I}} b\psi_j}{\theta_j} \psi_j, \quad \delta = \rho \sum_{j=1}^{\infty} \frac{\int_{\mathscr{I}} a\psi_j}{\rho^2\theta_j + \lambda} \psi_j, \quad \Delta = \rho \sum_{j=1}^{\infty} \frac{\int_{\mathscr{I}} A\psi_j}{\rho^2\theta_j + \lambda} \psi_j, \tag{19}$$

where the functions a and A are given in § 4·1,

$$b(t) = \int_{\mathscr{I}} m^*(s)\phi(s|t)f(s)ds, \tag{20}$$

and the scalar θ_j and function ψ_j are obtained from the spectral decomposition of ψ:

$$\psi(u, v) = \sum_{j=1}^{\infty} \theta_j \psi_j(u)\psi_j(v). \tag{21}$$

Here, $\theta_1 \geqslant \theta_2 \geqslant \ldots > 0$ denote the eigenvalues of the operator ψ, and $\psi_1, \psi_2, \ldots$ are the respective orthonormal eigenvectors. Since all the eigenvalues must be strictly positive, because the operator ψ is strictly positive definite by Assumption 3, the functions ψ_j form an orthonormal basis for the space of all square-integrable functions; see Chapter 4 of Indritz (1963) for an introduction to spectral theory for symmetric linear operators. Note that, for a given choice of spline function, the quantities θ_j and the functions ψ_j are known and can be computed, while a, b and A are unknown and depend on the data.

Recall from § 3 that if Assumption 2 holds then $S_1 \equiv \int_{\mathscr{I}^2} \psi^2 < \infty$. This in turn implies that $\sum_j \theta_j^2 < \infty$, and indeed the sum equals S_1. Even though the series $\sum_j \theta_j$ may not converge, the infinite series at (21) converges absolutely, uniformly in $(x, y) \in \mathscr{I}^2$; see Indritz (1963, p. 217). The second and third identities in (19) imply that if $\lambda > 0$ then $\int_{\mathscr{I}} \delta^2 < \infty$ and $\int_{\mathscr{I}} E(\Delta^2) < \infty$, provided $\int_{\mathscr{I}} a^2 < \infty$ and $\int_{\mathscr{I}} E(A^2) < \infty$ respectively. The first of the latter two conditions follows directly from Assumption 4, since $a = -\beta^0/\rho$; the second is a consequence of Assumption 2. Therefore, Assumptions 2 and 4 are the key to $\hat{m}$, given by (18), having finite mean integrated squared error.

The function Δ, being linear in functions with zero mean, has vanishing expectation, while δ is a non-stochastic function. In view of (18) and (19),

$$\hat{m}(t) - m(t) = \text{bias}\,(t) + \text{stoch}\,(t), \tag{22}$$

where 'bias (t)' and 'stoch (t)' denote respectively the bias and error-about-the-mean evaluated at t, and

$$\text{bias}\,(t) = \rho^2\lambda \sum_{j=1}^{\infty} (\rho^2\theta_j + \lambda)^{-1} \left(\int_{\mathscr{I}} a\psi_j \right) \left\{ \int_{\mathscr{I}} \psi_j(s)\phi(t|s)ds \right\}, \tag{23}$$

$$\text{stoch}\,(t) = \rho^2 \sum_{j=1}^{\infty} (\rho^2\theta_j + \lambda)^{-1} \left(\int_{\mathscr{I}} A\psi_j \right) \left\{ \int_{\mathscr{I}} \psi_j(s)\phi(t|s)ds \right\}, \tag{24}$$

the latter having zero expected value. From the definition of e_t in § 3, the random variables $N_j \equiv \int_{\mathscr{I}} A\psi_j$, appearing in (24), are such that the following holds.

