Testing for Cubic Smoothing Splines under Dependent Data

Tapio Nummi^{1,*}, Jianxin Pan², Tarja Siren³, and Kun Liu²

¹Tampere School of Public Health, FI-33014 University of Tampere, Finland

²School of Mathematics, The University of Manchester, M13 9PL, UK

³Department of Mathematics and Statistics, FI-33014 University of Tampere, Finland

**email:* tapio.nummi@uta.fi

SUMMARY: In most research on smoothing splines the focus has been on estimation, while inference, especially hypothesis testing, has received less attention. By defining design matrices for fixed and random effects and the structure of the covariance matrices of random errors in an appropriate way, the cubic smoothing spline admits a mixed model formulation, which places this nonparametric smoother firmly in a parametric setting. Thus nonlinear curves can be included with random effects and random coefficients. 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 an exact F-test for the situation and investigate its performance in a real pine stem data set and by simulation experiments. Under certain conditions the suggested methods can also be applied when the data are dependent.

KEY WORDS: Curve fitting; Forest harvesting; F-test; Linear regression.

1. Introduction

Smoothing splines have been introduced into the linear mixed model framework in recent years (e.g., Verbyla et al., 1999; Brumback and Rice, 1998; Zhang, Lin, and Sowers, 1998; Wang, 1998). This methodology enables flexible nonlinear curves to be fitted with fixed and random effects and correlated random errors. When applying statistical models simple linear models are often preferred due to their simplicity and interpretability. If some particular hypothesized form for the model and the alternative can be given a simple lack of fit test can often be easily performed. However, the hypothesized model may be difficult or impossible to test against all possible kinds of relevant alternatives. Therefore, efficient tests which perform well in a more general manner are needed. Most of the existing methods for testing general departures from a parametric model are based on nonparametric regression models such as kernel estimation (Azzalini and Bowman, 1993), local polynomial regression (Cleveland and Devlin, 1998) and smoothing spline (Green and Silverman, 1994). Our main focus in this paper is testing in a simple linear regression situation against the spline alternative, but this approach can also easily be extended to more general spline testing situations. One advantage of the methods proposed is that the distribution of the proposed test statistics is exact and it can also be applied under certain types of correlated sequences. The method is also computationally very simple.

In Section 2 we present some basic results needed in the estimation of cubic smoothing splines. We also present an approximation technique showing how to consider the estimation under the ordinary linear model framework by using projection matrices. In Section 3 a test statistic based on F-test is proposed. In Section 4 we apply the suggested test for a real data set arising from the forest industry. Moreover, we investigated under fairly general conditions how well the approximation utilized works when performing the F-test. The power of the

suggested testing method is also compared with 6 other statistical tests proposed for the situation.

2. Fitting a cubic smoothing spline

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ be the $n \times 1$ vector of observations. Suppose that design points (knots) $x_1, x_2, \dots, x_n$ are given on some interval $[a, b]$ satisfying $a < x_1 < x_2 < \dots < x_n < b$. Observations are assumed to satisfy

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

where $g(\cdot)$ is a smooth twice-differentiable curve and ε_i are independent and identically normally distributed random errors with zero mean and variance σ^2 , i.e., $\varepsilon_i \sim N(0, \sigma^2)$. In vector form the model (1) is written as $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$, where $\mathbf{g} = (g(x_1), g(x_2), \dots, g(x_n))'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Let $h_j = x_{j+1} - x_j$, $j = 1, 2, \dots, (n-1)$. Define the non-zero elements of banded $n \times (n-2)$ and $(n-2) \times (n-2)$ matrices $\boldsymbol{\nabla}$ and $\boldsymbol{\Delta}$ respectively as

$$\nabla_{k,k} = \frac{1}{h_k}, \quad \nabla_{k+1,k} = -\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right), \quad \nabla_{k+2,k} = \frac{1}{h_{k+1}}$$

and

$$\Delta_{k,k+1} = \Delta_{k+1,k} = \frac{h_{k+1}}{6}, \quad \Delta_{k,k} = \frac{h_k + h_{k+1}}{3},$$

