Biometrika (2005), **92,** 1, *pp.* 91–103

Printed in Great Britain

Exact likelihood ratio tests for penalised splines

By CIPRIAN CRAINICEANU

Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 221205, U.S.A.

ccrainic@jhsph.edu

DAVID RUPPERT

School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853, U.S.A.

davidr@orie.cornell.edu

GERDA CLAESKENS

Departement OR & Business Statistics, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium

claeskens@econ.kuleuven.ac.be

and M. P. WAND

Department of Statistics, School of Mathematics, University of New South Wales, Sydney 2052, Australia

wand@maths.unsw.edu.au

Summary

Penalised-spline-based additive models allow a simple mixed model representation where the variance components control departures from linear models. The smoothing parameter is the ratio of the random-coefficient and error variances and tests for linear regression reduce to tests for zero random-coefficient variances. We propose exact likelihood and restricted likelihood ratio tests for testing polynomial regression versus a general alternative modelled by penalised splines. Their spectral decompositions are used as the basis of fast simulation algorithms. We derive the asymptotic local power properties of the tests under weak conditions. In particular we characterise the local alternatives that are detected with asymptotic probability one. Confidence intervals for the smoothing parameter are obtained by inverting the tests for a fixed smoothing parameter versus a general alternative. We discuss F and R tests and show that ignoring the variability in the smoothing parameter estimator can have a dramatic effect on their null distributions. The powers of several known tests are investigated and a small set of tests with good power properties is identified. The restricted likelihood ratio test is among the best in terms of power.

Some key words: Linear mixed model; Penalised spline; Smoothing; Zero variance component.

1. Introduction

Penalised-spline additive models described in Marx & Eilers (1998), Ruppert & Carroll (2000) and Aerts et al. (2002) have the advantage that they require only a small set of spline basis functions for each covariate and can be represented as mixed models (Brumback et al., 1999). Testing for simplifying assumptions, such as no covariate effect or linear covariate effect, can be reduced to testing for zero variance components. The key idea, to embed the parametric regression function, such as a polynomial, into a larger, nonparametric family, such as penalised splines, has been used by others, such as Cleveland & Devlin (1988), Azzalini & Bowman (1993), Hart (1997), Härdle et al. (1998), Aerts et al. (1999), Ruppert et al. (2003) and Crainiceanu & Ruppert (2004a, b). When the null hypothesis is fully parametric, the null distribution of the test statistics is usually easy to obtain. Our approach is to use restricted likelihood ratio statistics and derive their exact finite-sample null distributions, when the alternative is modelled flexibly by penalised splines. The distributions can be computed within a few seconds and we can avoid asymptotic approximations which, in smoothing, are often of only moderate accuracy.

In the linear mixed models framework the fitted penalised splines are best linear unbiased predictors and the smoothing parameters are ratios of variance components which can be estimated by maximum likelihood or restricted maximum likelihood. Testing for a polynomial regression model is equivalent to testing that a variance component is zero. Likelihood ratio tests for null variance components are nonstandard because the null value of the parameter is on the boundary and data are not independent, at least under the alternative. General finite-sample and asymptotic results for linear mixed models were obtained by Crainiceanu & Ruppert (2004c) for a fixed, but large, number of knots. Under some additional assumptions, asymptotic results were also obtained by Claeskens (2004) when the number of knots increases with the sample size. Confidence intervals for the smoothing parameter are obtained by inverting the tests for a fixed smoothing parameter versus a general alternative. These intervals can be viewed as confidence intervals for the number of degrees of freedom of the nonparametric fit.

2. Framework

Consider the partially linear model $y_i = W_i^{\mathrm{T}}\gamma + m(x_i) + \varepsilon_i$, where the ε_i are independent identically distributed $N(0, \sigma_\varepsilon^2)$, ε_i is independent of W_i and x_i, W_i is a vector of covariates that enter the model linearly, x_i is another covariate, and $m(.)$ is a smooth function. We consider the x's to be random but their joint distribution is unspecified and could be degenerate; for example the x's might be equally spaced. We want to test if $m(.)$ is a $(p-q)$th degree polynomial, $m(x, \theta) = \beta_0 + \beta_1 x + \ldots + \beta_{p-q}x^{p-q}$, where $0 \leqslant q \leqslant p$ and $\theta = (\beta_0, \ldots, \beta_{p-q})^{\mathrm{T}}$. To define an alternative that is both flexible enough to describe a large class of functions and is suitable for testing, we consider the class of spline functions

$$m(x, \Theta) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^p,$$

where $\Theta = (\theta^{\mathrm{T}}, \beta_{p-q+1}, \ldots, \beta_p, b_1, \ldots, b_K)^{\mathrm{T}}$ is the vector of regression coefficients, and $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ are fixed knots. Following Ruppert (2002), we consider a number of

knots that is large enough, typically 5 to 20, to ensure the desired flexibility, and κ_k is the sample quantile of x's corresponding to probability $k/(K+1)$, but our results hold for any other choice of knots. To avoid overfitting, we minimise

$$\sum_{i=1}^{n} \{y_i - W_i^{\mathrm{T}}\gamma - m(x_i, \Theta)\}^2 + \frac{1}{\lambda}\Theta^{\mathrm{T}} D\Theta, \tag{1}$$

where λ is the smoothing parameter and D is a known positive semidefinite penalty matrix. The penalty $\int \{m^{(2)}(x, \Theta)\}^2 dx$ used for smoothing splines can be achieved with D equal to the sample second moment matrix of the second derivatives of the spline basis functions. However, in this paper we focus on matrices D of the form

$$D = \begin{bmatrix} 0_{(p+1)\times(p+1)} & 0_{(p+1)\times K} \\ 0_{K\times(p+1)} & \Sigma^{-1} \end{bmatrix},$$

where Σ is a positive definite matrix and $0_{m\times l}$ is an $m \times l$ matrix of zeros. This matrix D penalises the coefficients of the spline basis functions $(x-\kappa_k)_+^p$ only and will be used in the remainder of the paper. A standard choice is $\Sigma = I_K$.