Assumption 5. The random variables $N_1, N_2, \ldots$ are jointly normally distributed with zero means and covariances, and

$$\operatorname{cov}(N_j, N_k) = \frac{1}{n} \int_{\mathscr{I}} h_j h_k, \tag{25}$$

where

$$h_j(s) = v(s)^{\frac{1}{2}} f(s) \int_{\mathscr{I}} \phi(s|t)\psi_j(t)dt.$$

Using (22)–(24) and Assumption 5, we may derive sufficient conditions for uniform convergence. For example, if v is bounded, if $\sum_j |\int_{\mathscr{I}} \beta^0 \psi_j| + \sum_j (\theta_j \log j)^{\frac{1}{2}} < \infty$ and if $\lambda = \lambda(n)$ decreases to zero, as n increases, sufficiently slowly that

$$n^{-\frac{1}{2}} \sum_{j=1}^{\infty} \frac{(\theta_j \log j)^{\frac{1}{2}}}{\theta_j + \lambda} \to 0, \tag{26}$$

then, with probability 1,

$$\sup_{t \in \mathscr{I}} |\hat{m}(t) - m(t)| \to 0. \tag{27}$$

A derivation is given in the Appendix.

4·3. *Mean integrated squared error*

As in § 4·2, we assume that ρ is constant. Provided we incorporate the design density, f, as a weight, formulae (22)–(25) lead to a simple expression for mean integrated squared error:

$$\int_{\mathscr{I}} E(\hat{m} - m)^2 f = (\rho^2 \lambda)^2 \sum_{j=1}^{\infty} \frac{\theta_j \omega_{1j}^2}{(\rho^2 \theta_j + \lambda)^2} + \frac{\rho^4}{n} \sum_{j=1}^{\infty} \frac{\theta_j \omega_{2j}}{(\rho^2 \theta_j + \lambda)^2}, \tag{28}$$

where $\omega_{1j} = \int_{\mathscr{I}} a\psi_j$ and $\omega_{2j} = \int_{\mathscr{I}} h_j^2$; see the Appendix for a proof, and for derivations of the formula

$$\omega_{1j} = -\rho^{-1} \int_{\mathscr{I}} \beta^0 \psi_j \tag{29}$$

and the following result.

Property 1. If $c_1 \leqslant \inf_{t \in \mathscr{I}} v(t)f(t) \leqslant \sup_{t \in \mathscr{I}} v(t)f(t) \leqslant c_2$ then $\omega_{2j} \in [c_1\theta_j, c_2\theta_j]$ for all $j \geqslant 1$.

The two terms on the right-hand side of (28) denote, respectively, integrated squared bias and integrated variance, weighted in each case by the density f. Together, (28), (29) and Property 1 imply that, provided v is bounded away from zero and infinity, we may write

$$\int_{\mathscr{I}} E(\hat{m} - m)^2 f = (\rho\lambda)^2 \sum_{j=1}^{\infty} \frac{\theta_j (\int_{\mathscr{I}} \beta^0 \psi_j)^2}{(\rho^2 \theta_j + \lambda)^2} + \frac{\chi_1(\lambda)}{n} \sum_{j=1}^{\infty} \frac{\theta_j^2}{(\rho^2 \theta_j + \lambda)^2}, \tag{30}$$

where $\chi_1(\lambda)$ depends on λ, not on n, and is bounded away from zero and infinity uniformly in $\lambda \geqslant 0$.

The first term on the right-hand side of (30) does not depend on n. Provided β^0 is square-integrable, the first term converges to zero at rate $O(\lambda)$ as $\lambda \to 0$. Indeed, since $4\rho^2\lambda\theta_j \leqslant (\rho^2\theta_j + \lambda)^2$,

$$\int_{\mathscr{I}} (\text{bias})^2 f = (\rho\lambda)^2 \sum_{j=1}^{\infty} \frac{\theta_j (\int_{\mathscr{I}} \beta^0 \psi_j)^2}{(\rho^2\theta_j + \lambda)^2} \leqslant \frac{1}{4}\lambda \sum_{j=1}^{\infty} (\int_{\mathscr{I}} \beta^0 \psi_j)^2 = \frac{1}{4}\lambda \int_{\mathscr{I}} (\beta^0)^2,$$

