Statistical Modelling 2002; **4**: 333–349

Flexible smoothing with *P*-splines: a unified approach

ID Currie[1] and M Durban[2]
[1] Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK
[2] Departamento de Estadistica y Econometria, Universidad Carlos III de Madrid, Madrid, Spain

Abstract: We consider the application of *P*-splines (Eilers and Marx, 1996) to three classes of models with smooth components: semiparametric models, models with serially correlated errors, and models with heteroscedastic errors. We show that *P*-splines provide a common approach to these problems. We set out a simple nonparametric strategy for the choice of the *P*-spline parameters (the number of knots, the degree of the *P*-spline, and the order of the penalty) and use mixed model (REML) methods for smoothing parameter selection. We give an example of a model in each of the three classes and analyse appropriate data sets.

Key words: heterogeneity; mixed models; *P*-splines; REML; semiparametric models; serial correlation.

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received: October 2001; **revisited:** June 2002, October 2002; **accepted:** October 2002

1 Introduction

There are several approaches to nonparametric modelling, among which spline methods are well known for their flexibility to fit curves without choosing in advance a rigid form for the underlying function. We focus on regression splines, which can be seen as a compromise between linear regression and nonparametric regression models; we concentrate on low-rank smoothers and, more specifically, on *B*-splines with penalties, known as *P*-splines (Eilers and Marx, 1996). Low-rank smoothers offer significant computational advantages, since they involve inversion of matrices of order less than the number, n, of data points (smoothing splines use matrices of order n), and so they can allow the analysis of larger data sets. Eilers and Marx (1996) showed how *P*-splines can be used in many different contexts and illustrate their remarks with examples on density estimation and nonparametric smoothing.

We are attracted by the idea that a smoothly varying trend can be thought of as the sum of a polynomial trend of low degree (usually linear or quadratic) and a random component. This approach leads to a mixed model representation that allows (a) the level of smoothing to be chosen by residual maximum likelihood (REML) as an alternative to the Akaike information criterion (AIC) or generalized cross-validation (GCV) (which are known to undersmooth the data, especially when errors are correlated (Altman, 1990)), and (b) the performance of likelihood ratio tests for model selection.

Several authors (Speed, 1991; Wang, 1998a; Zhang *et al.*, 1998; Brumback and Rice, 1998; Verbyla *et al.*, 1999) have derived a connection between smoothing splines and a linear mixed effects model. Wang (1998b) extended the mixed model setup to cases when

Address for correspondence: ID Currie, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, UK. E-mail: I.D.Currie@ma.hw.ac.uk

10.1191/1471082x02st039ob

errors were correlated, and Lin and Zhang (1999) introduced generalized additive mixed models to cope with overdispersed and correlated non-Gaussian data. Similar work has been developed for penalized splines. Brumback *et al.* (1999) introduced the best linear unbiased predictor (BLUP) representation of penalized splines, Coull *et al.* (2001a) used penalized splines in the context of additive models and Coull *et al.* (2001b) extended this model to allow for correlated errors. Parise *et al.* (2001) and Aerts *et al.* (2002) applied the mixed model approach to generalized additive models. The work presented in these papers uses truncated lines as the *P*-spline basis. Although truncated lines have a simple formulation, it is known that they do not have optimal numerical properties (Aerts *et al.*, 2002). In this paper we use the original *B*-spline basis of Eilers and Marx (1996) (which has better numerical properties and is preferable for computation) and show that *P*-splines also have a mixed model representation with this basis. We focus on three classes of models: semiparametric models, models with serially correlated errors, and models with heteroscedastic errors. These mixed models based on *B*-splines are currently being extended to additive models (Durban and Currie, 2002), and generalized additive models and smoothing models in two dimensions (Durban *et al.*, 2002).

The layout of the paper is as follows. In section 2 we extend the basic theory of *P*-splines to include both ordinary regression terms and an arbitrary error variance; in practice, for the latter, we have in mind either a simple time series error structure, e.g., serial correlation such as AR(1) or AR(2), or independent but heteroscedastic errors. In section 2 we also consider the choice of the *P*-spline parameters *ndx* (the number of knots parameter), *bdeg* (the degree of the *B*-spline basis) and *pord* (the order of the penalty). In section 3 we show how the extended model of section 2 can be expressed as a mixed model. Section 4 is the main part of the paper with applications which illustrate the theory laid out earlier. The paper concludes with a brief discussion in section 5.

2 A general approach to modelling with *P*-splines

Suppose the variable $\boldsymbol{y}' = (y_1, \ldots, y_n)$ depends smoothly on a single variable $\boldsymbol{x}' = (x_1, \ldots, x_n)$ then the nonparametric model for $\boldsymbol{y}$ can be written $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $f(\cdot)$ is a smoothly varying function and $\boldsymbol{\epsilon}$ is a vector of independent errors with variance $\sigma^2\boldsymbol{I}$. Eilers and Marx (1996) make the following two assumptions: first, they assume that $E(\boldsymbol{y}) = \boldsymbol{Ba}$ where $\boldsymbol{B} = (B_1(\boldsymbol{x}), B_2(\boldsymbol{x}), \ldots, B_k(\boldsymbol{x}))$ is an $n \times k$ matrix of *B*-splines (k depends on the number of knots and the degree of the *B*-spline), and $\boldsymbol{a}$ is the vector of regression coefficients; secondly, they suppose that the coefficients of adjacent *B*-splines satisfy certain smoothness conditions that can be expressed in terms of finite differences of the a_is. Thus, from a least-squares perspective, the coefficients, $\boldsymbol{a}$, are chosen to minimise

$$S(\boldsymbol{a}) = (\boldsymbol{y} - \boldsymbol{Ba})'(\boldsymbol{y} - \boldsymbol{Ba}) + \lambda \boldsymbol{a}'\boldsymbol{D}'\boldsymbol{Da} \tag{2.1}$$

where $\boldsymbol{D}$ is a difference matrix and λ is a penalty. For given λ, the solution to this optimisation problem satisfies

$$(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})\hat{\boldsymbol{a}} = \boldsymbol{B}'\boldsymbol{y} \tag{2.2}$$

and then $\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{a}} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H}$ is the hat-matrix:

$$\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'. \tag{2.3}$$

Expressions (2.1) and (2.3) are at the heart of the reason why we find *P*-splines such an attractive and transparent method of smoothing. First, the term $(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a})$ in (2.1) corresponds to ordinary regression on the columns of $\boldsymbol{B}$ while the term $\lambda \boldsymbol{a}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{a}$ looks after the overparameterisation of the regression function by placing a penalty on the smoothness of the a_j. Secondly, there is a computational advantage with a small number of knots and this advantage may be substantial with large n since $\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}$ in (2.3) is $k \times k$. One important consequence of the *P*-spline approach is that the problem of choosing the number and position of the knots is largely overcome; as long as we choose enough knots the penalty function should ensure that the resulting fits are very similar. We discuss the choice of the number of knots later in this section.