Let $Y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, W be the matrix with the ith row equal to W_i^{T}, X be the matrix with the ith row $X_i = (1, x_i, \ldots, x_i^p)$, and Z be the matrix with ith row $Z_i = \{(x_i-\kappa_1)_+^p, \ldots, (x_i-\kappa_K)_+^p\}$.

3. Penalised splines as linear mixed models

If we divide (1) by the error variance we obtain

$$\frac{1}{\sigma_\varepsilon^2}\|Y - W\gamma - X\beta - Zb\|^2 + \frac{1}{\lambda\sigma_\varepsilon^2} b^{\mathrm{T}}\Sigma^{-1} b,$$

where $\beta = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ and $b = (b_1, \ldots, b_K)^{\mathrm{T}}$. Define $\sigma_b^2 = \lambda\sigma_\varepsilon^2$, and consider the vectors γ and β as fixed parameters and the vector b as a set of random parameters with $E(b) = 0$ and $\mathrm{cov}(b) = \sigma_b^2\Sigma$. If $(b^{\mathrm{T}}, \varepsilon^{\mathrm{T}})^{\mathrm{T}}$ is a normal random vector and b and ε are independent then one obtains an equivalent model representation of the penalised spline in the form of a linear mixed model (Brumback et al., 1999):

$$Y = W\gamma + X\beta + Zb + \varepsilon, \quad \mathrm{cov}\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{bmatrix} \sigma_b^2\Sigma & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{bmatrix}. \tag{2}$$

For this model $E(Y) = W\gamma + X\beta$ and $\mathrm{cov}(Y) = \sigma_\varepsilon^2 V_\lambda$, where $V_\lambda = I_n + \lambda Z\Sigma Z^{\mathrm{T}}$ and n is the total number of observations. In model (2) both W and X correspond to fixed effects. For simplicity we can redefine $X = [W|X]$, $\beta = (\gamma^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ and let $p+1$ be the dimension of the new vector β. With this new notation, twice the loglikelihood function of the observations Y, given model (2) with parameters β, σ_ε^2 and λ, is

$$\mathrm{L}(\beta, \sigma_\varepsilon^2, \lambda) = -\left[n\log(\sigma_\varepsilon^2) + \log\{\det(V_\lambda)\} + \frac{(Y-X\beta)^{\mathrm{T}} V_\lambda^{-1}(Y-X\beta)}{\sigma_\varepsilon^2} \right].$$

A second parameter estimation criterion for model (2) is the restricted loglikelihood

$$\text{REL}(\beta, \sigma_\varepsilon^2, \lambda) = \text{L}(\beta, \sigma_\varepsilon^2, \lambda) - (p+1)\log(\sigma_\varepsilon^2) - \log\{\det(X^{\mathrm{T}} V_\lambda^{-1} X)\}. \tag{3}$$

The joint maximisation of these criteria over $(\beta, \sigma_\varepsilon^2, \lambda)$ provides the maximum likelihood and restricted maximum likelihood estimators respectively.

Given the representation (2) of the penalised spline model, we test the null hypothesis

$$H_0: \beta_{p-q+1} = \ldots = \beta_p = 0, \quad \sigma_b^2 = 0 \quad (\lambda = 0) \tag{4}$$

against the alternative hypothesis H_A: $\beta_{p-q+1} \neq 0$ or $\ldots$ or $\beta_p \neq 0$ or $\sigma_b^2 > 0$, equivalently $\lambda > 0$. Since the coefficients b have mean zero and covariance matrix $\sigma_b^2 \Sigma$, the hypothesis that $\sigma_b^2 = 0$ in H_0 is equivalent to all spline coefficients b_i being zero. The condition that $\beta_{p-q+1} = \ldots = \beta_p = 0$ in H_0 is equivalent to the assumption that the true function is a polynomial of degree $p-q$. If $q = 0$ then we have the important particular case

$$H_0: \sigma_b^2 = 0 \quad (\text{equivalently, } \lambda = 0) \quad \text{versus} \quad H_A: \sigma_b^2 > 0 \quad (\text{equivalently, } \lambda > 0). \tag{5}$$

We have been assuming that γ is unconstrained under the null hypothesis, but it is trivial to impose constraints on γ; doing this merely increases the number of fixed effects specified by the null hypothesis.

A generalisation of (5) is

$$H_0: \lambda = \lambda_0 \quad \text{versus} \quad H_A: \lambda \in \Lambda \subset [0, \infty), \tag{6}$$

where Λ can be any subset of $[0, \infty)\backslash\{\lambda_0\}$. We discuss the following cases for Λ: $\{\lambda_1\}$ for a fixed $\lambda_1 \neq \lambda_0$, (λ_0, ∞) and $[0, \infty)\backslash\{\lambda_0\}$. In § 7 we show that testing for a fixed smoothing parameter, or equivalently for a fixed number of degrees of freedom, of a penalised spline regression is equivalent to testing (6). Tests of this type can be inverted to create confidence intervals for λ.

4. Likelihood ratio tests for polynomial regression

In this section we introduce likelihood ratio tests for the null hypothesis in (4). Define the likelihood ratio test statistic as

$$\text{LRT}_n = \sup_{H_A \cup H_0} \text{L}(\beta, \sigma_\varepsilon^2, \lambda) - \sup_{H_0} \text{L}(\beta, \sigma_\varepsilon^2, \lambda).$$

A similar test can be introduced based on the restricted maximum likelihood criterion (3). Since restricted maximum likelihood uses the likelihood of residuals after fitting the fixed effects, it is appropriate for testing only if the fixed effects are the same under the null and the alternative hypotheses. This should not be a problem since we can always choose $q = 0$, and

$$\text{RLRT}_n = \sup_{H_A \cup H_0} \text{REL}(\beta, \sigma_\varepsilon^2, \lambda) - \sup_{H_0} \text{REL}(\beta, \sigma_\varepsilon^2, \lambda).$$

Computing LRT_n or RLRT_n is simple using standard software, such as R and S-Plus using the `lme` function or SAS using the `MIXED` procedure.