using Parseval's identity (Indritz, 1963, p. 269). Moreover, unless β^0 vanishes, $\int_{\mathscr{I}} (\text{bias})^2 f$ never converges to zero more quickly than $O(\lambda^2)$, and it converges at that rate if and only if the generalised Fourier coefficients $\int_{\mathscr{I}} \beta^0 \psi_j$ converge to zero sufficiently quickly to ensure that $S_2 \equiv \sum_j \theta_j^{-1} (\int_{\mathscr{I}} \beta^0 \psi_j)^2 < \infty$; if this condition holds then $\int_{\mathscr{I}} (\text{bias})^2 f \sim \rho^{-2} S_2 \lambda^2$ as $\lambda \to \infty$.

The second term on the right-hand side of (30) can be written as $\chi(\lambda)/n$, where $\chi(\lambda) = \chi_1(\lambda)\chi_2(\lambda)$, $\chi_2(\lambda) = \sum_j (\rho^2 + \lambda\theta_j^{-1})^{-2}$ and $\chi(\lambda) \to \infty$ at a rate which is completely determined by the rate at which θ_j decreases. The faster the rate at which $\theta_j \to 0$, the slower is the rate at which $\chi(\lambda) \to \infty$. For example, if $\theta_j \asymp j^{-r}$ as $j \to \infty$ then $\chi(\lambda) \asymp \lambda^{-1/r}$ as $\lambda \to 0$, where '$a_r \asymp b_r$', for positive constants a_r and b_r, means that the ratio a_r/b_r is bounded away from zero and infinity as r approaches its limit. Therefore, if $\theta_j \asymp j^{-r}$ then

$$\int_{\mathscr{I}} \text{var}(\hat{m}) f \asymp (n\lambda^{1/r})^{-1} \tag{31}$$

as n increases and λ decreases.

The above bias and variance rates are valid for any generalised spline function ϕ. The next section gives more detailed results for the case of polynomial splines.

4·4. *Results for polynomial splines*

In the particular case of a polynomial spline, where $\phi(t|s) = (t-s)_+^p$, β^0 may be written down explicitly, as follows. Assume that m has $p+1$ square-integrable derivatives. If $\mathscr{I} = [a, b]$ then (13) has the form

$$m^*(t) = \int_a^t \beta^0(s)\rho(s)(t-s)^p ds,$$

which gives, if we differentiate both sides $p+1$ times,

$$\beta^0(t) = \frac{m^{*(p+1)}(t)}{p!\rho(t)} = \frac{m^{(p+1)}(t)}{p!\rho(t)}. \tag{32}$$

Of course, the fact that β^0 is proportional to the $(p+1)$st derivative of m implies that bias vanishes if m is exactly a polynomial of degree p.

The second identity in (32) confirms that the series $\sum_{0 \leqslant k \leqslant p} \beta_k \phi_k$, appearing at (8), plays no role in determining β^0. It also implies that, provided ρ is constant, which we shall assume below without further comment, the crucial condition in Assumption 4 is equivalent to m having $p+1$ well-defined, square-integrable derivatives on $\mathscr{I}$:

$$\int_{\mathscr{I}} (m^{(p+1)})^2 < \infty. \tag{33}$$

Clearly, Assumption 2 holds in the polynomial spline case. We show in the Appendix that, in the polynomial spline case, the operator ψ is nonsingular, as required by Assumption 3. Therefore, provided m satisfies (33), the theory developed in §§ 4·1–4·3 applies directly to polynomial splines.

It may be proved that in this case, provided f also has $p+1$ square-integrable derivatives, the eigenvalues θ_j are bounded between two positive, constant multiples of $j^{-2(p+1)}$, as $j \to \infty$, so that (31) holds for $r = 2(p+1)$. The derivation in the general case is complex and will not be given here. We briefly describe the procedure for the cases $p = 0$, 1, $f \equiv 1$ and $\mathscr{I} = [0, 1]$ here.