We now consider the more general model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + f(\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $\boldsymbol{X}\boldsymbol{\beta}$ is an ordinary regression term and $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$ for some variance matrix $\boldsymbol{\Sigma}$; in section 4 we will choose $\boldsymbol{\Sigma}$ to give either a model with serial correlation or one with heteroscedastic errors. Corresponding to (2.1) we find $\boldsymbol{\beta}$ and $\boldsymbol{a}$ by minimising

$$S(\boldsymbol{a}, \boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{a})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{a}) + \lambda \boldsymbol{a}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{a}. \tag{2.4}$$

We think of this as generalised least squares with a penalty on the smoothness of the a_j. If there is no regression term, i.e., if $\boldsymbol{X} = \boldsymbol{0}$, then the hat-matrix corresponding to (2.3) is

$$\boldsymbol{H}_V = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{V}^{-1}. \tag{2.5}$$

For given λ the solution to the optimisation problem (2.4) satisfies the linear system

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{B} \\ \boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{V}^{-1} \\ \boldsymbol{B}'\boldsymbol{V}^{-1} \end{bmatrix} \boldsymbol{y} \tag{2.6}$$

with hat-matrix $\boldsymbol{H}_V(\boldsymbol{X})$ given by

$$\boldsymbol{H}_V(\boldsymbol{X}) = \boldsymbol{H}_V + (\boldsymbol{I} - \boldsymbol{H}_V)\boldsymbol{X}\,[\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{H}_V)\boldsymbol{X}\,]^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{H}_V) \tag{2.7}$$

where $\boldsymbol{H}_V$ is defined in (2.5).

There are a number of choices that are implicit in the above development: the number of knots, *ndx*, (strictly $ndx - 1$ is the number of internal knots in the domain of x), the degree of the *B*-spline basis, *bdeg*, and the order of the penalty, *pord*. We outline two approaches to the choice of these parameters. The first approach is very simple: use one knot for every four or five observations up to a maximum of about 40 knots as suggested by Ruppert (2002); use cubic splines ($bdeg = 3$) and a quadratic penalty ($pord = 2$). These choices should work well in most examples and are reminiscent of what happens with cubic smoothing splines where a cubic basis and a quadratic penalty are used routinely.

The second approach is more systematic. We note first that it is not possible to use REML to choose *ndx*, *bdeg* and *pord* since different values of these parameters result in non-nested models. Instead, we follow Eilers and Marx (1996) who recommend plotting the AIC criterion against the trace of the hat-matrix $\boldsymbol{H}$ for various values of *ndx*, *bdeg*, and *pord*. The minimum position on this plot corresponds to optimal values of *ndx*, *bdeg*, and *pord*. However, this is unlikely to perform well in the semiparametric setting where there is considerable evidence that criteria such as AIC and GCV tend to lead to undersmoothing (Hurvich *et al.*, 1998). In the notation of Hurvich *et al.* we minimise criteria of the form $\log(\hat{\sigma}^2) + \psi(\boldsymbol{H})$ where $\hat{\sigma}^2 = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})^2\boldsymbol{y}/n$, $\psi(\cdot)$ is a penalty function and $\boldsymbol{H} = \boldsymbol{H}_V(\boldsymbol{X})$ is the hat-matrix defined in (2.7). The AIC_C criterion of Hurvich *et al.* introduces an additional penalty over GCV and AIC to account for increased model complexity and is designed to deal with the undersmoothing problem in the semiparametric setting. The penalty functions $\psi(\cdot)$ for GCV, AIC, and AIC_C are $-2\log(1-t)$, $2t$ and $(2t+\delta)/(1-t-\delta)$, respectively, where $t = \mathrm{tr}(\boldsymbol{H})/n$ and $\delta = 2/n$. We will illustrate the choice of the *P*-spline parameters *ndx*, *bdeg*, and *pord* with plots of the AIC_C criterion against $\mathrm{tr}(\boldsymbol{H})$ when we present our examples in section 4.

3 P-splines as mixed models

The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{a}$ in (2.6) have been obtained from a penalised least-squares perspective. However, equation (2.6) suggests a set of mixed model equations (Searle *et al.*, 1992, p. 276). Brumback *et al.* (1999) expressed *P*-splines in terms of truncated polynomials and used a mixed model approach to estimate the smoothing parameter. We now show that with our approach equation (2.6) also has a mixed model formulation. Thus we can take advantage simultaneously of existing mixed model theory and the good numerical properties of *B*-splines.

The idea is to express the nonparametric trend $\boldsymbol{Ba}$ as the sum of a fixed and random effect; this mirrors the well-known decomposition for the cubic smoothing spline, which can be expressed as the sum of a fixed linear effect (since the penalty is order 2 for a cubic smoothing spline) and a random effect; see Verbyla *et al.* (1999). We want to write $\boldsymbol{a} = \boldsymbol{Gb} + \boldsymbol{Zu}$ where $[\boldsymbol{G} : \boldsymbol{Z}]$ is square and non-singular and $\boldsymbol{G}$ and $\boldsymbol{Z}$ are such that, when $\boldsymbol{Gb} + \boldsymbol{Zu}$ is substituted for $\boldsymbol{a}$ in (2.6), a set of mixed model equations results. A suitable choice of $\boldsymbol{G}$ and $\boldsymbol{Z}$ is as follows: let $\boldsymbol{g}' = (1, 2, \ldots, k)$ where k is the number of columns of $\boldsymbol{B}$, and define $\boldsymbol{G} = (\boldsymbol{1}, \boldsymbol{g}, \boldsymbol{g}^2, \ldots, \boldsymbol{g}^{q-1})$ where q is the order of the penalty; define $\boldsymbol{Z} = \boldsymbol{D}'(\boldsymbol{DD}')^{-1}$. Then $[\boldsymbol{G} : \boldsymbol{Z}]$ is square and non-singular; furthermore, equation (2.6) reduces to the mixed model equations for the mixed model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{BGb} + \boldsymbol{BZu} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are fixed effects, $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma_u^2\boldsymbol{I})$ is a random effect and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$. The penalty λ is given by the ratio of the variance components, i.e., $\lambda = \sigma^2/\sigma_u^2$. Note that the substitution $\boldsymbol{a} = \boldsymbol{Gb} + \boldsymbol{Zu}$ decomposes the trend $\boldsymbol{Ba}$ into a fixed polynomial trend $\boldsymbol{BGb}$ of degree $q-1$ and a random component $\boldsymbol{BZu}$.

Mixed model methodology can now be used to estimate the smoothing parameter since the residual likelihood (REML) leads to estimates of the variance components σ_u^2 and σ^2 and so, in terms of λ and σ^2, we maximise

$$l(\sigma^2, \lambda) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}}| - \frac{1}{2}\boldsymbol{y}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1})\boldsymbol{y} \tag{3.2}$$

where, from (3.1), $\tilde{\boldsymbol{X}} \rightarrow [\boldsymbol{X} : \boldsymbol{BG}]$ and $\boldsymbol{\Sigma} \rightarrow \sigma^2(\boldsymbol{V} + \lambda^{-1}\boldsymbol{BZZ}'\boldsymbol{B}')$. Our calculation of (3.2) uses the QR decomposition (Pinheiro and Bates, 2000), which not only speeds up the calculation, but also, in the case that $\boldsymbol{V}$ is diagonal, avoids the inversion of the potentially large matrix in this expression.

4 Some examples

In this section we demonstrate the versatility of the above theory by looking at three examples: a randomised block design, a data set with heteroscedastic errors, and one with serial correlation.