5. Null distributions of the test statistics

5·1. *Preliminary considerations*

Suppose, for example, that we want to test for a constant mean against a general alternative of a piecewise-constant spline with K knots. This is equivalent to setting $q=0$ and $p=0$ and testing $H_0\colon \sigma_b^2=0$ $(\lambda=0)$ versus $H_A\colon \sigma_b^2>0$ $(\lambda>0)$. Since the parameter under the null is on the boundary, the classical result that $\text{LRT}_n \to \chi_1^2$ in distribution under H_0 does not hold. One may be tempted to believe that the asymptotic theory of Self & Liang (1987, 1995) and Stram & Lee (1994) for boundary problems applies. These results suggest that, in distribution, under H_0,

$$\text{LRT}_n \to 0{\cdot}5\chi_0^2 + 0{\cdot}5\chi_1^2, \tag{7}$$

where χ_0^2 denotes a point mass at zero. However, (7) is derived under the assumption that the response variable vector can be partitioned into J independent identically distributed subvectors with $J \to \infty$.

C. M. Crainiceanu, D. Ruppert and T. J. Vogelsang, in as yet unpublished work, show that the asymptotic probability masses at zero for LRT_n and RLRT_n are practically constant over a wide range of number of knots K, and are approximately 0·95 for LRT_n and 0·65 for RLRT_n. These values are much larger than 0·5 but provide excellent approximations of the finite sample probabilities.

Crainiceanu & Ruppert (2004c) found that the finite-sample distributions of LRT_n and RLRT_n converge very quickly to an asymptotic distribution different from $0{\cdot}5\chi_0^2 + 0{\cdot}5\chi_1^2$, the latter providing a very conservative approximation of the null finite sample and asymptotic distributions. Although the asymptotics in Crainiceanu & Ruppert (2004c) could be used, they are unnecessary since it is easy to compute exact critical values.

5·2. *Finite sample null distribution of the test statistics*

The finite sample and asymptotic distributions of LRT_n and RLRT_n for testing null hypotheses including zero random-effects variance in linear mixed models with one variance component were derived by Crainiceanu & Ruppert (2004c). These results are applicable in the penalised splines framework because of the linear mixed-model representation (2).

Let $\mu_{s,n}$ and $\xi_{s,n}$ be the K eigenvalues of the $K \times K$ matrices $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} P_0 Z \Sigma^{\frac{1}{2}}$ and $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} Z \Sigma^{\frac{1}{2}}$ respectively, where $P_0 = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Then, in distribution,

$$\text{LRT}_n = n\left(1 + \frac{\sum_{s=1}^{q} u_s^2}{\sum_{s=1}^{n-p-1} w_s^2}\right) + \sup_{\lambda \geqslant 0}\left[n \log\left\{1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right\} - \sum_{s=1}^{K} \log(1+\lambda\xi_{s,n})\right], \tag{8}$$

where u_s, for $s=1,\ldots,K$, and w_s, for $s=1,\ldots,n-p-1$, are independently $N(0,1)$,

$$N_n(\lambda) = \sum_{s=1}^{K} \frac{\lambda\mu_{s,n}}{1+\lambda\mu_{s,n}} w_s^2, \quad D_n(\lambda) = \sum_{s=1}^{K} \frac{w_s^2}{1+\lambda\mu_{s,n}} + \sum_{s=K+1}^{n-p-1} w_s^2.$$

If $q=0$ and we test $\sigma_b^2=0$, then, in distribution,

$$\text{RLRT}_n = \sup_{\lambda \geqslant 0}\left[(n-p-1)\log\left\{1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right\} - \sum_{s=1}^{K} \log(1+\lambda\mu_{s,n})\right]. \tag{9}$$

Although they look complicated, these distributions are easily simulated. Crainiceanu & Ruppert (2004c) provide a fast simulation algorithm that takes advantage of the spectral representation of LRT_n and RLRT_n.

6. Local power properties

We now focus on the local asymptotic power properties of LRT_n and RLRT_n for polynomial regression. We state the theorem for the more general case of linear mixed models with one random-effects variance component.

Theorem 1. *Consider testing hypotheses described in equation* (5) *for a linear mixed model with one variance component as described in equation* (2). *Assume that there exists* $a > 0$ *such that, for every* s, *the* K *eigenvalues* $\mu_{s,n}$ *of the matrix* $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} P_0 Z \Sigma^{\frac{1}{2}}$ *satisfy* $\lim_{n\to\infty}(n^{-a}\mu_{s,n}) = \mu_s$, *where not all* μ_s *are zero.*

(i) *If the true value of the variance ratio* $\sigma_b^2/\sigma_\varepsilon^2$ *is* $\lambda_{n,0} = n^{-a} d_0$ *then for every* x

$$\lim_{n\to\infty} \text{pr}(\text{RLRT}_n > x) = \text{pr}(X_{d_0} > x),$$

where, in distribution,

$$X_{d_0} = \sup_{d \geqslant 0} \left\{ \sum_{s=1}^{K} \frac{(1 + d_0\mu_s) d\mu_s}{1 + d\mu_s} w_s^2 - \sum_{s=1}^{K} \log(1 + d\mu_s) \right\},$$

and w_s *are independent* $N(0, 1)$ *random variables.*

(ii) *If the true value of the variance ratio* $\sigma_b^2/\sigma_\varepsilon^2$ *is* $\lambda_{n,0} = n^{-b} d_0$, *with* $0 \leqslant b < a$, *then, for every* $x > 0$,

$$\lim_{n\to\infty} \text{pr}(\text{RLRT}_n > x) = 1.$$

The complete proof is contained in a technical report available from the authors. These results generalise the null asymptotic results obtained by Crainiceanu & Ruppert (2004c, Theorem 2) for RLRT_n when $d_0 = 0$. Similar results hold true for LRT_n.

