

# Exact Likelihood Ratio Tests for Penalized Splines

BY CIPRIAN CRAINICEANU, DAVID RUPPERT, GERDA CLAESKENS, M.P. WAND

*Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA*

*School of Oper. Res. and Ind. Eng., Cornell University, Rhodes Hall, NY 14853, USA*

*Institute of Statistics, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium*

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

e-mail: ccrainic@jhsph.edu; davidr@orie.cornell.edu

e-mail: claeskens@stat.ucl.ac.be; wand@maths.unsw.edu.au

## SUMMARY

Penalized 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 between 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, (R)LRTs, for testing polynomial regression versus a general alternative modeled by penalized splines. Their spectral decompositions are used as the basis of fast simulation algorithms. We derive the asymptotic local power properties of (R)LRTs under weak conditions. In particular we characterize the local alternatives that are detected with asymptotic probability 1. Confidence intervals for the smoothing parameter are obtained by inverting the (R)LRT 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 power of several known tests is investigated and a small set of tests with good power properties is identified.

*Some key words:* Linear mixed models, penalized splines, smoothing, zero variance components.

## 1. INTRODUCTION

Penalized spline additive models have been described in Marx and Eilers (1998), Ruppert and Carroll (2000), and Aerts, Claeskens and Wand (2002). Only a small set of spline basis functions is needed for each covariate. Another advantage is the existence of mixed model representations (Brumback et al. 1999). Testing for simplifying assumptions, such as no or a linear covariate effect, can be reduced to testing for zero variance components. The key idea, to embed the parametric regression function (e.g. a polynomial) into a larger, non-parametric family (e.g. penalized splines), has been used by others, e.g., Cleveland and Devlin (1988), Azzalini and Bowman (1993), Hart (1997), Härdle, Mammen and Müller (1998), Aerts, Claeskens and Hart (1999), Ruppert, Wand and Carroll, (2003). 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 model uses penalized splines. The test 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 (LMMs) framework the fitted penalized splines are best linear unbiased predictors (BLUPs) and the smoothing parameters are ratios of variance components which can be estimated by maximum likelihood (ML) or restricted maximum likelihood (REML). Testing for a polynomial regression model is equivalent to testing that a variance component is zero. LRTs for null variance components are non-standard 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 LMMs were obtained by Crainiceanu and Ruppert (2004) for a fixed, but large number of knots. Under some additional assumptions, asymptotic results were also obtained by Claeskens (2002) when the number of knots increases with the sample size.

In §6 we derive the asymptotic local power properties of (R)LRTs under weak conditions. We show that the if the true value of the smoothing parameter converges to zero

slower than the eigenvalues of some design matrices, then (R)LRTs detect the alternative with asymptotic power 1.

In §7 we consider the problem of testing the null hypothesis of a fixed smoothing parameter (not necessarily zero) versus a general alternative using (R)LRTs. These results generalize those of Cantoni and Hastie (2002) who consider simple alternative hypotheses and impose restrictions on the design matrix. In §8 we obtain confidence intervals for the smoothing parameter by inverting the (R)LRT. By examining the splines with  $\lambda$  at each end of the interval, one can see the range of smoothness that is consistent with the data. We also investigate F and R tests in §9. Some of these tests use a smoothing parameter estimated under the alternative. We show that ignoring the variability of the smoothing parameter can have serious effects on the null distribution of the test statistic. Moreover, if the smoothing parameter is fixed a-priori, then the power is significantly diminished.

In §10 the power functions of tests for polynomial regression are compared in a simulation study and a small set of test statistics with good power properties is identified. The RLRT is among the best in terms of power. The F and R tests appear to be more powerful when they use a smoothing parameter selected by REML rather than ML, GCV or fixed a-priori.