4.1 A randomised block design

We use some data from a sugar beet breeding programme at the Plant Breeding Institute, Cambridge, UK, in the late 1970s. This large data set was analysed by Durban *et al.* (2001) where there was particular interest in the joint modelling of fertility trend and interplot competition. For simplicity, in this paper we use data from one of the trials, which exhibited trend but not competition. This trial consisted of three replicate blocks of 36 varieties with each block laid out in a linear array of plots. Tractor wheelings passed between each set of six plots as follows: between plots 1 and 2, and 5 and 6, between plots 7 and 8, and 11 and 12, etc. Arnold and Kempton (1979) showed that such wheelings can give substantial advantage to the centre plots 3 and 4, 9 and 10, etc, over the outer plots 1 and 6, 7 and 12, etc., so we allowed for this by fitting a factor `wheel` with values 1,2,3,3,2,1, etc. The analysis of variance table for the model

$$\texttt{yield} = \texttt{variety} + \texttt{block} + \texttt{wheel} + \texttt{linear} + \texttt{error} \tag{4.1}$$

is given in Table 1 and the variety effects, the wheel effects and the linear effects of plots within blocks are all highly significant. However, a plot of the residuals from (4.1) suggests that there may be some non-linear trend across plots and we investigate this by modelling this trend with *P*-splines.

Table 1 ANOVA for `yield = variety + block + wheel + linear + error`

Source	SS	DF	MS	F
Varieties	4426.4	35	126.5	11.7
Blocks	8.9	2	4.5	0.4
Wheel	128.0	2	64.0	5.9
Linear	137.8	3	45.9	4.2
Residual	704.9	65	10.8	

SS = Sum of squares, DF = Degrees of freedom, MS = Mean squares, F = F statistic.

In the case of a randomised block design with r blocks we want to fit a separate smooth function $\boldsymbol{Ba}_i$, $i = 1, \ldots, r$, to model underlying fertility in each block. We choose $\boldsymbol{\theta}$, the vector of variety and wheel parameters, and $\boldsymbol{a}' = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_r')$ by minimising

$$S(\boldsymbol{a}, \boldsymbol{\theta}) = \sum_{i=1}^{r} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\theta} - \boldsymbol{Ba}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\theta} - \boldsymbol{Ba}_i) + \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{a}_i. \tag{4.2}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ are the yields and design matrix for block i. We make two practical points with regard to (4.2). First, since $\boldsymbol{B}1 = 1$, it follows that $\boldsymbol{H}1 = 1$, and so 1 must not be in the span of $\boldsymbol{X}$; in particular, intercept and block effects must not be fitted in $\boldsymbol{X}$ since they will be fitted in the trend $\boldsymbol{Ba}_i$. Secondly, expression (4.2) allows us to have a distinct penalty λ_i for each block should each block require smoothing on a different scale. In our example we have $r = 3$, so (2.6) becomes

$$\boldsymbol{V} \to \boldsymbol{I}, \quad \boldsymbol{B} \to \begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}, \quad \lambda\boldsymbol{D}'\boldsymbol{D} \to \begin{bmatrix} \lambda_1\boldsymbol{D}'\boldsymbol{D} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_2\mathbf{D}'\mathbf{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \lambda_3\boldsymbol{D}'\boldsymbol{D} \end{bmatrix}, \tag{4.3}$$

while if we take a mixed model approach then we use (3.2) with

$$\tilde{\boldsymbol{X}} \to \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{BG} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{0} & \boldsymbol{BG} & \boldsymbol{0} \\ \boldsymbol{X}_3 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{BG} \end{bmatrix} \tag{4.4}$$

and

$$\boldsymbol{\Sigma} \to \sigma^2 \left(\boldsymbol{I} + \begin{bmatrix} \lambda_1\boldsymbol{BZZ}'\boldsymbol{B}' & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_2\boldsymbol{BZZ}'\boldsymbol{B}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \lambda_3\boldsymbol{BZZ}'\boldsymbol{B}' \end{bmatrix} \right). \tag{4.5}$$

We investigate appropriate values of *ndx*, *bdeg*, and *pord* by plotting AIC_C against $\mathrm{tr}(\boldsymbol{H})$, as discussed in section 2. Figure 1 suggests that we should use $ndx = 5$ (i.e., 4 internal knots per block), $bdeg = 3$, and $pord = 1$. Figure 1 also illustrates the computational advantage that *P*-splines can provide. The main computational demand comes from the matrix on the left side of (2.6) which is 61×61 when $ndx = 5$ but 154×154 when $ndx = 36$ (with

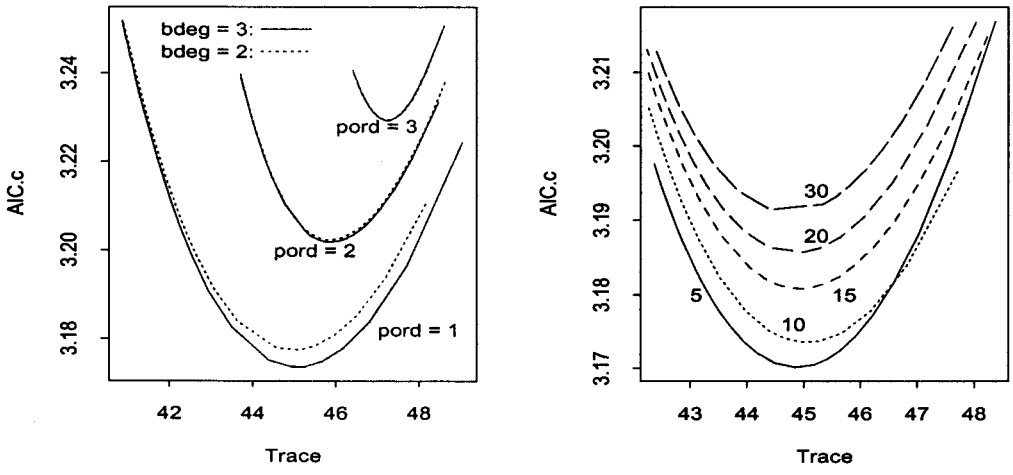

Figure 1 AIC_C *vs* tr(H) plots. Left panel: $ndx = 10$, $bdeg = 2, 3$, $pord = 1, 2, 3$; right panel: $bdeg = 3$, $pord = 1$, $ndx = 5, 10, 15, 20, 30$

$bdeg = 3$, as in the case of cubic smoothing splines); of course with the appropriate penalty both these fitted models have an effective dimension of approximately 45, i.e., three degrees of freedom (df) for the non-linear block effects. Indeed, all preferred models in Figure 1 fit about three degrees of freedom for the non-linear effects whatever the values of *ndx*, *bdeg*, and *pord* so, at least at this level, all selected models are roughly equivalent.

We now use REML for smoothing parameter selection within the family of models with $ndx = 5$, $bdeg = 3$, and $pord = 1$. Suppose first we use a common value of λ across plots, then we find $\hat{\lambda} = 0.82$ with $\mathrm{tr}(\boldsymbol{H}_X) = 47.2$, slightly larger than the model selected by AIC_C with $\mathrm{tr}(\boldsymbol{H}_X) = 45$. The left panels in Figure 2 show fitted trends and residuals for the model $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{Ba}$ while the right panels show a standardised residual plot. The trend across plots is particularly strong in block 3 while the residual plot looks satisfactory.