If $p = 0$ then the spectral decomposition (21) becomes

$$\psi(u, v) = \frac{1}{3} + \sum_{j=1}^{\infty} (j\pi)^{-2} 2^{\frac{1}{2}} \cos(j\pi u) 2^{\frac{1}{2}} \cos(j\pi v),$$

so that the orthonormal eigenvectors are the constant function 1 and the functions $2^{\frac{1}{2}} \cos(j\pi u)$, for $j \geqslant 1$. When $p = 1$ the eigenvalues can be written as $\theta_j = d_j^{-4}$, where $0 < d_1 < d_2 < \ldots$ denote the successive positive solutions, in $d = d_j$, of the equation $2 + (e^d + e^{-d}) \cos d = 0$. For large j, consecutive values of d_j are successively closer to elements of the sequence of distinct values of $2k\pi \pm \frac{1}{2}\pi$, where k is a positive integer; as a consequence, θ_j is indeed asymptotic to a constant multiple of j^{-4}.

We may deduce from (30) and the fact that $\theta_j \asymp j^{-2(p+1)}$ that

$$\int_{\mathscr{I}} E(\hat{m} - m)^2 f = O\{(n\lambda^{1/r}) + \lambda\}, \tag{34}$$

where $r = 2(p+1)$. Although (34) applies to mean integrated squared error weighted by f, the result in the unweighted case follows directly since we have assumed that f is bounded. The same asymptotic rates for the integrated variance and bias terms have been obtained for smoothing spline regression with penalised squared $(p+1)$th derivative (Cox, 1983).

The right-hand side of (34) is minimised by taking λ to be a constant multiple of $n^{-r/(r+1)}$, which gives a mean squared convergence rate of $O(n^{-r/(r+1)})$. This rate is achieved uniformly over any class of regression means, m, for which $\int_{\mathscr{I}} (m^{(p+1)})^2 \leqslant C$, for any fixed $C > 0$. The rate is optimal for functions with $p+1$ square-integrable derivatives; see Stone (1982). This result shows that, at least in its white-noise representation studied here, penalised spline regression achieves the best possible rate of convergence attainable by nonparametric regression methods.

An exact formula for mean integrated squared error, in the polynomial spline case, may of course be deduced from (28). In the case where $\mathscr{I}$ is the unit interval, implying that $\rho = 1$ since we have assumed that ρ is constant, the errors all have variance equal to σ^2, and the design points are equally spaced on $\mathscr{I}$, or more generally $f \equiv 1$, we have that

$$\int_{\mathscr{I}} E(\hat{m} - m)^2 f = \frac{\lambda^2}{(p!)^2} \sum_{j=1}^{\infty} \frac{\theta_j (\int_{\mathscr{I}} m^{(p+1)} \psi_j)^2}{(\theta_j + \lambda)^2} + \frac{\sigma^2}{n} \sum_{j=1}^{\infty} \frac{\theta_j^2}{(\theta_j + \lambda)^2}. \tag{35}$$

From this expression, it is clear that the variance component of the mean squared error is undefined when $\lambda = 0$. Hence, if a Taylor series around $\lambda = 0$ is used to construct an approximation to the second term on the right-hand side of (35), representing variance, then the first term in its Taylor expansion will be infinite. A similar difficulty is likely to

arise in the discrete data case if K is allowed to grow, making it difficult to justify a Taylor-expansion approximation. This explains the difference between our results and those of Wand (1999).

Formula (35) can potentially be used to choose λ adaptively via a plug-in rule. To be specific, we would construct a pilot estimator of $m^{(p+1)}$, and use the known values of θ_j and ψ_j to approximate the right-hand side of (35). However, the general formula for the eigenvalues ψ_j is too complex for this approach to be attractive.

Clearly, a drawback of the white-noise representation used in this paper is that any effects of number of basis functions on the properties of the estimator are ignored. A further investigation of *P*-splines regression that fully incorporates the effects of both the number of basis functions and the penalty remains an open issue.