## 2. FRAMEWORK

Consider the partially linear model  $y_i = W_i^T \gamma + m(x_i) + \epsilon_i$ , where  $\epsilon_i$  are i.i.d.  $N(0, \sigma_\epsilon^2)$ ,  $\epsilon_i$  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  $m(\cdot)$  is a smooth function. We consider the  $x$ 's to be random but their joint distribution is unspecified and could be degenerate, e.g., the  $x$ 's are equally spaced. We want to test if  $m(\cdot)$  is a  $p - q$ -th degree polynomial  $m(x, \theta) = \beta_0 + \beta_1 x + \dots + \beta_{p-q} x^{p-q}$ , where  $0 \leq q \leq p$  and  $\theta = (\beta_0, \dots, \beta_{p-q})^T$ . 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  $m(x, \Theta) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{k=1}^K b_k (x - \kappa_k)_+^p$ , where  $\Theta = (\theta^T, \beta_{p-q+1}, \dots, \beta_p, b_1, \dots, b_K)^T$  is the vector of regression coefficients, and

$\kappa_1 < \kappa_2 < \dots < \kappa_K$  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  $\kappa_k$  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 minimize

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

where  $\lambda$  is the smoothing parameter and  $D$  is a known positive semi-definite 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  $\Sigma$  is a positive definite matrix and  $0_{m \times l}$  is an  $m \times l$  matrix of zeros. This matrix  $D$  penalizes 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  $Y = (y_1, y_2, \dots, y_n)^T$ ,  $W$  be the matrix with the  $i$ -th row equal to  $W_i^T$ ,  $X$  be the matrix with the  $i$ -th row  $X_i = (1, x_i, \dots, x_i^p)$ , and  $Z$  be the matrix with  $i$ -th row  $Z_i = \{(x_i - \kappa_1)_+^p, \dots, (x_i - \kappa_K)_+^p\}$ .

### 3. PENALIZED SPLINES AS LINEAR MIXED MODELS

If we divide (1) by the error variance one obtains

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

where  $\beta = (\beta_0, \dots, \beta_p)^T$  and  $b = (b_1, \dots, b_K)^T$ . Define  $\sigma_b^2 = \lambda\sigma_\epsilon^2$ , consider the vectors  $\gamma$  and  $\beta$  as fixed parameters and the vector  $b$  as a set of random parameters with  $E(b) = 0$  and  $\text{cov}(b) = \sigma_b^2 \Sigma$ . If  $(b^T, \epsilon^T)^T$  is a normal random vector and  $b$  and  $\epsilon$  are independent then one obtains an equivalent model representation of the penalized spline in the form

of a linear mixed model (Brumback et al., 1999):

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

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

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

A second parameter estimation criterion for model (2) is the restricted log-likelihood

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

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

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

$$H_0 : \beta_{p-q+1} = \dots = \beta_p = 0 \text{ and } \sigma_b^2 = 0 \text{ (} \lambda = 0 \text{)} \quad (4)$$

against the alternative hypothesis  $H_A : \beta_{p-q+1} \neq 0$  or  $\dots$  or  $\beta_p \neq 0$  or  $\sigma_b^2 > 0$  ( $\lambda > 0$ ). Because 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  $\beta_{p-q+1} = \dots = \beta_p = 0$  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:

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

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

A generalization of (5) is

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

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

#### 4. (R)LRT TESTS FOR POLYNOMIAL REGRESSION

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

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

A similar test can be introduced based on the REML criterion (3). Because REML 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

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

Computing  $\text{LRT}_n$  (or  $\text{RLRT}_n$ ) is simple using standard software, such as R and S-PLUS (`lme` function) or SAS (`MIXED` procedure).

#### 5. NULL DISTRIBUTIONS OF (R)LRTs

While computing  $\text{LRT}_n$  and  $\text{RLRT}_n$  is easy, obtaining the null (finite sample or asymptotic) distribution is more challenging, as we discuss in the following.

### 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  $H_0 : \sigma_b^2 = 0$  ( $\lambda = 0$ ) vs.  $H_A : \sigma_b^2 > 0$  ( $\lambda > 0$ ). Because the parameter under the null is on the boundary, the classical result that  $\text{LRT}_n \Rightarrow \chi_1^2$  under  $H_0$ , where “ $\Rightarrow$ ” denotes weak convergence, does not hold. One may be tempted to believe that the asymptotic theory of Self and Liang (1987, 1995) and Stram and Lee (1994) for boundary problems apply. These results suggest that

$$\text{LRT}_n \Rightarrow 0.5\chi_0^2 + 0.5\chi_1^2 \text{ under } H_0, \quad (7)$$

where  $\chi_0^2$  denotes a point mass at zero. However, (7) is derived under the assumption that the response variable vector can be partitioned into  $J$  i.i.d. subvectors with  $J \rightarrow \infty$ .