This analysis assumes a common λ across plots within blocks but it is possible that the trend in each block might require smoothing on a different scale, that is, distinct λ. With distinct λ we find $\hat{\lambda}_1 = 0.18$, $\hat{\lambda}_2 = \infty$, and $\hat{\lambda}_3 = 0.64$ and the fitted trends are shown on Figure 2. One advantage of the mixed model approach is that we can use REML to conduct formal tests of the nested hypotheses: $H_1 : \lambda_1 = \lambda_2 = \lambda_3 = \infty$ (flat trend since $pord = 1$), $H_2 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda$, (common λ) and $H_3 : \lambda_1 \neq \lambda_2 \neq \lambda_3$ (distinct non-linear trend). The parameter is on the boundary of the parameter space under H_1 ($\lambda = \infty \Leftrightarrow \sigma_u^2 = 0$), and thus the asymptotic null distribution of the likelihood ratio test statistic is a 50 : 50 mixture of χ_0^2 and χ_1^2 (Liang and Self, 1996), that is, the distribution of the statistic is the same as that of $Z^2 I(Z > 0)$ (where Z has the standard normal distribution and I is the indicator function). The test statistic computed from (3.2) for H_1 *vs* H_2 is 11.78 and the 95% point of the mixture is 2.706; for H_2 *vs* H_3 it is 3.50 with 2 df. The non-linear trend in Figure 2 is confirmed, but the evidence that we need distinct λ is not strong.

What is the effect of modelling trend on estimated variety means and the standard errors of differences? The effect on estimated variety means and varietal rankings can be substantial. For example, the largest reduction in the estimated mean is for variety 34,

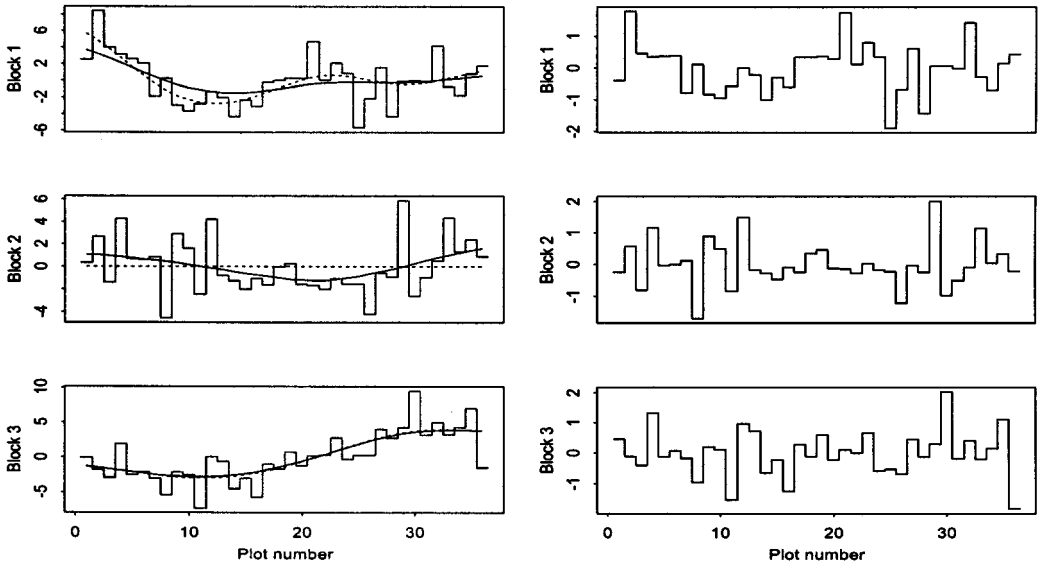

Figure 2 Model $\mathbf{y} = \boldsymbol{X}\theta + \boldsymbol{Ba}$ fitted by REML ($\hat{\lambda} = 0.82$, $\hat{\sigma}^2 = 8.37$) with $ndx = 5$, $bdeg = 3$, and $pord = 1$. Left panels: (fitted trends (—) $\hat{\boldsymbol{f}}_i = \boldsymbol{B}\hat{\boldsymbol{a}}_i$ and residuals $\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\theta}}$ fitted trends with $\hat{\lambda}_1 = 0.18$, $\hat{\lambda}_2 = \infty$, $\hat{\lambda}_3 = 0.64$ are shown (– – –)); right panels: residuals $(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{f}}_i)/\hat{\sigma}$, $i = 1, 2, 3$

which is changed from 10.87 kg to 8.14 kg; the associated change in rank is from 12th from bottom to 6th from bottom. The randomisation assigned variety 34 plot positions 2, 34, and 36 in blocks 1, 2, and 3, respectively, and Figure 2 confirms that these are advantageous plot positions. Modelling trend will also improve the precision of the experiment. The standard error of the estimate of varietal differences (SED) is 2.87 kg if the basic ANOVA model `variety+block+wheel` is fitted. There is little gain in efficiency if linear effects for blocks are fitted (the average SED is 2.75 kg). However, if trend is accounted for by the model specified in Figure 2 then the average SED is 2.50 kg, an efficiency gain of around 10% over simple blocks.

4.2 Smoothing and heteroscedasticity

The data are taken from a simulated experiment to test crash helmets and give $n = 133$ head accelerations in units of g and times in milliseconds (ms) after impact. Silverman (1985) used these data to illustrate smoothing in the presence of heteroscedastic errors; the data set is given in full by Hand *et al.* (1994, p. 276).

A plot of the accelerations against time shows clear evidence of heteroscedastic errors (see Figure 4). We used AIC_C *vs* $\mathrm{tr}(\boldsymbol{H})$ plots similar to Figure 1 with $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ to select the basic *P*-spline parameters *ndx*, *bdeg*, and *pord*. There was clear evidence that $ndx > 5$ and some evidence against $pord = 1$, and our final selection was $ndx = 20$, $bdeg = 3$, and $pord = 2$, very similar to the values obtained with Ruppert's (2002) rule of thumb mentioned in section 2.

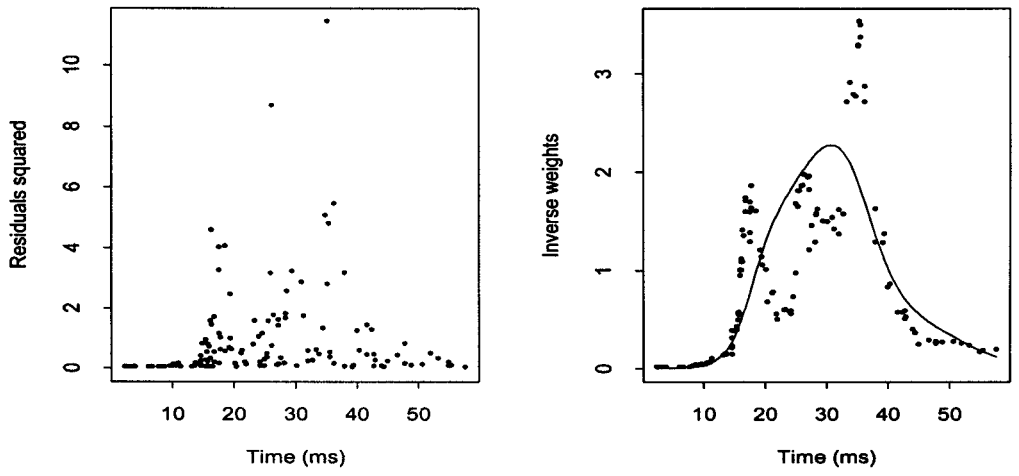

Figure 3 Left panel: squared residuals r_i^2; (right panel: reciprocals of weights, w_i^{-1}, given by smooth of squared) residuals r_i^2 (a) *P*-splines (–) (b) moving average (···)

The left panel of Figure 3 shows the squared residuals

$$r_i^2 = (y_i - \hat{y}_i)^2/\hat{\sigma}^2 \tag{4.6}$$

from the preliminary fit. This suggests that we fit the model $\boldsymbol{y} = \boldsymbol{Ba} + \boldsymbol{\epsilon}$ but with $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V} = \boldsymbol{W}^{-1}$ and $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_n)$ is a diagonal matrix of weights. Appropriate weights w_i are proportional to reciprocals of local variances and precise estimation of the weights is not necessary; for convenience, we take $\sum w_i^{-1} = n$. We used *P*-splines to smooth $R_i = \log r_i^2$ and then took $w_i^{-1} \propto \exp(\hat{R}_i)$. The right panel of Figure 3 shows the reciprocals of the resulting weights. Silverman used reciprocals of a moving average of the r_i^2 to estimate the weights and reciprocals of these weights are also shown.