When applied to the particular case of penalised splines, the first result of the theorem provides the asymptotic power of RLRT_n to detect alternatives when the true smoothing parameter converges to zero at rate n^a. By setting x equal to the $1 - \alpha$ quantile of the null asymptotic distribution corresponding to $d_0 = 0$ and varying d_0, we obtain the asymptotic power curve, which could be used for power comparisons. When $a = 1$ the condition on the asymptotic behaviour of the eigenvalues of the $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} P_0 Z \Sigma^{\frac{1}{2}}$ matrix is the standard condition used to ensure the asymptotic consistency of estimators of parameters. Note that $a = 1$ for designs that are generated from a random sample.

If the true parameter converges to zero more slowly than n^{-a} then the RLRT_n rejection regions have asymptotic probability 1.

7. Testing for a fixed smoothing parameter

We now test the hypothesis (6) about $\lambda = \sigma_b^2/\sigma_\varepsilon^2$. Suppose that λ_0 is the true value of λ. Crainiceanu & Ruppert (2004c) showed that the RLRT_n statistic for testing (6) has null spectral decomposition such that

$$\text{RLRT}_n = \sup_{\lambda \in \Lambda} \left[(n - p - 1) \log\left\{ 1 + \frac{N_n(\lambda, \lambda_0)}{D_n(\lambda, \lambda_0)} \right\} - \sum_{s=1}^{K} \log\left(\frac{1 + \lambda\mu_{s,n}}{1 + \lambda_0\mu_{s,n}} \right) \right], \qquad (10)$$

in distribution, where

$$N_n(\lambda, \lambda_0) = \sum_{s=1}^{K} \frac{(\lambda - \lambda_0)\mu_{s,n}}{1 + \lambda\mu_{s,n}} w_s^2, \quad D_n(\lambda, \lambda_0) = \sum_{s=1}^{K} \frac{1 + \lambda_0\mu_{s,n}}{1 + \lambda\mu_{s,n}} w_s^2 + \sum_{s=K+1}^{n-p-1} w_s^2,$$

in which w_s, for $s = 1, \ldots, n-p-1$, are independent $N(0, 1)$, and $\mu_{s,n}$ are the K eigenvalues of the $K \times K$ matrix $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} P_0 Z \Sigma^{\frac{1}{2}}$. A similar result holds for LRT_n.

The null finite sample distributions of RLRT_n depend only on λ_0, $\mu_{s,n}$ and Λ. One consequence is that it is invariant only to reparameterisations that leave $\Sigma^{\frac{1}{2}} Z^{\mathrm{T}} P_0 Z \Sigma^{\frac{1}{2}}$ invariant. This shows that changing the basis in the linear space generated by $1, x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_K)_+^p$ and leaving the penalty matrix unchanged would affect the null distributions of RLRT_n as well as those of the other parameter estimators. In particular, a change of basis to make $Z^{\mathrm{T}} X = 0$, as assumed by Cantoni & Hastie (2002), would change the distribution. For penalised likelihood models it is the combination of the design matrix $Z^{\mathrm{T}} P_0 Z$ and the penalty matrix Σ that determines the distributions. The distribution of RLRT_n can be simulated using an algorithm similar to the one for the case $\lambda_0 = 0$.

Several relevant results can be obtained as a by-product of equation (10). For example, when $\lambda_0 = 0$ and $\Lambda = (0, \infty)$ we obtain the result in (9). If $\Lambda = \{\lambda_1\}$ we obtain the finite sample distribution of RLRT_n for testing the number of degrees of freedom for penalised splines, as described in Cantoni & Hastie (2002), but without their assumption that $Z^{\mathrm{T}} X = 0$. For $\Lambda = (\lambda_0, \infty)$ we obtain the distribution of RLRT_n for testing

$$H_0: \lambda = \lambda_0 \quad \text{versus} \quad H_A: \lambda > \lambda_0. \tag{11}$$

For $\Lambda = [0, \infty) \backslash \{\lambda_0\}$ we obtain the distribution of RLRT_n for testing

$$H_0: \lambda = \lambda_0 \quad \text{versus} \quad H_A: \lambda \neq \lambda_0. \tag{12}$$

8. Confidence intervals for the smoothing parameter

Our ability to simulate quickly the null finite sample distribution of RLRT_n allows us to obtain confidence intervals for the smoothing parameter by inverting the RLRT_n. Indeed, we define the $(1 - \alpha)$-level confidence interval for λ as

$$\mathrm{CI}_\alpha = \{\lambda_0 | p(\lambda_0) \geqslant \alpha\}, \tag{13}$$

where $p(\lambda_0)$ is the p-value for the RLRT_n statistic for testing (12). Since $p(\lambda_0)$ can be obtained in seconds, CI_α can be obtained by computing $p(\lambda_0)$ on a relatively fine grid.

To illustrate this methodology, consider the logarithm of Janka hardness of a sample of Australian timbers versus the density of the timber (Williams, 1959, p. 43). The data are displayed in Fig. 1(a). Linear penalised splines with $K = 15$ knots were used. We obtained $\hat{\lambda}_{\mathrm{REML}} = 0{\cdot}0056$ corresponding to 4·13 degrees of freedom of regression. We used 100 000 simulations for the null distribution of RLRT_n to calculate $p(\lambda_0)$ for each λ_0 on a grid. The grid described earlier was fine enough to determine the regions where the cut-off points are situated but was not fine enough to pinpoint them. Therefore, in a second stage we calculated $p(\lambda_0)$ on very fine grids around the cut-off points. We obtained $\mathrm{CI}_\alpha = [0{\cdot}0014, 0{\cdot}0870]$, which corresponds to a confidence interval $[3{\cdot}32, 6{\cdot}83]$ for the degrees of freedom. Figure 1(b) shows the plot of p-values versus $\log_{10}(\lambda)$. Figure 1(a) shows the curves corresponding to the lower and upper bounds of CI_α.