Crainiceanu, Ruppert and Vogelsang (2003) show that the asymptotic probability mass at zero for  $\text{LRT}_n$  (and  $\text{RLRT}_n$ ) is practically constant over a wide range of number of knots  $K$ , and is approximately 0.95 for LRT and 0.65 for RLRT. These values are much larger than 0.5 but provide excellent approximations of the finite sample probabilities.

Crainiceanu and Ruppert (2004) found that the finite sample distributions of  $(\text{R})\text{LRT}_n$  converge very quickly to an asymptotic distribution different from  $0.5\chi_0^2 + 0.5\chi_1^2$ , the latter providing a very conservative approximation of the null finite sample and asymptotic distributions. Although the asymptotics in Crainiceanu and Ruppert (2004) could be used, they are unnecessary since exact critical values are easy to compute.

### 5.2. Finite sample null distribution of $(\text{R})\text{LRT}$

The finite sample and asymptotic distributions of  $\text{LRT}_n$  and  $\text{RLRT}_n$  for testing null hypotheses including zero random effects variance in LMMs with one variance component were derived by Crainiceanu and Ruppert (2004). These results are applicable in the penalized splines framework due to the LMM representation (2).

Let  $\mu_{s,n}$  and  $\xi_{s,n}$  be the  $K$  eigenvalues of the  $K \times K$  matrices  $\Sigma^{1/2} Z^T P_0 Z \Sigma^{1/2}$  and

$\Sigma^{1/2} Z^T Z \Sigma^{1/2}$  respectively, where  $P_0 = I_n - X(X^T X)^{-1} X^T$ . Then

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

where  $u_s$  for  $s = 1, \dots, K$ ,  $w_s$  for  $s = 1, \dots, n - p - 1$ , are independent  $N(0, 1)$ . Here “ $\stackrel{\mathcal{D}}{=}$ ” denotes equality in distribution and

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

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

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

Although appearing complex, these distributions are easily simulated. Crainiceanu and Ruppert (2004) provide a fast simulation algorithm that takes advantage of the spectral representation of (R)LRT $_n$ .

## 6. LOCAL POWER PROPERTIES

We now focus on the local asymptotic power properties of (R)LRT $_n$  for polynomial regression. We state the theorem for the more general case of LMMs with one random effects variance component.

**THEOREM 6.1.** *Consider testing hypotheses described in equation (5) for a LMM with one variance component as described in equation (2). Assume that there exists a  $a > 0$  such that for every  $s$ , the  $K$  eigenvalues  $\mu_{s,n}$  of the matrix  $\Sigma^{1/2} Z^T P_0 Z \Sigma^{1/2}$  satisfy  $\lim_{n \rightarrow \infty} (n^{-a} \mu_{s,n}) = \mu_s$ , where not all  $\mu_s$  are zero. Then*

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

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

where

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

and  $w_s$  are i.i.d.  $N(0, 1)$  random variables.



2. If the true value of the variance ratio  $\sigma_b^2/\sigma_\epsilon^2$  is  $\lambda_{n,0} = n^{-b}d_0$  with  $0 \leq b < a$  then for every  $x > 0$

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

The complete proof is contained in the attached Technical report (Crainiceanu et al., 2004). These results generalize the null asymptotic results obtained by Crainiceanu and Ruppert (2004) (Theorem 2) for RLRT when  $d_0 = 0$ . Similar results hold true for  $\text{LRT}_n$ .

When applied to the particular case of penalized splines, the first result of the theorem provides the asymptotic power of  $\text{RLRT}_n$  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 behavior of the eigenvalues of the  $\Sigma^{1/2}Z^T P_0 Z \Sigma^{1/2}$  matrix is the standard condition used to ensure the asymptotic consistency of parameters estimates. Note that  $a = 1$  for designs that are generated from a random sample.