We now use REML to refit the model where $\tilde{\boldsymbol{X}} \to \boldsymbol{BG}$ and $\boldsymbol{\Sigma} \to \sigma^2(\boldsymbol{W}^{-1} + \lambda^{-1}\boldsymbol{BZZ}'\boldsymbol{B}')$ in (3.2). We find $\hat{\lambda} = 0.344$ with $\hat{\sigma}^2 = 508.1$, not greatly changed from the unweighted estimates, and Figure 4 shows the weighted fit together with a 95% confidence interval. With moving average weights we found $\hat{\lambda} = 0.282$ and $\hat{\sigma}^2 = 445.6$; the resulting fit was almost identical to the *P*-spline fit. Silverman (1985) used smoothing splines with generalised cross validation to choose the smoothing parameter. Our fit and confidence intervals are very similar to those of Silverman.

4.3 Smoothing and serial correlation

It is well known that automatic smoothing parameter selection methods such as AIC and GCV lead to underestimation of the smoothing parameter in the presence of positive serial correlation (Wang 1998b). In this section we show that the general theory laid out above takes care of the undersmoothing problem. First, we present a simulation study in which we

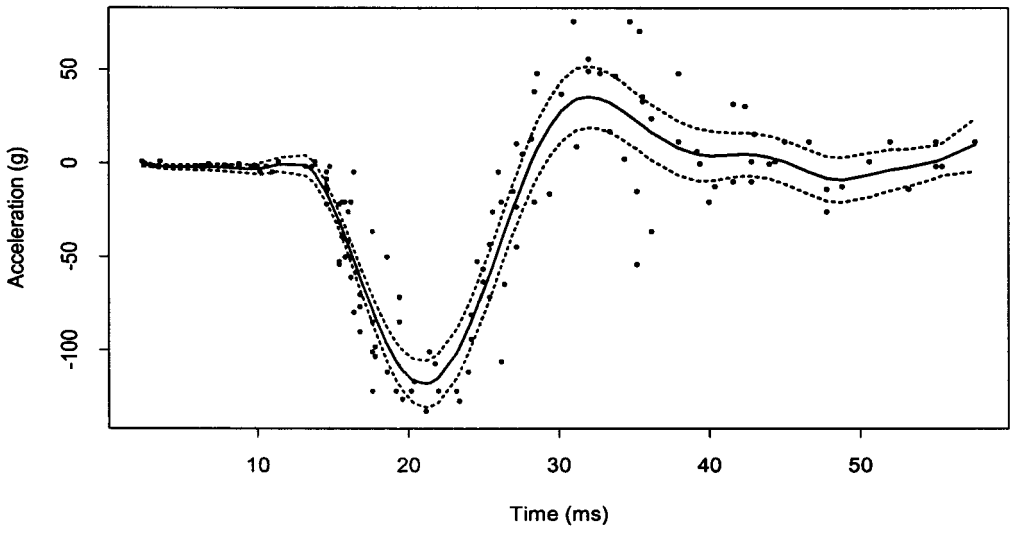

Figure 4 Fitted smooth, weighted fit ($\hat{\lambda} = 0.344$, $\hat{\sigma}^2 = 508.1$), with approximate pointwise 95% confidence interval

use *P*-splines to estimate an underlying trend in serially correlated data. Secondly, we analyse data on the profile of a block of wood that has been subject to a grinding process.

4.3.1 A simulation study

We simulated data from the model

$$y_i = \sin(2\pi i/n) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4.7}$$

where the ϵ_i are generated by a first-order autoregressive process with mean 0, standard deviation σ, and first-order correlation ρ; this is the same basic scheme as used by Wang (1998b). The aims of our study were threefold: first to check if *ndx*, the number of knots parameter, has an influence on the estimation of the correlation parameter and the amount of smoothness of the fitted curve; second, to compare the performance of REML and the modified AIC criterion in the selection of the smoothing parameter; and third, to compare the performance of *P*-splines with that of cubic smoothing splines in a correlated error context.

We considered $\rho = 0.3, 0.55$, $\sigma^2 = 0.01, 0.09$ and $n = 100$ (as in Wang (1998b)), and took *P*-spline parameters $bdeg = 3$, $pord = 2$, and $ndx = 10, 20, 40, 60$ and 80; we also fitted cubic smoothing splines. We used REML for smoothing parameter selection where in (3.2) we take $\tilde{\boldsymbol{X}} \rightarrow \boldsymbol{BG}$ and $\boldsymbol{\Sigma} \rightarrow \sigma^2(\boldsymbol{V} + \lambda^{-1}\boldsymbol{BZZ'B'})$ with $v_{ij} = \rho^{|i-j|}$. There were $m = 100$ replications of the simulation for each parameter set.

We examine the effect of *ndx* on the estimation of (a) the parameters ρ and σ^2, and (b) the true underlying curve. Figure 5 provides boxplots of the estimated sampling distributions of