where $k = 1, 2, \dots, (n-2)$. If we denote $\mathbf{K} = \boldsymbol{\nabla} \boldsymbol{\Delta}^{-1} \boldsymbol{\nabla}'$, the solution (at the design points) to the penalized least squares criterion $Q = (\mathbf{g} - \mathbf{y})'(\mathbf{g} - \mathbf{y}) + \alpha \mathbf{g}' \mathbf{K} \mathbf{g}$ is the cubic smoothing spline (e.g., Green and Silverman, 1994)

$$\hat{\mathbf{g}} = (\mathbf{I} + \alpha \mathbf{K})^{-1} \mathbf{y} = \mathbf{S}_\alpha \mathbf{y}. \quad (2)$$

The matrix $(\mathbf{I} + \alpha \mathbf{K})^{-1} = \mathbf{S}_\alpha$ is called the smoother matrix with smoothing parameter α .

The effective degrees of freedom for a smoother matrix are often defined as $\text{edf} = \text{tr } \mathbf{S}_\alpha$.

We now present a brief review of interesting connection of smoothing splines and mixed models (see, e.g., Verbyla et al., 1999). This can be shown as follows. Let $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$, where $\mathbf{x} = (x_1, \dots, x_n)'$ and by the special form of ∇ we note that $\mathbf{X}'\nabla = \mathbf{0}$ and after some manipulations

$$\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{\Delta}^{-1})^{-1}\mathbf{Z}',$$

where $\mathbf{Z} = \nabla(\nabla'\nabla)^{-1}$. The fitted spline is now $\hat{\mathbf{g}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{\Delta}^{-1})^{-1}\mathbf{Z}'\mathbf{y}$. These estimates can be seen as the BLUE and BLUP solutions of the mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ and $\mathbf{Z}$ are defined as before and $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{\Delta})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with smoothing parameter as the variance ratio $\alpha = \sigma^2/\sigma_u^2$. We may also rewrite $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_*\mathbf{u}_* + \boldsymbol{\varepsilon}$, where $\mathbf{Z}_* = \mathbf{Z}\mathbf{\Delta}^{1/2}$ and $\mathbf{u}_* = \mathbf{\Delta}^{-1/2}\mathbf{u}$ with $\mathbf{u}_* \sim N(\mathbf{0}, \sigma_u^2\mathbf{I})$, where $\mathbf{\Delta}^{1/2}$ is a matrix with $\mathbf{\Delta}^{1/2}\mathbf{\Delta}^{1/2} = \mathbf{\Delta}$. The mixed model formulation of spline solution has a nice interpretation. The fixed part $\mathbf{X}\boldsymbol{\beta}$ of the model is a straight line and the random part $\mathbf{Z}\mathbf{u}$ reflects the “spline features” around the line. If σ_u^2 is small the smoothing parameter α is large and the spline features of the curve are weak. Then the fitted curve $\hat{\mathbf{g}}$ is very close to a straight line. If, on the other hand, $\alpha \rightarrow 0$ the fit is an interpolating spline.

We now consider the case of dependent errors, i.e., $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{R}$. In this setting the penalized least squares criterion is $Q = (\mathbf{g} - \mathbf{y})'\mathbf{H}(\mathbf{g} - \mathbf{y}) + \alpha\mathbf{g}'\mathbf{K}\mathbf{g}$, with $\mathbf{H} = \mathbf{R}^{-1}$ and the analog of (2) is $\hat{\mathbf{g}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{y}$. In line with Nummi and Koskela (2007), when $\mathbf{R}$ has an appropriate special form, it can be seen that the equation for $\hat{\mathbf{g}}$ reduces to (2). This is a very important simplification, since spline estimates in this case are simple linear functions of the observations. One obvious condition that leads to this simplification is

$$\mathbf{K}\mathbf{R} = \mathbf{K}. \quad (3)$$

Indeed, if this condition holds, we have $\mathbf{K} = \mathbf{KH}$ and therefore $\hat{\mathbf{g}} = (\mathbf{H} + \alpha\mathbf{KH})^{-1}\mathbf{H}\mathbf{y} = (\mathbf{I} + \alpha\mathbf{K})^{-1}\mathbf{y}$. Unfortunately, covariance structures arising from the general times series models, e.g., AR(1), may not satisfy the condition (3).

