Hypothesis Testing in Smoothing Spline Models

Anna Liu and Yuedong Wang*

October 10, 2002

Abstract

This article provides a unified and comparative review of some existing test methods for the hypothesis of a parametric regression function using smoothing spline models. Some tests such as the locally most powerful (LMP) test by Cox, Koh, Wahba and Yandell (1988), the generalized maximum likelihood ratio (GML) test and the generalized cross validation (GCV) test by Wahba (1990) were developed from the corresponding Bayesian models. Their frequentist properties have not been studied. We conduct simulations to evaluate and compare finite sample performances. Simulation results show that the performances of these tests depend on the shape of the true function. The LMP and GML tests are more powerful for low frequency functions while the GCV test is more powerful for high frequency functions. For all test statistics, distributions under the null hypothesis are complicated. Computationally intensive Monte Carlo methods can be used to calculate null distributions. We also propose approximations to these null distributions and evaluate their performances by simulations.

Key words and phrases: Bayesian models for smoothing splines, connections between linear mixed effects models and smoothing splines, GCV test, GML test, F-test, LMP test, SKL test.

**Anna Liu is a graduate student in the Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106 (e-mail: annaliu@pstat.ucsb.edu). Yuedong Wang is an associate professor in the same department (e-mail: yuedong@pstat.ucsb.edu).* *This research was supported by NIH Grants R01 GM58533.*

1 Introduction

As a popular nonparametric regression method, spline smoothing has attracted a great deal of attention. Most research in the literature concentrates on estimation, while inference, especially hypothesis testing, has received less attention. Several test procedures were developed only for simple hypotheses of simple spline models. Their properties and performances are not well understood, one of the reasons that they are seldomly used in practice. The aim of this paper is to provide a unified and comparative review of the existing tests in a hope to promote further research, software development and application.

Consider the univariate nonparametric regression model

$$y_i = f(t_i) + \epsilon_i, \qquad i = 1, \cdots, n, \qquad 0 \leq t_i \leq 1, \tag{1}$$

where y_i is the response, ϵ_i is the random error and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. f is assumed to be in an infinite dimensional model space to be specified later.

One of the most useful applications of the nonparametric regression models is to check or suggest a parametric model. Parametric models, especially linear models, are preferred in practice because of their simplicity and interpretability. Diagnostic investigations of the departures from these parametric assumptions is necessary to avoid misleading results. If some specific alternative form is suspected, a simple lack of fit test can be performed. But this kind of test would not perform well for other departures from the parametric model, especially those orthogonal to the suspected alternative. For example, to detect departure from a linear model, one may consider a quadratic polynomial as the alternative. Then higher order departure may be missed. Tests performing well for general departures in a large model space are desired.

Most 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 1988) and smoothing spline. In this paper we focus on the tests based on smoothing spline models. The connection between smoothing spline models and Bayesian models (or mixed effects models) simplifies certain hypothesis tests. Also, the general form of smoothing spline models allows us to consider many different situations in a unified fashion.

Cox et al. (1988) showed that for the hypothesis of f being a polynomial of degree m ($m \geq 0$) versus f being smooth, there is no uniformly most powerful (UMP) test. Thus they proposed to use a locally most powerfully (LMP) test. Wahba (1990) proposed two tests based on the generalized maximum likelihood (GML) and the generalized cross validation (GCV) scores. For non-Gaussian data, Xiang and Wahba (1995) developed the symmetrized Kullback-Leibler (SKL) test based on the SKL distance between the function estimated under the null hypothesis and the function estimated under the alternative. We are going to examine the performance of these tests for Gaussian data.

Raz (1990) developed a permutation test for the hypothesis of independence between the response and the covariates without assuming any particular error distribution. Two generalized F tests were mentioned in Raz (1990) in the context of general nonparametric regression. However, no discussions about their performances were given.

In section 2, a brief introduction to smoothing splines is given. In section 3, we review some existing tests and develop approximations to null distributions. We evaluate and compare these tests and approximations in section 4. Section 5 concludes with some remarks and potential research topics.

2 Smoothing spline models

In this section we briefly review smoothing spline models, their corresponding Bayesian models and connections with linear mixed effects models. For simplicity, we limit our discussions to polynomial splines on $[0, 1]$. All tests in this paper can be written in terms of general spline models on arbitrary domains (Wahba 1990). Thus these tests can be used to test more complicated hypothesis under general spline models.

In model (1), assume that $f \in W_m$, where

$$W_m = \left\{ g | g, \cdots, g^{(m-1)} \text{ are absolutely continuous, } g^{(m)} \in \mathcal{L}_2[0, 1] \right\}.$$

The smoothing spline estimate of f, $\hat{f}_\lambda$, is the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du, \qquad (2)$$

where λ is a smoothing parameter which controls the trade-off between the goodness-of-fit and smoothness of the estimate.