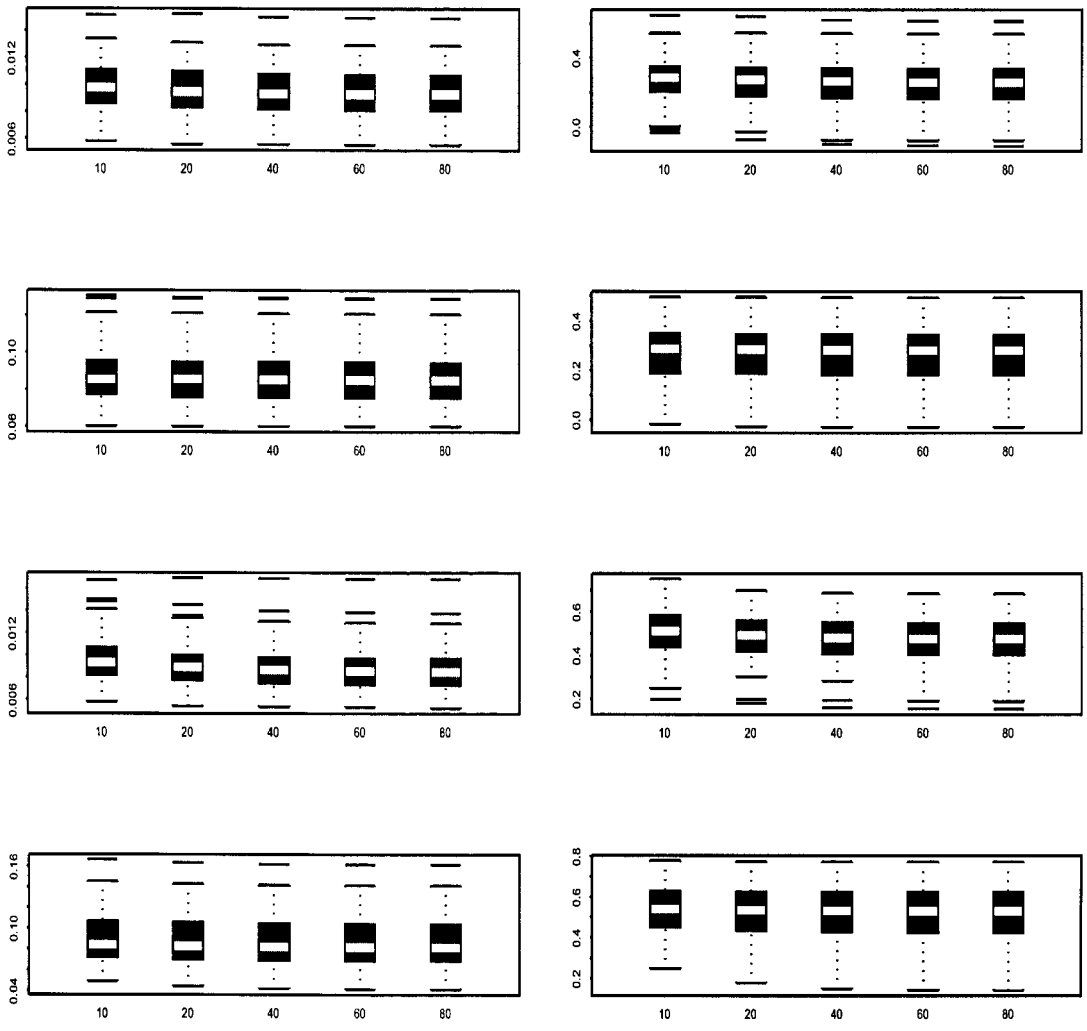

Figure 5 Boxplots of $\hat{\sigma}^2$ (left) and $\hat{\rho}$ (right) for different combinations of parameter values and $ndx = 10$, 20, 40, 60, and 80 (from top to bottom: $\sigma^2 = 0.01$, $\rho = 0.3$; $\sigma^2 = 0.09$; $\rho = 0.3$; $\sigma^2 = 0.01$; $\rho = 0.55$; $\sigma^2 = 0.09$, $\rho = 0.55$)

$\hat{\rho}$ and $\hat{\sigma}^2$ for ndx from 10 to 80. We conclude that there is no significant bias in the estimation of ρ or σ^2 and that ndx has little or no effect on the sampling distribution of $\hat{\rho}$ or $\hat{\sigma}^2$.

We now examine how well the underlying curve is recovered. We calculated the weighted mean square error (WMSE)

$$\text{WMSE} = \frac{1}{n} \| \hat{\boldsymbol{V}}^{-1/2} (\hat{\boldsymbol{f}} - \boldsymbol{f}) \|. \tag{4.8}$$

We found that the average WMSE values are almost independent of the range of values of *ndx* considered here. For example, with $\sigma^2 = 0.09$ and $\rho = 0.55$ we found average WMSE values ($\times 10^{-1}$) of 2.18, 2.17, 2.18, 2.15, and 2.15 for $ndx = 10$, 20, 40, 60, and 80, respectively; the WMSE value for cubic splines was 2.14. Similar results were obtained for the other combinations of parameters. A further check on the recovery of the underlying curve was carried out ($\sigma^2 = 0.09$, $\rho = 0.55$ only) for *P*-splines with $ndx = 40$ and cubic smoothing splines. We selected the 5th, 25th, 50th, 75th, and 95th best estimates of the underlying curve as determined by the ordering of the WMSE values; Figure 6 shows the resulting plot. We found that REML did not lead to any cases of interpolation of the data and was efficient in separating the smooth trend and the autocorrelation in the cases presented above; however, with higher values of σ^2 and ρ there was a tendency to undersmooth the fitted curve. Figure 6 also shows that there is no significant difference between estimates with *P*-splines (40 knots) and cubic smoothing splines (100 knots). In a small number of cases (usually when the smoothing parameter was large) convergence of the REML estimator was slow. We also note that selection of the smoothing parameter by the modified AIC criterion, AIC_C, tended to lead to interpolation of the data and so even this modification of AIC fails to correct the tendency of AIC to undersmooth correlated data; Wang (1998b) noted a similar tendency with GCV.

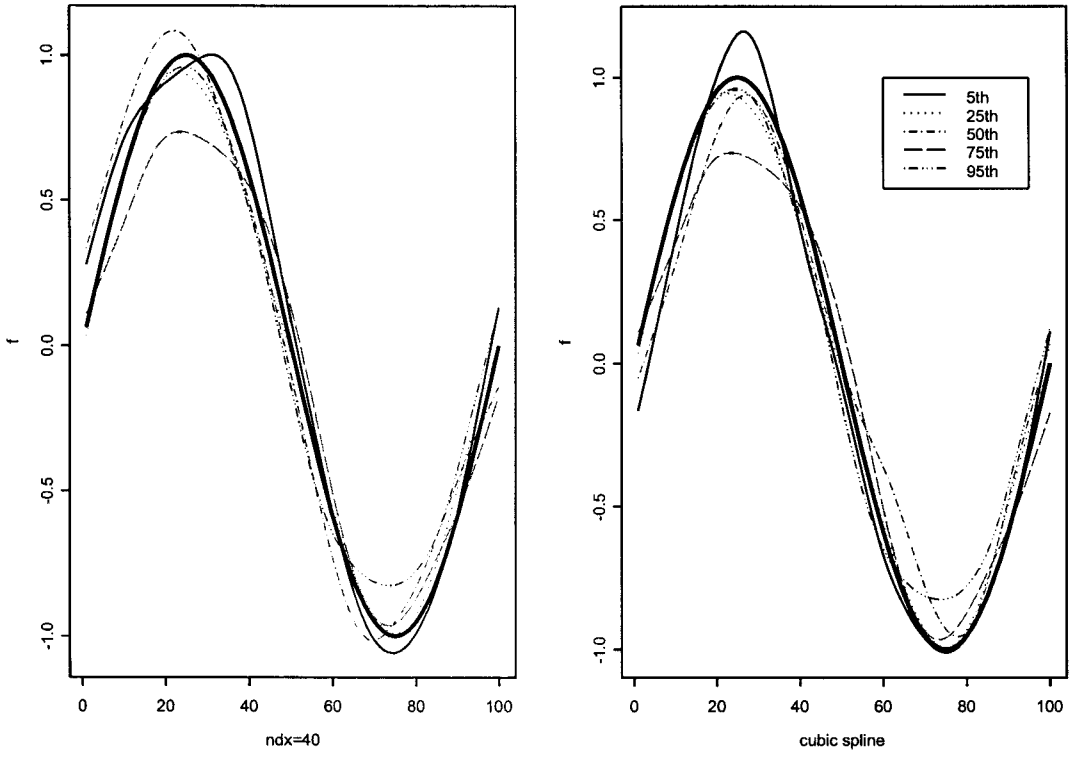

Figure 6 Fitted curves for $ndx = 40$ (left) and cubic smoothing spline (right) for $\sigma^2 = 0.09$, $\rho = 0.55$. Dashed lines correspond to 5th, 25th, 50th, 75th, and 95th best estimates ordered by WMSE; the solid bold line is the true function

In conclusion, in this study we found that using REML to select the smoothing parameter in *P*-splines gave good estimates of the underlying smooth function $f(\cdot)$ and the correlation coefficient ρ. It was not necessary to use as many knots as there are data points (as with cubic smoothing splines), and a maximum of 40 points (as in Ruppert (2002)) was sufficient. The *P*-spline method is as statistically efficient as that proposed by Wang (1998b), but is computationally more attractive.