A set of covariance matrices satisfying the condition (3) can be generated by using the formula $\mathbf{R} = \mathbf{I} + \mathbf{FM}$, for some $\mathbf{F}$, where $\mathbf{M} = \mathbf{I} - \nabla(\nabla'\nabla)^{-1}\nabla'$. Note that here $\mathbf{R}$ must be positive definite. One choice for $\mathbf{F}$ could be $\mathbf{F} = \mathbf{M}_*\mathbf{D}\mathbf{M}_*$, where the h ($\leq n$) columns of $\mathbf{M}_*$ are chosen from the column space of $\mathbf{M}$ and $\mathbf{D}$ is a $h \times h$ covariance matrix. A more familiar set of covariance structures is obtained when we choose $\mathbf{F} = \mathbf{X}_*\mathbf{D}\mathbf{X}_*'$, where the columns of $\mathbf{X}_*$ consists of either the first, the second, or both columns of $\mathbf{X}$. Such structures include independent $\mathbf{R} = \mathbf{I}$, uniform (or compound symmetry) $\mathbf{R} = d^2\mathbf{1}\mathbf{1}' + \mathbf{I}$ and linear structure $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$ as special cases. From this point on, unless otherwise stated, we assume that $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{R}$, where $\mathbf{R}$ takes the form $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$. The estimator utilized will thus have the simple form (2).

The matrix $(\mathbf{I} + \alpha\mathbf{K})^{-1}$ is not, in general, a projection matrix, but it is possible to plausibly approximate it by a projection matrix. Write $\mathbf{K}$ in the spectral form $\mathbf{K} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'$, where $\mathbf{T}$ is the matrix of n orthonormal eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_1, \dots, \lambda_n$ of $\mathbf{K}$. We then have $\mathbf{S}_\alpha = \mathbf{T}(\mathbf{I} + \alpha\mathbf{\Lambda})^{-1}\mathbf{T}'$. The set of eigenvectors of $\mathbf{S}_\alpha$ and $\mathbf{K}$ is the same and the eigenvalues are connected such that the eigenvalues of $\mathbf{S}_\alpha$ are $\gamma = 1/(1 + \alpha\lambda)$. Note that we can set $\mathbf{t}_1 = \mathbf{1}/\sqrt{n}$ and $\mathbf{t}_2$ is given by $\mathbf{t}_{2i} = (x_i - \bar{x})/S_x$, where $\bar{x}$ is the mean of the x_i 's and $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. Thus, $\mathbf{t}_1$ and $\mathbf{t}_2$ span the subspace corresponding to the straight line model. Next, it is assumed that the eigenvectors $\mathbf{T}$ are ordered according to eigenvalues of $\mathbf{S}_\alpha$. It is well known (see, e.g., Section 5.4.1 of Hastie, Tibshirani, and Friedman, 2001) that the first two eigenvalues of $\mathbf{S}_\alpha$ are always 1 and the sequence of eigenvectors $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$ appears to increase in complexity like a sequence of orthogonal polynomials defined on $\mathbf{x}$. This idea is utilized here to construct an approximation

for the spline fit $\hat{\mathbf{g}}$ by utilizing $c(\geq 3)$ first eigenvectors of $\mathbf{T}$. We denote the matrix consisting these c eigenvalues by $\mathbf{T}_*$. A more thorough consideration of spline approximations can be found in Hastie (1996). For approximating a penalized spline with a regression spline we refer to discussions in Wood (2006) and Hastie and Tibshirani (1990). In some cases an efficient algorithm is needed to compute the eigenvectors needed in spline approximation. For this purpose we may refer to the algorithms applied for the thin plate regression in Wood (2003), for example.