If the true parameter converges to zero slower than  $n^{-a}$  then the  $\text{RLRT}_n$  rejection regions have asymptotic probability 1.

## 7. TESTING FOR A FIXED SMOOTHING PARAMETER

We now test the hypothesis (6) about  $\lambda = \sigma_b^2/\sigma_\epsilon^2$ . Suppose that  $\lambda_0$  is the true value of  $\lambda$ . Crainiceanu and Ruppert (2004) showed that the  $\text{RLRT}_n$  statistic for testing (6) has the null spectral decomposition

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

where

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

and  $w_s$ , for  $s = 1, \dots, n - p - 1$ , are independent  $N(0, 1)$ , and  $\mu_{s,n}$  are the  $K$  eigenvalues of the  $K \times K$  matrix  $\Sigma^{1/2}Z^T P_0 Z \Sigma^{1/2}$ . A similar result holds for  $\text{LRT}_n$ .

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

Several relevant results can be obtained as a byproduct 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  $\text{RLRT}_n$  for testing the number of degrees of freedom for penalized splines, as described in Cantoni and Hastie (2002), but without their assumption that  $Z^T X = 0$ . For  $\Lambda = (\lambda_0, \infty)$  we obtain the distribution of  $\text{RLRT}_n$  for testing

$$H_0 : \lambda = \lambda_0 \text{ vs. } H_A : \lambda > \lambda_0 . \quad (11)$$

For  $\Lambda = [0, \infty) \setminus \{\lambda_0\}$  we obtain the distribution of  $\text{RLRT}_n$  for testing

$$H_0 : \lambda = \lambda_0 \text{ vs. } H_A : \lambda \neq \lambda_0 . \quad (12)$$

## 8. CONFIDENCE INTERVALS FOR THE SMOOTHING PARAMETER

Our ability to quickly simulate the null finite sample distribution of  $\text{RLRT}_n$  allows us to obtain confidence intervals for the smoothing parameter by inverting the  $\text{RLRT}_n$ . Indeed, we define the  $(1 - \alpha)$  - level confidence interval for  $\lambda$

$$\text{CI}_\alpha = \{\lambda_0 | p(\lambda_0) \geq \alpha\} , \quad (13)$$

where  $p(\lambda_0)$  is the p-value for the  $\text{RLRT}_n$  statistic for testing (12). Because  $p(\lambda_0)$  can be obtained in seconds,  $\text{CI}_\alpha$  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). The data are displayed in Figure 1–(a). Linear penalized splines with  $K = 15$  knots were used. We obtained  $\hat{\lambda}_{\text{REML}} = 0.0056$  corresponding to 4.13 degrees of freedom of regression. We used 100,000 simulations for the null distribution of  $\text{RLRT}_n$  to calculate  $p(\lambda_0)$  for each  $\lambda_0$  on a grid. The grid described earlier was enough to determine the regions where the cutoff points are situated but was not enough to pinpoint them. Therefore, in a second stage we calculated  $p(\lambda_0)$  on very fine grids around the cutoff points. We obtained  $\text{CI}_\alpha = [0.0014, 0.0870]$ , which corresponds to a confidence interval  $[3.32, 6.83]$  for the degrees of freedom. Figure 1–(b) shows a zoom in of the p-value plot versus  $\log_{10}(\lambda)$ . Figure 1–(a) shows the curves corresponding to the lower and upper bounds of  $\text{CI}_\alpha$ .