4.3.2 *Analysis of wood profile data*

The data, described in Pandit and Wu (1993), give 320 measurements of the profile of a block of wood that was subject to grinding. Figure 7 shows the profile height at different sampling distances; the profile variation follows a curve determined by the radius of the grinding stone. Pandit and Wu (1993) assumed that the trend is known *a priori* to be a circle, and used three parameters (two for the centre of the grinding stone and one for the radius) to fit it; they then iterated between estimating the deterministic trend and the error structure. We propose a more unified and general approach. We smooth the data using *P*-splines and fit the covariance structure simultaneously; we use REML to choose the smoothing and correlation parameters, and likelihood ratio tests to select the appropriate model for the error term. Similar analyses have been done previously by Altman (1990) and Wang (1998b). However, we extend these results by allowing for comparison between different forms of the covariance matrix V.

The simulation study in the previous section showed that AIC_C led to undersmoothing with correlated errors and so it could not be used to select *ndx*, *bdeg*, and *pord*. Instead we simply used Ruppert's (2002) rule and took $ndx = 40$, $bdeg = 3$ and $pord = 2$. We follow

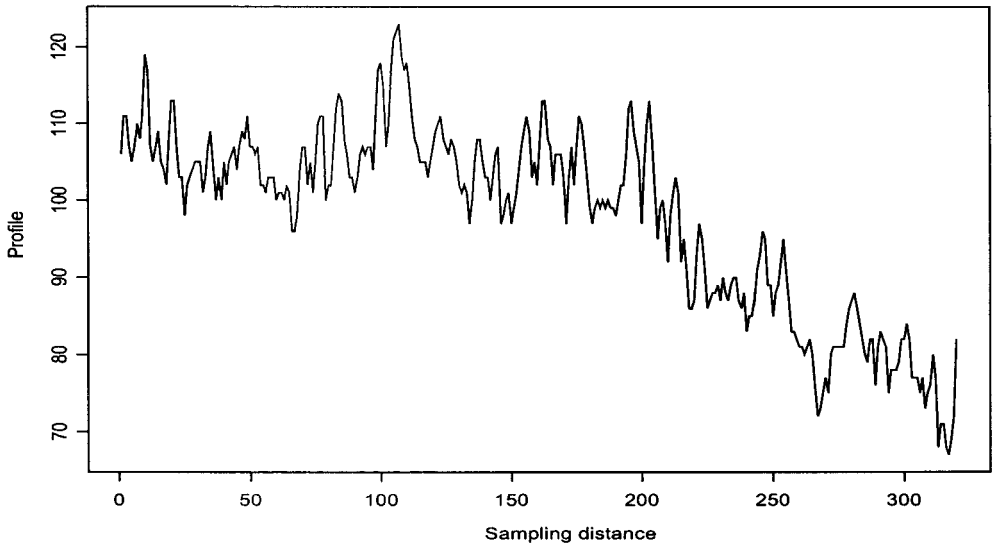

Figure 7 Observations on the profile of a block of wood subject to grinding

Table 2 Parameter and log-likelihood values for models with independent, AR(1), and AR(2) errors

Error variance model	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\rho}_2$	Number of variance parameters	log-likelihood
I	12.95	0.16			2	−632.08
AR(1)	28.40	256.65	0.804		3	−533.75
AR(2)	26.21	159.78	0.975	−0.238	4	−525.15

the method outlined in Section 4.3.1 and give results for three models for errors: independent, AR(1), and AR(2). Table 2 gives the estimated parameters and likelihood ratio tests.

A model with an AR(1) structure led to a significant change in likelihood, $-2\log\Lambda = 196.7$ on 1 df, and increased the value of the smoothing parameter fitting a smoother trend. However, residuals from this model still presented some correlation (see left panel of Figure 8), and so a second model with an AR(2) structure for errors was fitted. Now, the likelihood ratio test of $H_0 : \rho_2 = 0$ *vs* $H_1 : \rho_2 \neq 0$ computed from (3.2) resulted in $-2\log\Lambda = 17.2$ on 1 df, ample evidence that ρ_2 is non-zero. The right panel of Figure 8 shows that the autocorrelation function of the residuals is consistent with white noise, suggesting that an AR(2) model for errors would be appropriate. Further models (ARMA(2,1) and AR(3)) were fitted, but they did not improve significantly on an AR(2) model. Figure 9 shows the fitted smooths with and without the serial correlation; the smooth without the serial correlation shows the expected erratic behaviour. (Our findings in the simulation study were confirmed since the use of AIC_C gave an even more undersmoothed trend.) We also show 95% confidence intervals for the trend; this is computed easily since $\text{var}(\hat{\boldsymbol{y}}) = \sigma^2 \boldsymbol{H}_V \boldsymbol{V} \boldsymbol{H}_V'$ where $\boldsymbol{H}_V$ is defined in (2.5) and $\hat{\sigma}^2 = 26.21$.

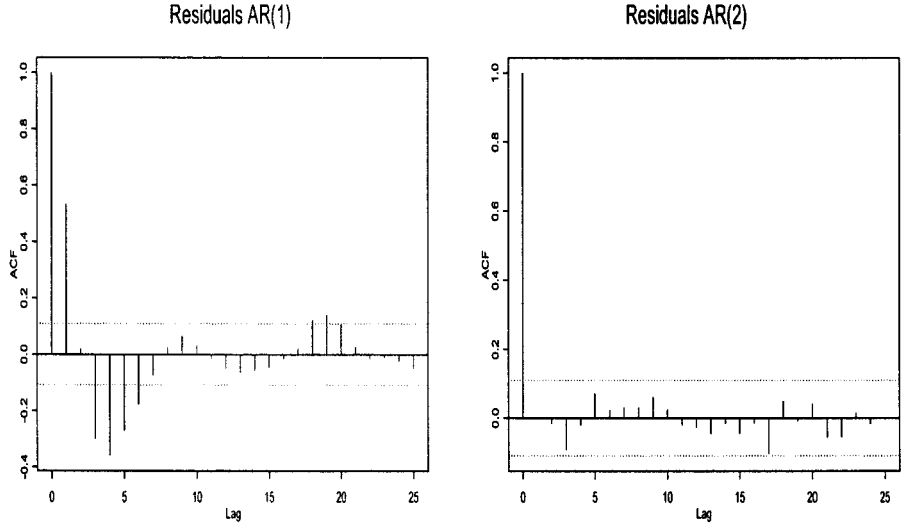

Figure 8 Autocorrelation function estimates and approximate 95% confidence intervals of residuals after fitting a model with trend and AR(1) model for errors (left) or AR(2) (right)