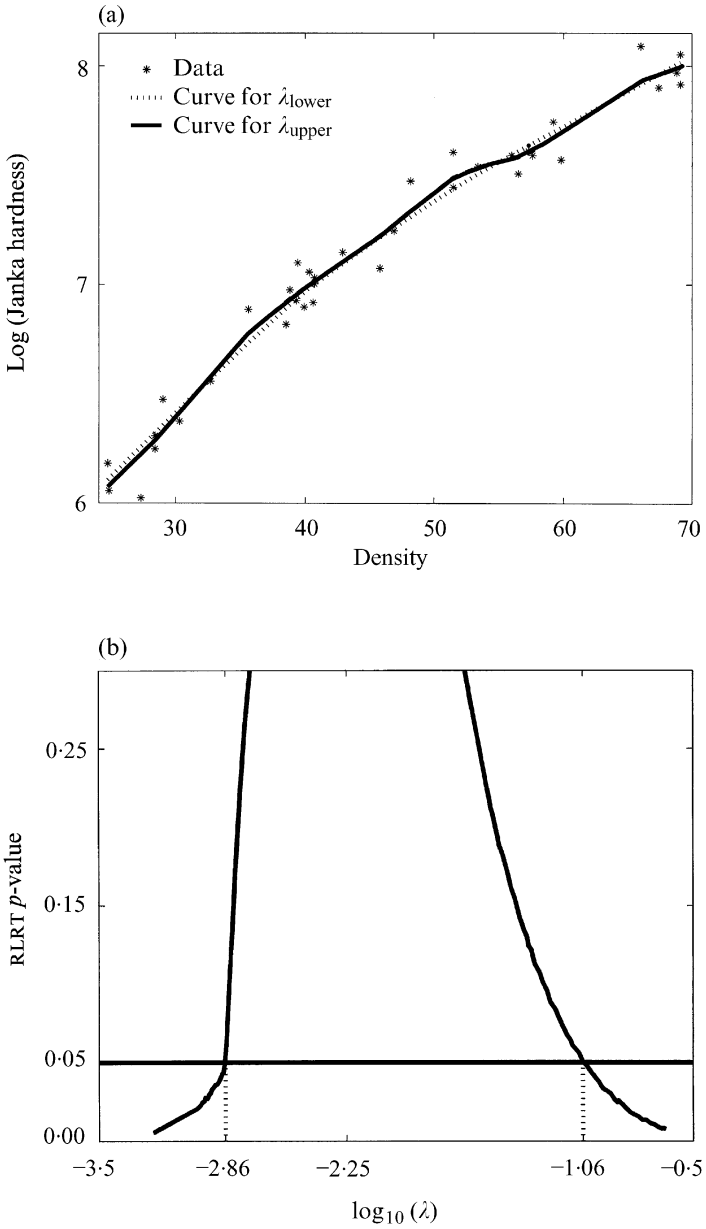

Fig. 1. (a) Logarithm of Janka hardness versus density for a sample of Australian timbers. Fitted curves using linear splines with $K = 15$ knots corresponding to the lower, dotted, and upper, dashed, bounds of the 95% confidence interval for λ. (b) Part of the plot of the p-value versus $\log_{10}(\lambda)$ for the statistic RLRT.

9. THE F AND R TESTS

9·1. *Description of the tests*

Suppose the estimated smooth is $\hat{Y}(\lambda) = S_\lambda Y$, where $S_\lambda = (\mathscr{X}^{\mathrm{T}}\mathscr{X} + \lambda^{-1}D)^{-1}\mathscr{X}^{\mathrm{T}}Y$ is the smoother matrix and $\mathscr{X} = [X|Z]$. The estimated residual vector is $\hat{\varepsilon}_\lambda = (I_n - S_\lambda)Y$. By analogy with linear regression, the degrees of freedom for residuals is $\gamma_\lambda = \mathrm{tr}\,\{(I_n - S_\lambda)^2\}$, where $\lambda \to \gamma_\lambda$ is a one-to-one function (Ruppert et al., 2003, pp. 97–100). The F statistic

for testing the simple null against the simple alternative,

$$H_0: \lambda = \lambda_0 \quad \text{versus} \quad H_A: \lambda = \lambda_1, \tag{14}$$

where $\lambda_0 < \lambda_1$, is defined as

$$F_{\lambda_0,\lambda_1} = \frac{(\text{RSS}_0 - \text{RSS}_1)/(\gamma_0 - \gamma_1)}{\text{RSS}_1/\gamma_1}, \tag{15}$$

where RSS_a is the residual sum of squares for the model corresponding to λ_a and we denote γ_{λ_0} by γ_0 and γ_{λ_1} by γ_1 (Hastie & Tibshirani, 1990, pp. 66–7). Cantoni & Hastie (2002) proposed the following related statistic for testing (14):

$$R_{\lambda_0,\lambda_1} = \frac{Y^{\mathrm{T}}(S_{\lambda_1} - S_{\lambda_0})Y}{Y^{\mathrm{T}}(I_n - S_{\lambda_1})Y}. \tag{16}$$

The finite-sample distributions of the F and R statistics have either been approximated (Hastie & Tibshirani, 1990) or derived (Cantoni & Hastie, 2002). Cantoni & Hastie (2002) suggest using these statistics for testing $H_0: \lambda = \lambda_0$ versus $H_A: \lambda > \lambda_0$, by replacing λ_1 in the F or R test with an estimator $\hat{\lambda}$. They suggested using $\hat{\lambda}_{\text{REML}}(\lambda_0)$, which is the estimated smoothing parameter using restricted maximum likelihood restricted to $[\lambda_0, \infty)$, but other criteria can be used as well. For simplicity we denote $\hat{\lambda}_{\text{REML}}(\lambda_0)$ by $\hat{\lambda}$.