Appendix

Technical details

Proof that first $p+1$ *basis terms are not needed in polynomial spline case.* Here we show, by appealing to the white-noise model, that for large K the first $p+1$ components in the basis at (4) are unnecessary. For this it is sufficient to prove that the operator ψ is nonsingular, or equivalently that, for each square-integrable function α which does not vanish almost everywhere, $\iint_{\mathscr{I}^2} \alpha(u)\alpha(v)\psi(u, v)du\,dv$ does not vanish; that is, in view of the definition of ψ at (9), ψ is nonsingular if and only if, whenever the square-integrable function α satisfies

$$\int_{\mathscr{I}} \left\{ \int_{\mathscr{I}} \phi(t|s)\alpha(s)ds \right\}^2 f(t)dt = 0 \quad (\text{for all } t \in \mathscr{I}), \tag{A1}$$

$\alpha = 0$ almost everywhere. In the polynomial spline case, $\int_{\mathscr{I}} \phi(t|s)\alpha(s)ds = (I^p\alpha)(t)$, where I denotes the functional taking α to $I\alpha$, defined by $(I\alpha)(t) = \int_{0<s<t} \alpha(s)ds$. Hence, (A1) implies that $I^p\alpha = 0$ almost everywhere, and therefore that $(d/dt)^p(I^p\alpha)(t) = \alpha(t) = 0$ almost everywhere.

Motivation for Assumption 3. Assume that the generalised spline basis functions $\phi(.|s)$ satisfy Assumption 2 and that Assumption 1 holds. Then the functions $\phi(.|s)$, for $s \in \mathscr{I}$, form a complete basis if any square-integrable function m on $\mathscr{I}$ can be uniquely represented, almost everywhere and in terms of β, in the form

$$m(t) = \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds;$$

compare (13). This is equivalent to asking that, for each m,

$$\int_{\mathscr{I}} \left\{ m(t) - \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds \right\}^2 f(t)dt$$

vanishes for a functional β of m, uniquely determined up to almost everywhere equivalence. A calculus of variations argument now shows that β must be given by the solution of

$$\int_{\mathscr{I}} \beta(s)\rho(s)\psi(s, s_0)ds = \int_{\mathscr{I}} m(t)\phi(t|s_0)f(t)dt \quad (\text{for all } s_0 \in \mathscr{I}),$$

where ψ is given by (9). Therefore, $\beta\rho$ is uniquely defined if and only if Assumption 3 holds. Since ρ is assumed nonvanishing, the same statement applies to β.

Derivation of (17). If we take $n = \infty$ and $\lambda = 0$, which are the values of these quantities in the asymptotic limit, then the right-hand side of (11) vanishes and $y_t = m(t)$, so that (11) becomes

$$\int_{\mathscr{I}} \left\{ m^*(t) - \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds \right\} \phi(t|s_0)f(t)dt = 0. \tag{A2}$$

Let $\beta = \beta^0$ denote the solution of (A2). It follows from the arguments above that, provided Assumptions 1–3 hold, β^0 is given equivalently by (13).

The value of β^0, plus the systematic error, i.e. bias, of $\hat{\beta}$, equals the value of β obtained by setting $n = \infty$ in formula (11) for $\hat{\beta}$. In this case (11) alters to

$$\int_{\mathscr{I}} \left\{ m^*(t) - \int_{\mathscr{I}} \beta(s)\rho(s)\phi(t|s)ds \right\} \phi(t|s_0)f(t)dt = \lambda\beta(s_0)/\rho(s_0). \tag{A3}$$

Let β_1 denote the solution of (A3). If we start from (A3), it can be derived that $\beta_1 - \beta^0 = \lambda\delta$, where δ is the solution of (14). Finally, we write $\hat{\beta} = \beta_1 + \Delta$, where Δ solves

$$\int_{\mathscr{I}} \left\{ e_t - \int_{\mathscr{I}} \Delta(s)\rho(s)\phi(t|s)ds \right\} \phi(t|s_0)f(t)dt = \lambda\Delta(s_0)/\rho(s_0).$$

Equivalently, Δ is the solution of (15).