Let $\mathbf{y} = (y_1, \cdots, y_n)'$, $\phi_\nu(t) = t^{\nu-1}/(\nu-1)!$, $\nu = 1, \cdots, m$, and $R^1(s, t) = \int_0^{min(s,t)} (s-u)^{m-1}(t-u)^{m-1} du/((m-1)!)^2$. Denote $T_{n \times m} = \{\phi_\nu(t_i)\}_{i=1}^{n}{}_{\nu=1}^{m}$ and

$\Sigma_{n\times n} = \{R^1(t_i, t_j)\}_{i=1}^{n}{}_{j=1}^{n}$. Kimeldorf and Wahba (1971) showed that the solution to (2) has the form

$$\hat{f}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i R^1(t, t_i) \ ,$$

where $\mathbf{c} = (c_1, \cdots, c_n)'$ and $\mathbf{d} = (d_1, \cdots, d_m)'$ are solutions to

$$\begin{pmatrix} T & \Sigma + n\lambda I \\ \mathbf{0} & T' \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}. \tag{3}$$

The system (3) is definite when T is of full column rank, which we assume to be true in this paper. Thus $\hat{\mathbf{f}}_\lambda = (\hat{f}_\lambda(t_1), \cdots, \hat{f}_\lambda(t_n))' = T\mathbf{d} + \Sigma\mathbf{c}$ is always unique. Let

$$T = (Q_1 \ \ Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix}$$

be the QR decomposition of T. One may check that $\hat{\mathbf{f}}_\lambda$ is a linear function of $\mathbf{y}$: $\hat{\mathbf{f}}_\lambda = A(\lambda)\mathbf{y}$, where $A(\lambda)$ is the "hat" matrix. It can be verified that

$$A(\lambda) = I - n\lambda Q_2(Q_2'(\Sigma + n\lambda)Q_2)^{-1}Q_2'. \tag{4}$$

Note that $A(\lambda)$ is symmetric but usually not idempotent.

The smoothing spline estimate can be obtained from the Bayesian point of view. Assume a prior for f as

$$F(t) = \sum_{\nu=1}^{m} \theta_\nu \phi_\nu(t) + b^{1/2} X(t),$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_\nu)' \sim N(0, aI)$, a and b are positive constants, and $X(t)$ is a zero mean Gaussian stochastic process independent of $\boldsymbol{\theta}$ with covariance $EX(s)X(t) = R^1(s, t)$.

Consider

$$y_i = F(t_i) + \epsilon_i, \qquad i = 1, \cdots, n, \qquad t_i \in [0, 1], \tag{5}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 I)$ and is independent of F. Wahba (1990) showed that with $\lambda = \sigma^2/nb$,

$$\lim_{a\to\infty} E(F(t)|\mathbf{y}) = \hat{f}_\lambda(t).$$

With $a \to \infty$, diffuse priors are assumed for the coefficients of the polynomials of degree less than m.

Smoothing spline models can also be connected to certain linear mixed effects models (LMM). Consider the following LMM

$$\mathbf{y} = T\mathbf{d} + \mathbf{u} + \boldsymbol{\epsilon}, \tag{6}$$

where $\mathbf{d}$ are the fixed effects, $\mathbf{u}$ are random effects and $\mathbf{u} \sim N(0, b\Sigma)$, $\boldsymbol{\epsilon}$ are random errors and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, and $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are independent. Wang (1998a) showed that the smoothing spline estimate evaluated at the design points, $\hat{\mathbf{f}}_\lambda$, is the same as the best linear unbiased prediction (BLUP) estimate in (6).

3 Existing test methods

Let $\mathcal{H}_0 = span\{\phi_\nu, \nu = 1, \cdots, m\}$. Often we are interested in testing the hypothesis that f is a polynomial of degree $m-1$ or less

$$\mathrm{H}_0 : f \in \mathcal{H}_0, \qquad \mathrm{H}_1 : f \in W_m \text{ and } f \notin \mathcal{H}_0. \tag{7}$$

It is easy to see that $\lambda = \infty$ in (2), or equivalently $b = 0$ in the corresponding Bayesian and mixed effects models, leads to $f \in \mathcal{H}_0$. Thus the hypothesis (7) can be re-expressed as

$$\mathrm{H}_0 : \lambda = \infty, \qquad \mathrm{H}_1 : \lambda < \infty, \tag{8}$$

or

$$\mathrm{H}_0 : b = 0, \qquad \mathrm{H}_1 : b > 0. \tag{9}$$

Notice that $\mathbf{y} \sim N(0, aTT' + b\Sigma + \sigma^2 I)$ under the Bayesian model (5) and $\mathbf{y} \sim N(T\mathbf{d}, b\Sigma + \sigma^2 I)$ under the mixed effects model (6). Let $\mathbf{w} = {Q_2}'\mathbf{y}$, then $\mathbf{w} \sim N(0, bQ_2'\Sigma Q_2 + \sigma^2 I)$ under both the mixed effects and Bayesian models. It is clear that the transformation ${Q_2}'\mathbf{y}$ eliminates contribution from the model under the null hypothesis. Thus $\mathbf{w}$ reflects signals, if any, from $W_m \ominus \mathcal{H}_0$.