While defining the tests by replacing λ_1 by $\hat{\lambda}$ is straightforward, deriving the null distributions of the test statistics is not. Estimating λ and ignoring the estimation variability when computing the finite sample distributions of $F_{\lambda_0,\hat{\lambda}}$ and $R_{\lambda_0,\hat{\lambda}}$, as suggested by Cantoni & Hastie (2002), inaccurately approximates null distributions.

The probability that $\hat{\lambda} = \lambda_0$ is equal to $\text{pr}(\hat{\lambda}_{\text{REML}} \leqslant \lambda_0)$, where $\hat{\lambda}_{\text{REML}}$ is the unconstrained restricted maximum likelihood estimator of λ. Crainiceanu & Ruppert (2004c) showed that this probability is greater than 0·5 and is equal to the null probability mass at λ_0 of RLRT_n and $R_{\lambda_0,\hat{\lambda}}$. Furthermore, for any fixed $\lambda_1 > \lambda_0$, R_{λ_0,λ_1} has no mass at λ_0. A similar reasoning shows that the $F_{\lambda_0,\hat{\lambda}}$ statistic has the same probability mass at $\lim_{\lambda \downarrow \lambda_0} F_{\lambda_0,\lambda} = c_F(\lambda_0) > 0$. If we consider testing $H_0: \lambda = \lambda_0$ versus $H_A: \lambda \neq \lambda_0$, the LRT_n, RLRT_n and $R_{\lambda_0,\hat{\lambda}}$ statistics will have probability mass at zero (Crainiceanu & Ruppert, 2004c), showing that the χ^2_1 approximation cannot be used even when $\lambda_0 > 0$.

9·2. *Estimated smoothing parameter*

To describe the F and R statistics denote by $\text{RSS}(q, 0)$ the residual sum of squares for the $(p-q)$th-degree polynomial fit corresponding to $\lambda = 0$, and by $\text{RSS}(p, \hat{\lambda})$ the residual sum of squares for the pth-degree penalised spline regression. The F statistic is defined as

$$F^{q,p}(\hat{\lambda}) = \frac{\text{RSS}(q, 0) - \text{RSS}(p, \hat{\lambda})}{\text{RSS}(p, \hat{\lambda})} \frac{\gamma_{\hat{\lambda}}}{n - p + q - 1 - \gamma_{\hat{\lambda}}},$$

and is used by Hastie & Tibshirani (1990, Ch. 3). Many of the test statistics used by other authors are quite similar to $F^{q,p}(\hat{\lambda})$. The quantity $\text{RSS}(q, 0) - \text{RSS}(p, \hat{\lambda})$ is, of course, the difference between the residual sums of squares under the null and alternative hypotheses. For nested linear models, this difference is $\|(I - S_0)Y\|^2 - \|(I - S_A)Y\|^2$, but also equals $\|(S_A - S_0)Y\|^2$ and $\|S_A(I - S_0)Y\|^2$, where S_0 and S_A are the hat matrices for the null and alternative models. For nonparametric models, these three quantities will differ somewhat but should be similar. The tests of Härdle et al. (1998) for generalised regression models are based on the quasilikelihood estimates of Severini & Staniswalis (1994). When

specialised to Gaussian responses, these tests are F statistics using $\|(S_A - S_0)Y\|^2$ in place of $\|(I - S_0)Y\|^2 - \|(I - S_A)Y\|^2$. The test statistic of Azzalini & Bowman (1993) uses $\|S_A(I - S_0)Y\|^2$, which is the sum of squared fitted values when the residuals from the null model are smoothed. Azzalini & Bowman (1993) use this statistic to overcome a peculiarity of kernel regression, which is in general biased even if the true regression is linear. Kernel regression is being displaced by local polynomial regression and penalised splines, which do not have this undesirable property. Therefore, tests based on $\|S_A(I - S_0)Y\|^2$ are no longer needed, though they can of course still be used. Raz (1990) used a similar test when testing for no effect, which is $q = p$ and $\sigma_b^2 = 0$ in our framework. Eubank & Spiegelman (1990) also use $\|S_A(I - S_0)Y\|^2$ since their test statistic is the sum of squared fitted values when a smoothing spline is fitted to residuals from a linear polynomial regression.

The R statistic is defined as

$$R^{q,p}(\hat{\lambda}) = \frac{Y^{\mathrm{T}}(S_{p,\hat{\lambda}} - S_{q,0})Y}{Y^{\mathrm{T}}(I_n - S_{p,\hat{\lambda}})Y},$$

where $S_{p,\hat{\lambda}}$ is the smoother matrix corresponding to the degree-p penalised spline with smoothing parameter $\hat{\lambda}$, and $S_{q,0}$ is the smoother matrix corresponding to the degree-$(p-q)$ polynomial regression. Here $\hat{\lambda}$ denotes a data-dependent smoothing parameter based on one of the available criteria. For reasons of convenience we consider only the maximum likelihood, restricted maximum likelihood and generalised crossvalidation criteria with the restricted maximum likelihood being used only when $q = 0$. The null distributions of $F^{q,p}(\hat{\lambda})$ and $R^{q,p}(\hat{\lambda})$ are not known in this context and they need to be bootstrapped.