## 9. F AND R TESTS

Suppose the estimated smooth is  $\hat{Y}(\lambda) = S_\lambda Y$ , where  $S_\lambda = (\mathcal{X}^T \mathcal{X} + \frac{1}{\lambda} D)^{-1} \mathcal{X}^T Y$  is the smoother matrix and  $\mathcal{X} = [X|Z]$ . The estimated residual vector is  $\hat{\epsilon}_\lambda = (I_n - S_\lambda)Y$ . By analogy with linear regression, the degrees of freedom for residuals is  $\gamma_\lambda = \text{tr} \left\{ (I_n - S_\lambda)^2 \right\}$ , where  $\lambda \rightarrow \gamma_\lambda$  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

$$H_0 : \lambda = \lambda_0 \text{ vs. } H_A : \lambda = \lambda_1 , \quad (14)$$

where  $\lambda_0 < \lambda_1$  is defined as

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

where for simplicity we denoted  $\gamma_0 = \gamma_{\lambda_0}$  and  $\gamma_1 = \gamma_{\lambda_1}$  (Hastie and Tibshirani, 1990). Cantoni and Hastie (2002) proposed the following related statistic for testing (14)

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

The finite sample distributions of F and R statistics have either been approximated (Hastie and Tibshirani, 1990) or derived (Cantoni and Hastie, 2002). Cantoni and Hastie

(2002) suggest using these statistics to testing  $H_0 : \lambda = \lambda_0$  vs.  $H_A : \lambda > \lambda_0$ , by replacing  $\lambda_1$  in the F or R tests with an estimator  $\hat{\lambda}$ . Cantoni and Hastie suggested using  $\hat{\lambda}_{\text{REML}}(\lambda_0)$ , which is the estimated smoothing parameter using REML restricted to  $[\lambda_0, \infty)$ , but other criteria can be used as well. For simplicity we denote  $\hat{\lambda} = \hat{\lambda}_{\text{REML}}(\lambda_0)$ .

While defining the tests by replacing  $\lambda_1$  by  $\hat{\lambda}$  is straightforward, deriving the null distribution of the test statistics is not. Estimating  $\lambda$  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 and Hastie (2002), inaccurately approximates null distributions.

The probability that  $\hat{\lambda} = \lambda_0$  is equal to  $\text{pr}(\hat{\lambda}_{\text{REML}} \leq \lambda_0)$ , where  $\hat{\lambda}_{\text{REML}}$  is the unrestricted REML of  $\lambda$ . Crainiceanu and Ruppert (2004) showed that this probability is greater than 0.5 and is equal to the null probability mass at  $\lambda_0$  of  $\text{RLRT}_n$  and of the  $R_{\lambda_0, \hat{\lambda}}$  statistics. Furthermore, for any fixed  $\lambda_1 > \lambda_0$ ,  $R_{\lambda_0, \lambda_1}$  has no mass at  $\lambda_0$ . 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  $H_0 : \lambda = \lambda_0$  vs.  $H_A : \lambda \neq \lambda_0$ , the (R)LRT $_n$  and  $R_{\lambda_0, \hat{\lambda}}$  statistics will have probability mass at zero (see Crainiceanu and Ruppert, 2004), showing that the  $\chi_1^2$  approximation cannot be used even when  $\lambda_0 > 0$ .

### 9.1. 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 ( $\lambda = 0$ ), and by  $\text{RSS}(p, \hat{\lambda})$  the residual sum of squares for the  $p$ -degree penalized spline regression. The F-statistic is defined as

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

$F^{q,p}(\hat{\lambda})$  is used by Hastie and Tibshirani (1990). 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 sum 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 semi or nonparametric models, these three quantities will differ somewhat but should be similar. The tests of Härdle, Mammen, and Müller (1998) for generalized regression models are based on the quasi-likelihood estimates of Severini and Staniswalis (1994). When specialized 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 and 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 and 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 penalized 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 and 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 fit to residuals from a linear polynomial regression.

The R-statistic is defined as

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

where  $S_{p,\hat{\lambda}}$  is the smoother matrix corresponding to the degree  $p$  penalized spline and 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 using one of the available criteria. For reasons of convenience we consider only the ML, REML, and GCV criteria with REML 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.