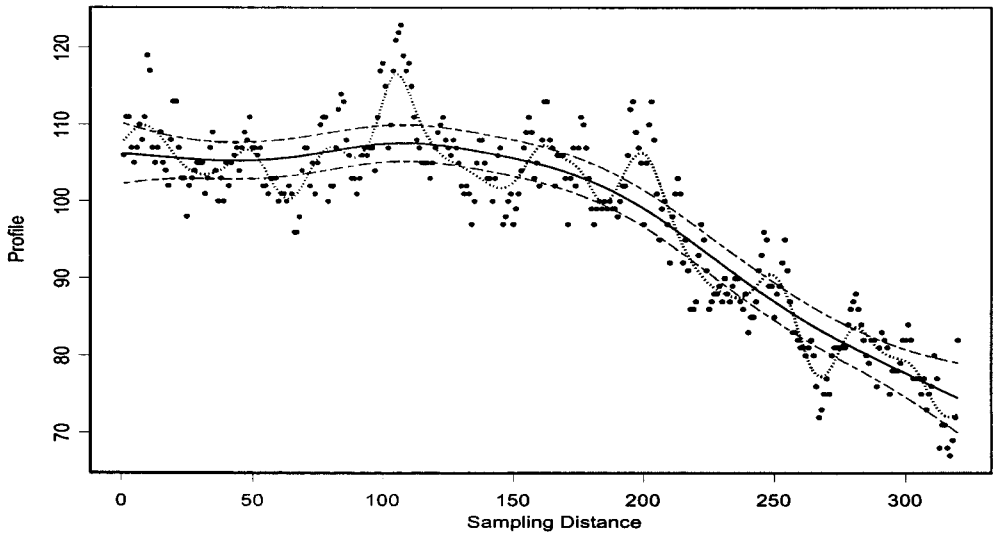

Figure 9 Fitted smooth assuming independence (dashed line) or AR(2) (solid line) model for errors with approximate 95% confidence intervals for trend

This data set is fairly large (320 observations), and so, by using smoothing splines as in Wang (1998b) to fit the trend and REML to estimate the four variance parameters simultaneously, (as in the case of the AR(2) model, for example), we would incur a computational burden that can be improved substantially with *P*-splines.

5 Concluding remarks

In this paper we have used *P*-splines to develop a unified approach to fitting various non-parametric models with a smoothly varying trend. One advantage of *P*-splines over other smoothers (such as cubic smoothing splines) is that *B*-splines with penalties offer significant computational efficiencies and thus allow the fitting of models to larger data sets. They proved to be an effective vehicle of analysis in the examples presented here.

The basis for fitting *P*-splines depends on the number of knots and the degree of the polynomial, and the actual fit depends on the order of the penalty. We have given a systematic approach to the selection of these parameters in the case of independent errors, and have shown in a simulation study that, when errors are correlated, the impact of these parameters on the estimation of both the correlation and the fitted curve is small. Our investigations give support to the simple approach of Ruppert (2002).

P-splines (with a *B*-spline basis) have a mixed model representation and this allows the use of existing theory for parameter estimation and model selection; further, the *B*-spline basis has sound numerical properties, an important feature that is not present in some other bases.

Correlation has important consequences for the selection of the smoothing parameter and for the statistical properties of the smoother. The simulation study showed that REML is a good method of estimation for *P*-spline fitting when errors are correlated (at least, for the parameter values used in our study). REML did not lead to significant bias of the parameter estimates and in no case did the fitted curve undersmooth the data; in contrast, the modified AIC criterion of Hurvich *et al.* (1998) led to curves that in many cases interpolated the data. REML performed as well with smoothing splines as with *P*-splines but low-rank smoothers such as *P*-splines have the benefit of being computationally faster to implement. Further work is needed to see how REML copes with higher values of σ^2 and ρ (a small simulation showed us that for large values of both parameters it will tend to undersmooth the data).

This unified approach to modelling with *P*-splines can readily be extended to cope with more general settings, such as missing values, generalized linear models, or models in more than one dimension. For example, in our current work we model Poisson data in two dimensions: the systematic part of the model is represented as the tensor product of two *B*-spline bases and the analysis can be carried out using the generalized linear mixed model framework (GLMM). The dimension-reducing property of *P*-splines plays a crucial part in the computations.

Acknowledgements

We would like to thank the two referees for their comments, which led to significant improvements to the paper. The work of Maria Durban was supported by DGES project BEC 2001-1270.

References

Aerts M, Claeskens G, Wand MP (2002) Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, **103**, 455–70.

Altman NS (1990) Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749–59.

Arnold, MH, Kempton RA (1979) Estimating the performance of sugar beet varieties. Proceedings 42nd Winter Congress Institut Internacional de Rescherches Bettaravieres, Brussels, 189–203.

Brumback BA, Rice JA (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–94.

Brumback BA, Ruppert D, Wand MP (1999) Comment on Shively, Kohn & Wood. *Journal of the American Statistical Association*, **94**, 794–7.

Coull BA, Ruppert D, Wand MP (2001a) Simple incorporation of interactions into additive models. *Biometrics*, **57**, 539–45.

Coull BA, Schwartz J, Wand MP (2001b) Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, **2**, 337–50.

Durban M, Currie ID (2002) A note on *P*-spline additive models with correlated errors. *Computational Statistics and Data Analysis*, in press.

Durban M, Currie ID, Eilers P (2002) Using *P*-splines to smooth two-dimensional Poisson data. *Proceedings of the 17th International Workshop on Statistical Modelling*, Crete, Greece, 207–14.

Durban M, Currie ID, Kempton RA (2001) Adjusting for fertility and competition in variety trials. *Journal of Agricultural Science, Cambridge*, **136**, 129–40.

Eilers PHC, Marx BD (1996) Flexible smoothing with *B*-splines and penalties. *Statistical Science*, **11**, 89–121.

Hand DJ, Daly F, Lunn AD, McConway KJ, Ostrowski E (1994) *A handbook of small data sets*. London: Chapman & Hall.

Hurvich CM, Simonoff JS, Tsai C-L (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271–93.

Liang K-Y, Self SG (1996) On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society, Series B*, **58**, 785–96.

Lin X, Zhang D (1999) Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381–400.

Pandit SM, Wu S-M (1993) *Time series and system analysis with applications*. Malabar, Florida: Krieger.

Parise H, Wand MP, Ruppert D, Ryan L (2001) Incorporation of historical controls using semiparametric mixed models. *Journal of the Royal Statistical Society, Series C*, **50**, 31–42.

Pinheiro JC, Bates DM (2000) *Mixed-effects models in S and S-plus*. New York: Springer-Verlag, 57–96.

Ruppert D (2002) Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, (in press).

Searle SR, Casella G, McCulloch CE (1992) *Variance components*. New York: John Wiley & Sons.

Silverman BW (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**, 1–52.

Speed T (1991) Comment on Robinson. *Statistical Science*, **6**, 42–44.

Verbyla AP, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Journal of the Royal Statistical Society, Series B*, **48**, 269–311.

Wand MP (1999) On the optimal amount of smoothing in penalised spline regression. *Biometrika*, **86**, 936–40.

Wang Y (1998a) Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–74.

Wang Y (1998b) Smoothing spline models with correlated errors. *Journal of the American Statistical Association*, **93**, 341–48.

Zhang D, Lin X, Raz J, Sowers M (1998) Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**, 710–19.