9·3. *Fixed smoothing parameter*

To avoid bootstrapping the null distribution of $F^{q,p}(\hat{\lambda})$ and $R^{q,p}(\hat{\lambda})$ one can use a fixed, not estimated, smoothing parameter under the alternative. For a given λ, the mean response is a pth-degree spline function with $\mathrm{tr}(S_\lambda)$ degrees of freedom (Ruppert et al., 2003, p. 81), where $\mathrm{tr}(S_\lambda)$ is an increasing, continuous function of λ from $p+1$ for $\lambda = 0$ to $p+1+K$ for $\lambda = \infty$. Ruppert et al. use λ^p where we use λ^{-1}. We can choose λ_1 to match any given number of degrees of freedom in the interval $[\,p+1, p+1+K\,]$. For example, for λ_1 corresponding to $p+2$ degrees of freedom of fit we define

$$F^{q,p}(1) = \frac{\mathrm{RSS}(q,0) - \mathrm{RSS}(p,\lambda_1)}{\mathrm{RSS}(p,\lambda_1)} \frac{\gamma_{\lambda_1}}{n - p + q - 1 - \gamma_{\lambda_1}}, \quad R^{q,p}(1) = \frac{Y^{\mathrm{T}}(S_{p,\lambda_1} - S_{q,0})Y}{Y^{\mathrm{T}}(I_n - S_{p,\lambda_1})Y},$$

with the usual notation. The '1' in the $F^{q,p}(1)$ and $R^{q,p}(1)$ notation stands for 'one degree of freedom more than the $p+1$ degrees of freedom of the pth-degree polynomial'. Similarly we can define $F^{q,p}(2)$, $R^{q,p}(2)$ and so on. A potential problem could be that, by fixing a priori the number of degrees of freedom under the alternative, one may lose power, especially when the fixed number is far from the 'true' number of degrees of freedom. We investigate this further in § 10.

We also used the von Neumann type test for unspecified alternatives described in Hart (1997, pp. 132–4). This test does not require the specification of an alternative hypothesis. Another class of possible tests are those of Durbin–Watson type (Munson & Jernigan, 1989), but these proved disappointing in a power study by Azzalini & Bowman (1993).

10. Power simulations

In §§ 4 and 9 several statistics were described for testing $(p-q)$th-degree polynomial regression. In this section we compare the power of these tests under different alternatives when the null hypothesis is a constant mean, $p-q=0$, or a linear polynomial, $p-q=1$. When testing for constant mean, the alternative hypothesis is modelled using either a piecewise constant, $p=0$, or a linear spline, $p=1$. When testing for a linear polynomial, the alternative hypothesis is modelled using either a linear, $p=1$, or a quadratic, $p=2$, spline.

One is interested in tests that perform well under various alternatives. We use three families of functions hoping that they are a cross-section of regression functions found in practice. For all these functions, a scalar d controls the degree of departure from the null hypothesis, with $d=0$ corresponding to H_0.

Case 1: *An increasing function.* Here we take

$$m_d(x)=\frac{d}{1+e^{10(0{\cdot}5-x)}} \quad (q=0), \quad m_d(x)=x+\frac{d}{1+e^{10(0{\cdot}5-x)}} \quad (q=1).$$

Case 2: *A concave function.* Here we take

$$m_d(x)=-d|0{\cdot}5-x|^{2{\cdot}5} \quad (q=0), \quad m_d(x)=1-d|0{\cdot}5-x|^{2{\cdot}5} \quad (q=1).$$

Case 3: *A function with a periodic component*. Here we take

$$m_d(x)=d\cos(4\pi x) \quad (q=0), \quad m_d(x)=1+2x+d\cos(4\pi x) \quad (q=1).$$

We set the sample size at $n=100$ and take x_i equally spaced on $[0,1]$, even though the results in this paper apply to arbitrarily spaced x_i. The error standard deviation is $\sigma_\varepsilon=0{\cdot}25$ and $K=20$ knots are used, where the knot κ_k is the $k/(K+1)$th sample quantile of the x's. For every family of functions m_d we simulate from the model $y_i=m_d(x_i)+\varepsilon_i$, where ε_i are independently $N(0,\sigma_\varepsilon^2)$, and we compute the test statistics described in previous sections for every pair of vectors (x, Y), where $x=(x_1,\ldots,x_n)^{\mathrm{T}}$. The computed values are then compared with the 95% quantile obtained by simulation for $d=0$, corresponding to H_0. Thus the power function for a fixed level $\alpha=0{\cdot}05$ is obtained for a given family of functions. Altogether 5000 simulations were used for each member of the family of functions $m_d(.)$.

For every family $m_d(.)$, it was seen in the simulations that the power curves did not cross each other over the range of d corresponding to values of power between 0·05 and 1. Therefore, without losing too much information, one may consider only one value of d for each family. We chose this value so that the power of tests that perform better is about 0·8. Table 1 presents the average, maximum and minimum power for testing constant mean against a general alternative for the three cases. The superscripts denote the degree of the null polynomial and the degree of the penalised spline used to model the alternative. For example, a $(0, 1)$ superscript means test for a constant mean versus a linear spline. For tests F and R using a data-dependent smoothing parameter we indicate the estimation criterion used as a subscript of $\hat{\lambda}$. We omitted the subscript n because all tests are compared in finite samples, with $n=100$.

When $p=q=0$ we only used $\mathrm{RLRT}^{0,0}$ because the large mass at zero of $\mathrm{LRT}^{0,0}$ makes a test with fixed size difficult to design and unlikely to have good power. When $p=q=1$ we used $\mathrm{LRT}^{0,1}$ instead of $\mathrm{RLRT}^{0,1}$ because the fixed effects under the null and alternative are different.

Table 1. *Comparison of power function for tests for constant mean for three alternatives. The tests have been ordered by average power. The superscripts denote the degree of the null polynomial and the degree of the penalised spline used to model the alternative*

Test	Average	Maximum	Minimum	Test	Average	Maximum	Minimum
$\text{RLRT}^{0,0}$	0·85	0·87	0·81	$R^{0,1}(\hat{\lambda}_{\text{GCV}})$	0·66	0·76	0·53
$F^{0,0}(\hat{\lambda}_{\text{REML}})$	0·84	0·87	0·82	$F^{0,0}(1)$	0·64	0·85	0·28
$R^{0,0}(\hat{\lambda}_{\text{REML}})$	0·84	0·95	0·69	$F^{0,1}(1)$	0·64	0·95	0·20
$F^{0,0}(\hat{\lambda}_{\text{GCV}})$	0·83	0·84	0·81	$R^{0,1}(1)$	0·61	0·92	0·10
$R^{0,1}(\hat{\lambda}_{\text{ML}})$	0·79	0·88	0·66	$R^{0,0}(\hat{\lambda}_{\text{GCV}})$	0·59	0·87	0·27
$F^{0,1}(\hat{\lambda}_{\text{ML}})$	0·76	0·84	0·62	$R^{0,0}(1)$	0·58	0·85	0·16
$\text{LRT}^{0,1}$	0·67	0·85	0·47	vN	0·46	0·65	0·25

REML, restricted maximum likelihood; GCV, generalised crossvalidation; ML, maximum likelihood; vN, von Neumann.