### 9.2. 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  $\lambda$ , the mean

response is a degree  $p$  spline function with  $\text{tr}(S_\lambda)$  degrees of freedom (Ruppert et al., 2003), where  $\text{tr}(S_\lambda)$  is an increasing, continuous function of  $\lambda$  from  $p + 1$  for  $\lambda = 0$  to  $p + 1 + K$  for  $\lambda = \infty$ . Ruppert et al. use  $\lambda^p$  where we use  $\lambda^{-1}$ . We can choose  $\lambda_1$  to match any given number of degrees of freedom in the interval  $[p + 1, p + 1 + K]$ . For example, for  $\lambda_1$  corresponding to  $p + 2$  degrees of freedom of fit we define

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

where we used the usual notations. The “1” in the  $F^{q,p}(1)$  and  $R^{q,p}(1)$  notations stands for “one degree of freedom more than the  $p + 1$  degrees of freedom of the  $p$ -degree polynomial”. Similarly we can define  $F^{q,p}(2)$ ,  $R^{q,p}(2)$ , etc. A potential problem could be that by a-priori fixing 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 (VN) type test for unspecified alternatives described in Hart (1997). This test does not require the specification of an alternative hypothesis. Another class of possible tests are those of Durbin-Watson type (Munson and Jernigan, 1989), but these proved disappointing in a power study by Azzalini and Bowman (1993).

## 10. POWER SIMULATIONS

In §4 and §9 several statistics were described for testing  $p - q$  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 modeled either using a piecewise constant ( $p = 0$ ) or a linear ( $p = 1$ ) spline. When testing for a linear polynomial, the alternative hypothesis is modeled either using 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 and  $d = 0$  corresponds to  $H_0$ .

**Increasing function**

$$m_d(x) = \frac{d}{1 + e^{10(0.5-x)}} \quad (q = 0), \quad m_d(x) = x + \frac{d}{1 + e^{10(0.5-x)}} \quad (q = 1)$$

**Concave function**

$$m_d(x) = -d|0.5 - x|^{2.5} \quad (q = 0), \quad m_d(x) = 1 - d|0.5 - x|^{2.5} \quad (q = 1)$$

**Function with periodic component**

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

We set the sample size  $n = 100$  and take  $x_i$  equally spaced on  $[0, 1]$ , even though results in this paper apply to arbitrarily spaced  $x_i$ . The error standard deviation is  $\sigma_\epsilon = 0.25$  and  $K = 20$  knots are used, where the  $\kappa_k$  knot is the sample  $k/(K + 1)$  quantile of  $x$ 's. For every family of functions  $m_d$  we simulate from the model  $y_i = m_d(x_i) + \epsilon_i$ , where  $\epsilon_i$  are independent  $N(0, \sigma_\epsilon^2)$  and compute the tests statistics described in previous sections for every pair of vectors  $(x, Y)$ , where  $x = (x_1, \dots, x_n)^T$ . The computed values are then compared with the 95% quantile obtained by simulation for  $d = 0$  ( $H_0$ ). Thus the power function for a fixed level  $\alpha = 0.05$  is obtained for a given family of functions. 5,000 simulations were used for each member of the family of functions  $m_d(\cdot)$ .

For every family,  $m_d(\cdot)$ , it was seen in the simulations that the power curves do 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 around 0.8. Table 1 presents the average, maximum, and minimum power for testing constant mean against a general alternative over the three alternative families. The superscripts denote the degree of the null polynomial and the degree of the penalized 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 eliminated the subscript  $n$  because all tests are compared in finite samples ( $n = 100$ ).

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

Table 1 shows that all tests using ML or REML estimation of the smoothing parameter are able to detect fairly well departures from a constant mean. Tests using REML perform better than tests using ML. GCV criterion for  $\lambda$  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 tests considered, including  $\text{LRT}^{0,1}$ . The higher power of  $\text{RLRT}^{0,0}$  compared to  $\text{LRT}^{0,1}$  is probably due to the large variance estimation bias when ML rather than REML is used. The main advantage of  $(\text{R})\text{LRT}_n$  statistics is that the null finite sample distribution of  $(\text{R})\text{LRT}_n$  can be obtained using the algorithm described in §5, whereas bootstrap needs to be used for F and R. R tests using REML to select  $\lambda$  seem more powerful than those using GCV. For example, the average power of  $\text{R}^{0,0}$  is 0.84 for REML and 0.59 for GCV. Similarly,  $\text{R}^{1,1}$  has an average power of 0.89 with REML but only 0.54 with GCV.