Our approximation minimizes the least squares criterion $(\mathbf{y} - \mathbf{T}_*\boldsymbol{\beta})'(\mathbf{y} - \mathbf{T}_*\boldsymbol{\beta})$ and the fit is simply computed by $\tilde{\mathbf{y}} = \mathbf{T}_*\mathbf{T}_*'\mathbf{y} = \mathbf{P}\mathbf{y}$, where $\mathbf{P} = \mathbf{T}_*\mathbf{T}_*'$ is now a projection matrix. Note that the fit $\tilde{\mathbf{y}}$ is also a natural spline, since it is a linear combination of natural splines $\mathbf{t}_1, \dots, \mathbf{t}_c$. The number c of columns $\mathbf{T}$ to be included in $\mathbf{T}_*$ can be determined using standard model selection techniques (e.g., the information criteria AIC or BIC). The smoother matrix $\mathbf{S}_\alpha$ and the smoothing parameter α need not be computed. If α is estimated we may set c equal to the smallest integer greater than $\text{edf} - 1$, for example. The approximation has several advantages. First, it is quite accurate for relatively smooth data. Secondly, $\mathbf{P}$ is a projection matrix, which makes possible the use of standard linear model theory and finally the approximation is computationally very simple, since the eigenvectors can be calculated directly from the penalty matrix $\mathbf{K}$. Note that in the approximation c eigenvalues are raised to 1 and $n - c$ eigenvalues are shrunk to 0. This means that the effect of some of the high degree components is shrunk to zero. For the fixed part $\mathbf{X}\boldsymbol{\beta}$ the smoothers $\mathbf{P}$ and $\mathbf{S}_\alpha$ give the same fit. In Section 4 the approximation is investigated more thoroughly using a simulation experiment.

3. Testing the spline model

Here we want to test the hypothesis of the form $H_0: \sigma_u^2 = 0$ (linear regression) against the alternative $H_A: \sigma_u^2 > 0$ (spline model). The approach utilized here is based on the spline

model $\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y}$. We assume that $\mathbf{R}$ is of the form $\mathbf{R} = \mathbf{X}_* \mathbf{D} \mathbf{X}_*' + \mathbf{I}$, and develop for this case a test whose exact distribution can be derived. When $\mathbf{R}$ is of the above special form, we have $\mathbf{P}\mathbf{X}_* = \mathbf{X}_*$ (the columns of $\mathbf{X}_*$ lies in the span of the columns of $\mathbf{P}$) and thus $(\mathbf{I} - \mathbf{P})\mathbf{R}(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \mathbf{P}$, it is easy to see that the distribution of $\sigma^{-2}S_{\min} = \sigma^{-2}\mathbf{y}'(\mathbf{I} - \mathbf{P})\mathbf{y}$ is (see also Demidenko, 2004, p. 137) $\sigma^{-2}S_{\min} \sim \chi_{n-c}^2$. Similarly, if we denote $\mathbf{P}_X$ as a projection matrix of a straight line model, $\sigma^{-2}S_{\text{OLS}} = \sigma^{-2}\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y} \sim \chi_{n-2}^2$ and $\sigma^{-2}(S_{\text{OLS}} - S_{\min}) \sim \chi_{c-2}^2$. Since $\mathbf{P}\mathbf{P}_X = \mathbf{P}_X\mathbf{P} = \mathbf{P}_X$ it is also easily seen that $(\mathbf{I} - \mathbf{P})(\mathbf{P} - \mathbf{P}_X) = \mathbf{0}$ and thus the sums of squares $S_{\min}$ and $S_{\text{OLS}} - S_{\min}$ are independent. Therefore the test statistic

$$F = \frac{(S_{\text{OLS}} - S_{\min})/(c-2)}{S_{\min}/(n-c)} \sim F(c-2, n-c).$$

Testing can now be based on the quantiles of the F-distribution. Observing a larger F-value than the quantile $F_{1-\alpha}(c-2, n-c)$ yields rejection of the null hypothesis. We emphasize that this exact distributional result holds only when $\mathbf{R}$ takes the special form $\mathbf{R} = \mathbf{X}_* \mathbf{D} \mathbf{X}_*' + \mathbf{I}$. For other forms of dependence the distributional result is only approximate. The testing method itself is also easily modified for more general spline testing situations. For example, extension to regression splines should be quite straightforward, since it is well known from the theory of linear models that estimation in regression splines involves projection matrices with same kind of properties for correlated sequences as those discussed in Section 2 (see e.g. Puntanen and Styan, 1989).