Table 1 shows that all tests using maximum likelihood or restricted maximum likelihood estimation of the smoothing parameter can detect fairly well departures from a constant mean. Tests using restricted maximum likelihood perform better than tests using maximum likelihood. The generalised crossvalidation criterion for λ gave more mixed results. However, none of these tests was uniformly most powerful even in the restricted class of alternatives considered. With respect to the average power, $\text{RLRT}^{0,0}$ outperforms the other statistics considered, including $\text{LRT}^{0,1}$. The higher power of $\text{RLRT}^{0,0}$ compared to $\text{LRT}^{0,1}$ is probably caused by the large variance estimation bias when maximum likelihood rather than restricted maximum likelihood is used. The main advantage of LRT_n and RLRT_n statistics is that their null finite sample distributions can be obtained using a fast simulation algorithm, whereas bootstrap must be used for F and R. The tests using restricted maximum likelihood to select λ seem more powerful than those using generalised crossvalidation. For example, the average power of $R^{0,0}$ is 0·84 for restricted maximum likelihood and 0·59 for generalised crossvalidation. Similarly, $R^{1,1}$ has an average power of 0·89 with restricted maximum likelihood but only 0·54 with generalised crossvalidation.

Tests that use a fixed value of the smoothing parameter, $F^{0,0}(1)$, $F^{0,1}(1)$, $R^{0,0}(1)$, $R^{0,1}(1)$, etc., can detect departures from the null, but their power is reduced when the parameter does not match the degrees of freedom of the mean function. The von Neumann test for unspecified alternatives performs poorly.

Results for testing a linear null against a general alternative were similar to those for constant mean. In addition $F^{1,2}(\hat{\lambda}_{\text{ML}})$ and $R^{1,2}(\hat{\lambda}_{\text{ML}})$ performed poorly, giving worse results than tests with a fixed smoothing parameter. This is because of a 'luckier' a priori choice of the smoothing parameter, but indicates potential power problems when the maximum likelihood criterion is employed. In contrast, $\text{RLRT}^{1,1}$ continues to perform well together with $F^{1,1}(\hat{\lambda}_{\text{REML}})$ and $R^{1,1}(\hat{\lambda}_{\text{REML}})$. Other a priori fixed degrees of freedom were also used. Power was observed to be low when they did not match the degree of nonlinearity of the true regression function.

Acknowledgement

The authors thank the editor and two anonymous referees for their careful reading of the paper and constructive comments. The research of Claeskens is supported by the Federal Office for Scientific, Technical and Cultural Affairs of the Belgian government.

References

Aerts, M., Claeskens, G. & Hart, J. D. (1999). Testing the fit of a parametric function. *J. Am. Statist. Assoc.* **94**, 869–79.

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

Azzalini, A. & Bowman (1993). On the use of nonparametric regression for checking linear relationships. *J. R. Statist. Soc.* B **55**, 549–57.

Brumback, B., Ruppert, D. & Wand, M. P. (1999). Comment on 'Variable selection and function estimation in additive nonparametric regression using data-based prior' by Shively, Kohn and Wood. *J. Am. Statist. Assoc.* **94**, 794–7.

Cantoni, E. & Hastie, T. J. (2002). Degrees of freedom tests for smoothing splines. *Biometrika* **89**, 251–63.

Claeskens, G. (2004). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *J. R. Statist. Soc.* B **66**, 909–26.

Cleveland, W. S. & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *J. Am. Statist. Assoc.* **83**, 597–610.

Crainiceanu, C. M. & Ruppert, D. (2004a). Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *J. Mult. Anal.* **91**, 35–52.

Crainiceanu, C. M. & Ruppert, D. (2004b). Restricted likelihood ratio tests in nonparametric longitudinal models. *Statist. Sinica.* **14**, 713–29.

Crainiceanu, C. M. & Ruppert, D. (2004c). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc.* B **66**, 165–85.

Eubank, R. L. & Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Am. Statist. Assoc.* **85**, 387–92.

Härdle, W., Mammen, E. & Müller, M. (1998). Testing parametric versus semiparametric modeling in generalised linear models. *J. Am. Statist. Assoc.* **93**, 1461–74.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests.* New York: Springer.

Hastie, T. J. & Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Marx, B. D. & Eilers, P. H. C. (1998). Direct generalised additive modeling with penalised likelihood. *J. Comp. Statist. Data Anal.* **28**, 193–209.

Munson, P. J. & Jernigan, R. W. (1989). A cubic spline extension of the Durbin–Watson test. *Biometrika* **76**, 39–47.

Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *J. Am. Statist. Assoc.* **85**, 132–8.

Ruppert, D. (2002). Selecting the number of knots for penalised splines. *Comp. Statist. Data Anal.* **11**, 735–57.

Ruppert, D. & Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. New Zeal. J. Statist.* **42**, 205–23.

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

Self, S. G. & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Am. Statist. Assoc.* **82**, 605–10.

Self, S. G. & Liang, K. Y. (1995). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. R. Statist. Soc.* B **58**, 785–96.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–7.

Williams, E. J. (1959). *Regression Analysis.* New York: Wiley.

[*Received December* 2002. *Revised May* 2004]