Tests that use a fixed value of the smoothing parameter,  $\text{F}^{0,0}(1)$ ,  $\text{F}^{0,1}(1)$ ,  $\text{R}^{0,0}(1)$ ,  $\text{R}^{0,1}(1)$ , 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 VN test for unspecified alternatives performs poorly.

Results for testing a linear null against a general alternative were similar to the ones 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 due to a “luckier” a-priori choice of the smoothing parameter, but indicates potential power problems when the ML 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 non-linearity of the true regression function.



## ACKNOWLEDGEMENTS

The authors thank the 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**, 869–879.
- AERTS, M., CLAESKENS, G., & WAND, M.P. (2002). Some theory for penalized spline additive models. *J. Statist. Plann. Inference*, **103**, 455–470.
- AZZALINI, A., & BOWMAN (1993). On the use of nonparametric regression for checking linear relationships. *J. R. Statist. Soc. B*, **55**(3), 549–557.
- 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. Amer. Stat. Assoc.*, **94**, 794–797.
- CANTONI, E., & HASTIE, T.J. (2002). Degrees of freedom tests for smoothing splines. *Biometrika*, **89**, 251–263.
- CLAESKENS, G. (2002). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *submitted*.
- CLEVELAND, W.S. & DEVLIN, S.J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting, *J. Amer. Stat. Assoc.*, **83**, 597–610.
- CRAINICEANU, C.M. (2003). Nonparametric likelihood ratio testing, *Cornell University Ph.D. Thesis*.
- CRAINICEANU, C. M., & RUPPERT, D., (2004a). Likelihood ratio tests in linear mixed models with one variance component. *J.R. Statist. Soc. B*, **66**, 165–185.
- CRAINICEANU, C. M. and Ruppert, D., Claeskens, G. & Wand, M.P. (2004b) Proofs for the paper “Exact Likelihood Ratio Tests for Penalized Splines”. Technical Report.
- CRAINICEANU, C.M., RUPPERT, D., & VOGELSANG, T.J. (2003). Some properties of Likelihood Ratio Tests in Linear Mixed Models. *submitted*, available at [www.orie.cornell.edu/~davidr/papers](http://www.orie.cornell.edu/~davidr/papers).
- EUBANK, R. L., & SPIEGELMAN, C. H. (1990). Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models. *J. Amer. Stat. Assoc.*, **85**, 387–392.
- HÄRDLE, W., MAMMEN, E., & MULLER, M. (1994). Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models. *J. Amer. Stat. Assoc.*, **93**, 1461–1474.

- HART, J.D. (1997). *Nonparametric smoothing and lack-of-fit tests*. New York: Springer.
- HASTIE, T.J. (1996). Pseudosplines. *J. R. Statist. Soc. B*, **58**, 379–396.
- HASTIE, T.J., & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- MARX, B.D., & EILERS, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. *J. Comp. Statist. & Data Anal.*, **28**, 193–209.
- MUNSON, P.J. and 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. Amer. Stat. Assoc.*, **85**, 132–138.
- RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comp. Statist. & Data Anal.*, **11**, 735–757.
- RUPPERT, D., & CARROLL, R.J. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand J. Statist.*, **42**(2), 205–223.
- RUPPERT, D., WAND, M.P., & CARROLL, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: 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. Amer. Stat. Assoc.*, **82**, 605–610.
- SELF, S.G., & LIANG, K.Y. (1995). On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. *J. R. Statist. Soc. B*, **58**, 785–796.
- SEVERINI, T.A., & STANISWALIS, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. of the Amer. Stat. Assoc.*, **89**, 501–511.
- STRAM, D.O., & LEE, J.W. (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, **50**, 1171–1177.
- WILLIAMS, E.J. (1959). *Regression Analysis*, New York: Wiley.

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 and upper bounds of the 95% CI for  $\lambda$ . (b) – Zoom-in of the most interesting region of the RLRT p-value graph versus  $\log_{10}(\lambda)$  graph.