4. Numerical Results

4.1 Modeling Stem Curves

Cubic smoothing splines have important applications in forestry. One important application is to model the decrease of stem diameter as a function of stem height. Traditionally various parametric models have been tested for this purpose, but since many factors difficult to control including tree species, site type, soil, climate, genetics etc., affect to the form of the

curve, the usefulness of such models may be limited. Koskela, Nummi, Wentzel and Kivinen (2006) showed how cubic smoothing splines can be utilized in stem curve prediction, which is an important part of the process of crosscutting a stem into smaller logs. In the following example we fitted a cubic smoothing spline for 59 pine stem diameter measurements of the first stem of the data set. This is part of the data introduced by Nummi and Möttönen (2004), where stem diameter measurements of 25 pine trees are taken as a function of stem height starting at the height of 150 cm with 30 cm intervals.

The spline approximation based on the eigenvectors of $\mathbf{K}$ was first tested for $c = 2, 3, 4$, and 5. The corresponding values of the information criteria (BIC) were 395.1403, 369.7435, 368.1361, and 372.1688 respectively. The minimum value is obtained for $c = 4$. In Figure 1 we plotted the original data points and the corresponding spline approximation. Testing for the pine stem data gives $S_{\text{OLS}} = 2276.669$ and $S_{\text{min}} = 1253.509$. This gives

$$F = \frac{1023.160/2}{1253.509/55} = 23.244651.$$

If this is compared to $F_{.95}(2, 55) = 3.164993$ we conclude that the null hypothesis of the linear regression model for pine stem data is clearly rejected if tested against the spline alternative.

[Figure 1 about here.]

4.2 Quality of the Spline Approximation

Since the approximation $\mathbf{P}\mathbf{y}$ of $\mathbf{S}_\alpha\mathbf{y}$ has an important role in testing here, we want to ascertain that the approximation works reasonably well to be implemented in practice. Therefore we conducted a simulation study where observations were generated from the known spline model with a small amount of added random noise. Suppose that by using the mixed model presentation of the spline solution we generated N independent normally distributed random vectors from the mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_*\mathbf{u}_* + \boldsymbol{\varepsilon}$ with

$$\text{Var} \begin{pmatrix} \mathbf{u}_* \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{R} \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \mathbf{X} \mathbf{D} \mathbf{X}' + \mathbf{I}.$$

In the experiment we set $N = 100$, $\sigma^2 = 1$, $\mathbf{D} = \mathbf{I}_2$, $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, $\mathbf{x} = (1, 2, \dots, 50)'$ and $\boldsymbol{\beta} = (500, 1)'$ with $\mathbf{Z}^* = \nabla(\nabla' \nabla)^{-1} \boldsymbol{\Delta}^{1/2}$ (see Section 2). Now since $\alpha = \sigma^2 / \sigma_u^2$ with $\sigma^2 = 1$ and $\text{edf} = \text{tr}(\mathbf{S}_\alpha)$ we can solve for each value of edf the corresponding value of σ_u^2 by a one-dimensional search, for example. Note that we prefer to investigate the approximation as a function edf instead of σ_u^2 , since edf gives a clear interpretation of the model complexity in the original spline context. Random vectors can now be generated using the model above. For each generated random vector $\mathbf{y}$ the fitted values of the spline model $\hat{\mathbf{y}} = \mathbf{S}_\alpha \mathbf{y}$ and of the approximated spline model $\tilde{\mathbf{y}} = \mathbf{P} \mathbf{y}$ were compared by using $\text{RMSE}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}) = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (\hat{y}_i - \tilde{y}_i)^2}$, where $\tilde{y}_i$ and $\hat{y}_i$ are i th components of the vectors $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$, respectively. We also calculated the relative measure $\text{RMSE}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}) / \text{RMSE}(\hat{\mathbf{y}}, \mathbf{y})$, where original observations y_i were taken as a point of comparison for the spline approximation $\tilde{y}_i$.