Derivation of (19). To obtain the formula for β^0 in (19), note that, if we multiply both sides of (13) by $\phi(t|u)f(t)$ and integrate over $t \in \mathscr{I}$, we obtain $b(u) = \int_{\mathscr{I}} \beta^0(s)\rho(s)\psi(u, s)ds$. Expanding b on the left-hand side in its generalised Fourier series corresponding to the basis $\psi_1, \psi_2, \ldots$, and expanding ψ on the right-hand side using (21), we deduce that $\sum_j (\int_{\mathscr{I}} b\psi_j)\psi_j = \sum \theta_j (\int_{\mathscr{I}} \beta^0\rho\psi_j)\psi_j$. Equating coefficients of ψ_j we deduce that $\int_{\mathscr{I}} \beta^0\rho\psi_j = \theta_j^{-1} \int_{\mathscr{I}} b\psi_j$. Therefore, $\beta^0\rho = \sum_j \theta_j^{-1} (\int_{\mathscr{I}} b\psi_j)\psi_j$.

Next we derive the formulae for δ and Δ in (19). Note that (14) and (15) have the form

$$\rho \int_{\mathscr{I}} \xi(s)\psi(s, t)ds + \rho^{-1}\lambda\xi(t) = \eta(t) \quad \text{(for all } t \in \mathscr{I}\text{)}, \tag{A4}$$

where the equation is to be solved for ξ. If we express ξ and η in terms of the basis $\{\psi_j\}$ as $\xi = \sum_{j \geqslant 1} \xi_j\psi_j$ and $\eta = \sum_{j \geqslant 1} \eta_j\psi_j$, respectively, (A4) is identical to

$$\rho \sum_{j=1}^{\infty} \theta_j\xi_j\psi_j + \rho^{-1}\lambda \sum_{j=1}^{\infty} \xi_j\psi_j = \sum_{j=1}^{\infty} \eta_j\psi_j.$$

Equating coefficients of ψ_j we deduce that $\xi_j = \eta_j/(\rho\theta_j + \rho^{-1}\lambda)$, or that

$$\xi = \rho \sum_{j=1}^{\infty} \frac{\int_{\mathscr{I}} \eta\psi_j}{\rho^2\theta_j + \lambda} \psi_j. \tag{A5}$$

Taking $(\xi, \eta) = (\delta, a)$ or (Δ, A) we see that the formulae for δ and Δ in (19) follow from (A5).

Derivation of (27). In view of (18) and (19), we have for each $t \in \mathscr{I}$ that

$$\begin{aligned} |E\hat{m}(t) - m(t)| &= \lambda\rho \left| \int_{\mathscr{I}} \delta(s)\psi(s, t)ds \right| = \lambda\rho \left| \sum_{j=1}^{\infty} \frac{\int_{\mathscr{I}} \beta^0\psi_j}{\rho^2\theta_j + \lambda} \int_{\mathscr{I}} \psi_j(s)\psi(s, t)ds \right| \\ &\leqslant \lambda\rho\alpha(t) \sum_{j=1}^{\infty} \frac{|\int_{\mathscr{I}} \beta^0\psi_j|}{\rho^2\theta_j + \lambda} = O\left(\sum_{j=1}^{\infty} \frac{\lambda|\int_{\mathscr{I}} \beta^0\psi_j|}{\theta_j + \lambda} \right), \end{aligned} \tag{A6}$$

with

$$\alpha(t)^2 = \int_{\mathscr{I}} \int_{\mathscr{I}} \phi(v|s)^2\phi(v|t)^2 f(v)dvds < \infty$$

by Assumption 2, so that the last identity in (A6) holds uniformly in $t \in \mathscr{I}$. Therefore, provided that $\sum_j |\int_{\mathscr{I}} \beta^0\psi_j| < \infty$,

$$\sup_{t \in \mathscr{I}} |E\hat{m}(t) - m(t)| \to 0 \tag{A7}$$

as $\lambda \to 0$.