Let the spectral decomposition of $Q_2'\Sigma Q_2$ be UDU', where $D = diag(\lambda_{\nu n}, \nu = 1, \cdots, n-m)$, and $\lambda_{\nu n}$'s are the eigenvalues of $Q_2'\Sigma Q_2$ ordered so that $\lambda_{1n} \geq \lambda_{2n} \geq \cdots \geq \lambda_{n-m\ n}$. Let $\mathbf{z} = U'\mathbf{w}$, then

$$\mathbf{z} \sim N(0, bD + \sigma^2 I). \tag{10}$$

Note that z_ν, the νth component of $\mathbf{z}$, is the projection of $\mathbf{w}$ in the direction of the νth column (eigenvector) of U ($Q_2'\Sigma Q_2$).

3.1 LMP tests

Cox et al. (1988) showed that the UMP test does not exist for hypothesis (9) under model (5). When σ^2 is known, they proposed a LMP test which rejects the null hypothesis for large values of

$$t_{LMP} = \sum_{\nu=1}^{n-m} \lambda_{\nu n} z_\nu^2. \qquad (11)$$

When σ^2 is unknown they proposed an approximate LMP which rejects the null hypothesis for large values of

$$t_{appLMP} = \sum_{\nu=1}^{n-m} \lambda_{\nu n} z_\nu^2 / \sum_{\nu=1}^{n-m} z_\nu^2. \qquad (12)$$

Let $l(b, \sigma^2|\mathbf{z})$ denote the likelihood of b and σ^2 given $\mathbf{z}$. Then

$$l(b, \sigma^2|\mathbf{z}) = -\frac{n-m}{2}\log(2\pi) - \frac{1}{2}\sum_{\nu=1}^{n-m}\log(b\lambda_{\nu n} + \sigma^2) - \frac{1}{2}\sum_{\nu=1}^{n-m}\frac{z_\nu^2}{b\lambda_{\nu n} + \sigma^2}.$$

Note that $\mathbf{z}$ is based on an orthogonal contrast of the original observations which eliminates the fixed effects. Thus $l(b, \sigma^2|\mathbf{z})$ is the so called restricted likelihood in the mixed effects literature.

It is not difficult to check that the LMP test is equivalent to the score test (Cox and Hinkley 1974) defined by

$$t_{score} = U_b(0, \sigma^2)/\sqrt{I_{bb}(0, \sigma^2)}, \qquad (13)$$

where $U_b(b, \sigma^2)$ is the efficient score defined as $\partial l(b, \sigma^2|\mathbf{z})/\partial b$ and $I_{bb}(b, \sigma^2)$ is the Fisher information of b. When it is unknown, σ^2 can be replaced by the MLE under the null hypothesis (9), $\hat{\sigma}^2 = \sum_{\nu=1}^{n-m} z_\nu^2/(n-m)$, which leads to the approximate LMP test.

The test statistic t_{appLMP} does not follow a simple distribution under H_0. It is straightforward to simulate the null distribution (Wahba 1990). We can also approximate its null distribution as follows. First, the numerator can be approximated using Satterthwaite method by $d_2\chi^2_{d_1}$ with $d_2 = \sum_{\nu=1}^{n-m}\lambda_{\nu n}^2/\sum_{\nu=1}^{n-m}\lambda_{\nu n}$ and $d_1 = (\sum_{\nu=1}^{n-m}\lambda_{\nu n})^2/\sum_{\nu=1}^{n-m}\lambda_{\nu n}^2$. The denominator follows the χ^2_{n-m} distribution. Then

$$F_{appLMP} = \frac{\sum_{\nu=1}^{n-m}\lambda_{\nu n} z_\nu^2/(d_1 d_2)}{\sum_{\nu=1}^{n-m} z_\nu^2/(n-m)} = \frac{n-m}{\sum_{\nu=1}^{n-m}\lambda_{\nu n}} t_{appLMP} \qquad (14)$$

is approximated by an F distribution with degrees of freedom d_1 and $n-m$. This approximation is compared with the simulated null distribution in Section 5. The p-value, $P(t_{appLMP} > t^{obs}_{appLMP}) = P(\sum_{\nu=1}^{n-m}(\lambda_{\nu n} - t^{obs}_{appLMP})z_\nu^2 > 0)$, can be calculated numerically using the algorithm in Davies (1980). We find that this numerical method is very fast and agrees with the Monte Carlo method.