In the simulations we investigated two different settings. First, both α and c were estimated for each edf used in the study. The smoothing parameter α was estimated by generalized cross-validation criteria and the number of columns c in the spline approximation was determined using the information criteria BIC. Secondly, the known values of α and c used in simulations were used to construct the smoothers. In Figure 2 the mean values of RMSE were presented for each value of edf utilized in simulation. The RMSE values seem to increase quite steadily as a function of edf . Clearly, RMSE values for the known values of α and c are slightly lower than those of the estimated α and c . For small values of $\text{edf} (\leq 20)$ when α and c are estimated the relative RMSE is about 15-50 %. Thus also the relative RMSE is quite good for moderate values of edf . However, for very large $\text{edf} (> 35)$ the values of relative RMSE $\approx 100\%$ (note that for computational reasons we set the number of columns c utilized in the approximation ≤ 40). On the whole it seems that the approximation works reasonably

well for testing purposes discussed in this paper. In the next section we investigate the actual performance of the test proposed.

[Figure 2 about here.]

4.3 Power Studies

In this section we compare the method proposed to some existing spline tests. For this purpose we utilized the study by Liu and Wang (2004), where the performance of locally most powerful (LMP) test (Cox et al., 1988), generalized maximum likelihood (GML) test (Wahba, 1990), generalized cross validation (GCV) test (Wahba, 1990), two F-type test (Raz, 1990) and the SKL test (Xiang and Wahba, 1995) were investigated. In this study 100 observations were generated from three functions:

$$f(t) = 1 + t + at^2, \quad (4)$$

$$f(t) = 1 + t + 3a(t - 0.5)^3, \quad (5)$$

$$f(t) = 1 + t + \sqrt{2} \times a \times \cos(6\pi t), \quad (6)$$

with design points $t_i = (i - 1)/99$, $i = 1, \dots, 100$ and independent random errors $\varepsilon_i \sim N(0, 0.2^2)$ for the first two models and $\varepsilon_i \sim N(0, 1)$ for the last one. For (4) the coefficient a takes one of the five values: 0, 0.2, 0.5, 0.7, or 1. For (5) a takes one of the four values: 0, 0.5, 1, or 1.3 and for (6) a takes the values 0, 0.3, 0.5, or 1. For all the test statistics in the study by Liu and Wang (2004) the distribution under the null hypothesis is complicated. The null distributions were generated by the Monte Carlo method with a simulation sample size of 40 000. For the proposed F-test the significance levels were calculated from the F-distribution. The test was constructed from the c (≥ 4) first eigenvectors of $\mathbf{S}_\alpha$ computed from $\mathbf{K}$, where c was selected according to the minimum value of the information criterion BIC. The power was estimated with significance level 0.05 by counting the percentage of rejections in the 1000 repetitions.

[Table 1 about here.]

The results of the comparisons with the proposed test are summarized in Table 1. For the model (4) the proposed exact F-test performed almost equally well as the best LMP and GML tests. For the model (5) the proposed exact F-test performed clearly better than the other tests in the study. Interestingly, the LMP test, which was superior for the model (4), did not perform well for (5). The second best tests in this case were GML and GCV tests. For the high degree model (6) the exact F-test performed equally well as the GCV test, which was the best test for the high degree model in the study by Liu and Wang (2004). From the point of view of the general performance for low and high degree models the GCV test and exact F-test were the best alternatives. Of these two the exact F-test performed slightly better and is also a computationally more appealing alternative.

We also conducted a small simulation experiment to investigate the robustness of the exact F-test. Some of the models tested in Liu and Wang (2004) were considered. According to our results there is some indication that the proposed exact F-test may be quite robust to the violation of the normality assumption, but not to the violation of the dependence assumption other than conditions discussed in Section 2. These initial observations seem quite apparent and are also in line with findings of Liu and Wang (2004). However, a more comprehensive study with comparable settings is needed to confirm these observations.