From the definition of the N_j and the Borel–Cantelli Lemma, it follows that

$$\sup_{j \geqslant 1} N_j^2 (EN_j^2 \log j)^{-1} < \infty$$

with probability 1. Therefore, noting (22), (24), and (25) and using Assumption 2, we deduce that

$$|\hat{m}(t) - E\hat{m}(t)| = O\left[n^{-1/2} \sum_{j=1}^{\infty} \frac{\{(\log j) \int_{\mathscr{I}} h_j^2\}^{1/2}}{\theta_j + \lambda} \right] = O\left\{ n^{-1/2} \sum_{j=1}^{\infty} \frac{(\theta_j \log j)^{1/2}}{\theta_j + \lambda} \right\},$$

uniformly in t, with probability 1. Hence, provided (26) holds, $\sup_{\mathscr{I}} |\hat{m} - E\hat{m}| \to 0$ with probability 1. Property (27) follows from this result and (A7).

Derivation of (28), (29) *and Property* 1. For any real numbers $r_1, r_2, \ldots$ such that the function $r(t) = \sum_r r_j \int_{\mathscr{I}} \psi_j(s) \phi(t|s) ds$ is well-defined in an L_2 sense,

$$\int_{\mathscr{I}} r(t)^2 f(t) dt = \sum_{j=1}^{\infty} \theta_j r_j^2.$$

Applying this result twice with $r_j = (\rho^2 \theta_j + \lambda)^{-1} \int_{\mathscr{I}} a\psi_j$ and $r_j = (\rho^2 \theta_j + \lambda)^{-1} \int_{\mathscr{I}} A\psi_j$, we deduce (28) from (22)–(24) and Assumption 5.

To derive (29), note that, by the definition of the function a given in § 4·1,

$$-\rho^2 \omega_{1j} = -\rho^2 \int_{\mathscr{I}} a\psi_j = \rho \int_{\mathscr{I}} \beta^0 \psi_j = \theta_j^{-1} \int_{\mathscr{I}} b\psi_j,$$

where the last identity uses the first formula in (19). To obtain Property 1, observe that

$$\omega_{2j} = \int_{\mathscr{I}} h_j^2 = \int_{\mathscr{I}} \left\{ \int_{\mathscr{I}} \phi(s|t) \psi_j dt \right\}^2 v(s) f(s)^2 ds,$$

and so if $c_1 \leqslant \inf_{\mathscr{I}} vf \leqslant \sup_{\mathscr{I}} vf \leqslant c_2$ then ω_{2j} lies between multiples c_1 and c_2 of

$$\int_{\mathscr{I}} \left\{ \int_{\mathscr{I}} \phi(s|t) \psi_j dt \right\}^2 f(s) ds = \iint_{\mathscr{I}^2} \psi_j(t_1) \psi_j(t_2) \psi(t_1, t_2) dt_1 dt_2 = \theta_j.$$

References

Aerts, M., Claeskens, G. & Wand, M. (2002). Some theory for penalized spline generalized additive models. *J. Statist. Plan. Infer.* **103**, 455–70.

Brown, L. D. & Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384–98.

Brown, L. D., Cai, T. T., Low, M. G. & Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30**, 688–707.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *Ann. Statist.* **11**, 530–51.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Huang, J. Z. & Stone, C. J. (2003). Extended linear modelling with splines. In *Nonlinear Estimation and Classification*, Ed. D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick and B. Yu, pp. 213–34. New York: Springer.

Indritz, J. (1963). *Methods in Analysis*. New York: Macmillan.

Nychka, D. & Cummins, D. (1996). Comment on 'Flexible smoothing with B-splines and penalties' by P. H. C. Eilers and B. D. Marx. *Statist. Sci.* **89**, 104–5.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with Discussion). *Statist. Soc.* **1**, 502–27.

Parker, R. L. & Rice, J. A. (1985). Comment on 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting'. *J. R. Statist. Soc.* B **47**, 40–2.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comp. Graph. Statist.* **11**, 735–57.
RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.
SPEED, T. (1991). Comment on 'That BLUP is a good thing: the estimation of random effects'. *Statist. Sci.* **6**, 42–4.
STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–53.
WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: SIAM.
WAND, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika* **86**, 936–40.
WAND, M. P. (2003). Smoothing and mixed models. *Comp. Statist.* **18**, 223–49.
WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

[*Received October* 2003. *Revised May* 2004]