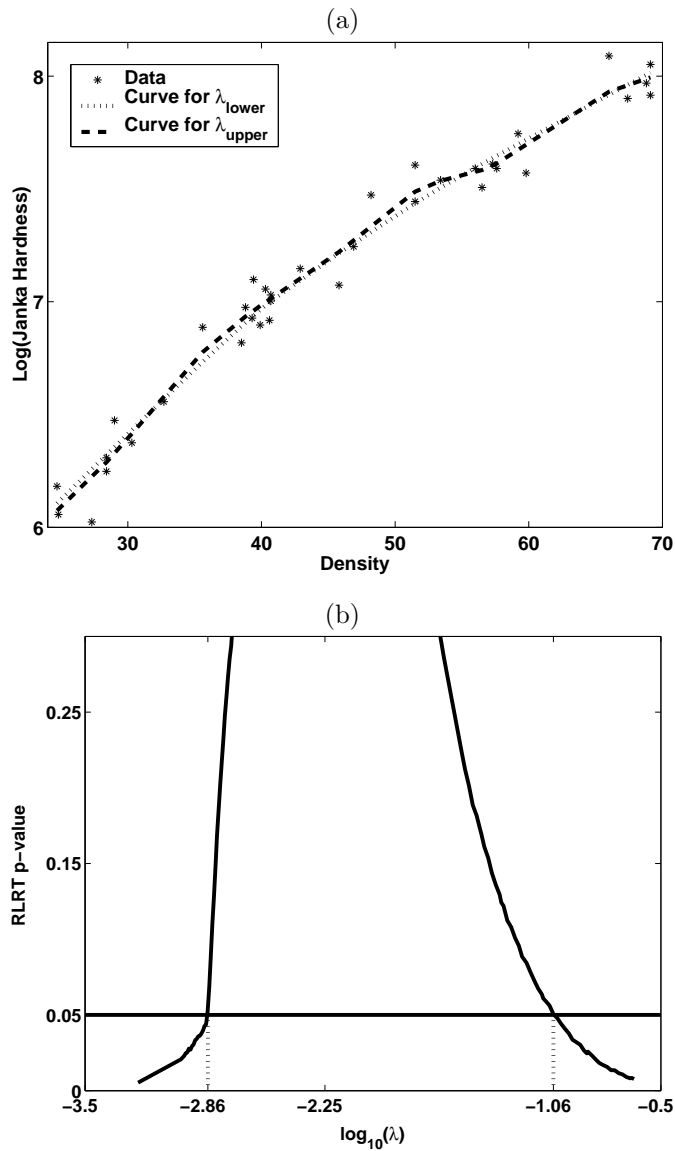


Table 1. Comparison of power function for tests for constant mean for three alternatives. The tests have been ordered by average power. The superscripts denote the degree of the null polynomial and the degree of the penalized spline used to model the alternative.

Test	Average	Maximum	Minimum
$\text{RLRT}^{0,0}$	0.85	0.87	0.81
$\text{F}^{0,0}(\hat{\lambda}_{\text{REML}})$	0.84	0.87	0.82
$\text{R}^{0,0}(\hat{\lambda}_{\text{REML}})$	0.84	0.95	0.69
$\text{F}^{0,0}(\hat{\lambda}_{\text{GCV}})$	0.83	0.84	0.81
$\text{R}^{0,1}(\hat{\lambda}_{\text{ML}})$	0.79	0.88	0.66
$\text{F}^{0,1}(\hat{\lambda}_{\text{ML}})$	0.76	0.84	0.62
$\text{LRT}^{0,1}$	0.67	0.85	0.47
$\text{R}^{0,1}(\hat{\lambda}_{\text{GCV}})$	0.66	0.76	0.53
$\text{F}^{0,0}(1)$	0.64	0.85	0.28
$\text{F}^{0,1}(1)$	0.64	0.95	0.20
$\text{R}^{0,1}(1)$	0.61	0.92	0.10
$\text{R}^{0,0}(\hat{\lambda}_{\text{GCV}})$	0.59	0.87	0.27
$\text{R}^{0,0}(1)$	0.58	0.85	0.16
VN	0.46	0.65	0.25
$\text{F}^{0,1}(\hat{\lambda}_{\text{GCV}})$	0.40	0.69	0.09