5. Concluding Remarks

The most common approaches in spline testing yield a test statistic whose exact distribution is unknown. Our approach has the advantage that the exact distribution can be used, rather than an approximation, and the results also apply to certain kinds of correlated sequences. The method performed very well in comparison to other tests and is also computationally very simple.

ACKNOWLEDGEMENTS

The study was supported by the project “Nonparametric and Robust Methods for Mixed Models” (Project 111447) funded by the Academy of Finland and the project “Modeling of Mean-covariance Structures for Growth Data Using Smoothing Splines” (project R102706) funded by the Royal Society of the UK. The authors are very grateful to referees, the Associate Editor and the Co-editor for their kind advice that led to considerable improvements in the paper.

REFERENCES

- Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, Series B* **55**, 549–557.
- Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association* **93**, 961–994.
- Cleveland, W. S. and Devlin, S. J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 597–610.
- Cox, D., Koh, E., Wahba, G., and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Annals of Statistics* **16**, 113–119.
- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New Jersey: Wiley.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
- Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B*, **58**, 379–396.

- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.
- Koskela, L., Nummi, T., Wentzel, S., and Kivinen, V. (2006). On the analysis of cubic smoothing spline-based stem curve prediction for forest harvesters. *Canadian Journal of Forest Research* **36**, 2909–2920.
- Liu, A. and Wang, Y. (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation* **74**, 581–597.
- Nummi, T. and Koskela, L. (2007). Analysis of growth curve data using cubic smoothing splines. *Journal of Applied Statistics* **35**, 1–11.
- Nummi, T. and Möttönen, J. (2004). Estimation and prediction for low-degree polynomial models under measurement errors with an application to forest harvester. *Journal of the Royal Statistical Society, Series C* **53**, 495–505.
- Puntanen, S. and Styan, G.P.H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator (with discussion). *The American Statistician*, **43**, 153–164.
- Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *Journal of the American Statistical Association* **85**, 132–138.
- Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Journal of the Royal Statistical Society, Series C* **48**, 269–311.
- Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.
- Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal*

Statistical Society, Series B **60**, 233–249.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, **65**, 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall.

Xiang, D. and Wahba, G. (1995). Testing the generalized linear model null hypothesis versus ‘smooth’ alternatives. Technical report, Dept. Statistics, Univ. Wisconsin–Madison.

Zhang, D., Lin, X., and Sowers, M. (1998). Semiparametric stochastic mixed effect models for longitudinal data. *Journal of the American Statistical Association*, **93**, 710–719.

Received ?? 2010. Revised ?? 2010. Accepted ?? 2010.

Figure 1. Stem curve measurements (dots) and the spline approximation (solid curve).

Figure 2. Mean RMSE and mean relative RMSE as a function of edf. Black dot correspond to estimated α and c and white dot correspond to the known α and c .

Table 1
Proportion of the rejections in 1000 repetitions under the models (4), (5), and (6).

model	a	LMP	GML	GCV	F1	F2	SKL	Exact F
(4)	0	0.055	0.051	0.051	0.052	0.049	0.049	0.052
	0.2	0.103	0.093	0.083	0.102	0.1	0.047	0.118
	0.5	0.483	0.454	0.325	0.441	0.436	0.104	0.400
	0.7	0.759	0.734	0.582	0.702	0.7	0.245	0.671
	1	0.978	0.969	0.9	0.931	0.934	0.554	0.914
(5)	0	0.049	0.048	0.048	0.048	0.044	0.041	0.057
	0.5	0.070	0.165	0.161	0.118	0.155	0.08	0.261
	1	0.096	0.538	0.553	0.425	0.507	0.261	0.718
	1.3	0.136	0.826	0.840	0.720	0.806	0.525	0.915
(6)	0	0.053	0.054	0.055	0.052	0.053	0.045	0.062
	0.5	0.054	0.094	0.362	0.076	0.161	0.327	0.352
	1	0.058	0.229	0.872	0.483	0.699	0.849	0.891
	1.3	0.047	0.985	1.000	1.000	1.000	1.000	1.000