3.2 GML test

Since $b = \sigma^2/n\lambda$, the log likelihood from $\mathbf{z}$ can be re-expressed as

$$l(\lambda, b|\mathbf{z}) = -\frac{1}{2}(n-m)\log b - \frac{1}{2}\sum_{\nu=1}^{n-m}\log(\lambda_{\nu n} + n\lambda) - \frac{1}{2b}\sum_{\nu=1}^{n-m}\frac{z_\nu^2}{\lambda_{\nu n} + n\lambda} + C,$$

where C is a constant.

For fixed λ, maximizing the log likelihood with respect to b, we have

$$\hat{b}_\lambda = \frac{1}{n-m}\sum_{\nu=1}^{n-m}\frac{z_\nu^2}{\lambda_{\nu n} + n\lambda}.$$

Then the profiled likelihood of λ is

$$L(\lambda|\mathbf{z}) = exp(l(\lambda, \hat{b}_\lambda|\mathbf{z})) = C_1\left(\frac{\sum_{\nu=1}^{n-m} z_\nu^2/(\lambda_{\nu n} + n\lambda)}{\Pi_{\nu=1}^{n-m}(\lambda_{\nu n} + n\lambda)^{-\frac{1}{n-m}}}\right)^{-\frac{n-m}{2}}, \qquad (15)$$

where C_1 is a constant.

The generalized maximum likelihood (GML) estimate of λ, $\hat{\lambda}_{GML}$, is the maximizer of (15). Wahba (1990) defined the GML test statistic for the hypothesis (8) as

$$t_{GML} = \left[\frac{L(\hat{\lambda}_{GML}|\mathbf{z})}{L(\infty|\mathbf{z})}\right]^{-\frac{2}{n-m}} = \frac{\sum_{\nu=1}^{n-m} z_\nu^2/(\lambda_{\nu n} + n\hat{\lambda}_{GML})}{\Pi_{\nu=1}^{n-m}(\lambda_{\nu n} + n\hat{\lambda}_{GML})^{-\frac{1}{n-m}}}\frac{1}{\sum_{\nu=1}^{n-m} z_\nu^2}. \qquad (16)$$

The null hypothesis is rejected when t_{GML} is too small.

It is difficult to derive the null distribution for t_{GML}. Standard theory for likelihood ratio tests does not apply because the parameter is on the boundary under the null hypothesis. The non-standard asymptotic theory developed by Self and Liang (1987), which states that $-(n-m)\log t_{GML}$ has an asymptotic null distribution of a 50:50 mixture of χ_1^2 and χ_0^2, does not apply either because of the lack of replicated observations. Crainiceanu, Ruppert and Vogelsang (2002) reported the same finding for P-spline models. In a subsequent paper, Crainiceanu and Ruppert (2002) provide the asymptotic distributions of likelihood ratio tests for linear mixed models. Monte Carlo methods are still needed to obtain quantiles of these asymptotic distributions.

The direct Monte Carlo method simulates l samples of $-(n-m)\log t_{GML}$ under the null hypothesis. Denote $-(n-m)\log t_{GML}$ based on data as x_0 and suppose that $x_0 > 0$. Then the true p-value is

$$p = P(-(n-m)\log t_{GML} > x_0|\mathrm{H}_0).$$

We generate l samples of $\mathbf{z}$ from $N(0, I)$ (without loss of generality, we set $\sigma^2 = 1$), calculate $\hat{\lambda}_{GML}$ for each sample, and construct t_{GML} for each sample. Let $x_1, \cdots, x_l$ denote the l samples of $-(n-m)\log t_{GML}$. Then p is estimated by

$$\hat{p} = \frac{1}{l}\sum_{i=1}^{l} I(x_i > x_0),$$

where $I(\cdot)$ is the indicator function. Then $E\hat{p} = p$ and $Var(\hat{p}) = p(1-p)/l$. This approach usually requires a very large l. For example, to have margin of error $2\sqrt{Var(\hat{p})}$ bounded by 0.005, l has to be at least 40000. Note that $\hat{\lambda}_{GML}$ is computed for each sample. Therefore, this approach is computationally intensive.

Our simulation results suggest that the null distribution of $-(n-m)\log t_{GML}$ can be well approximated by a mixture of χ^2_1 and χ^2_0, denoted by $r\chi^2_0 + (1-r)\chi^2_1$. However, the ratio r is not fixed. It depends on the order m, sample size n and the design points t_i's. Thus we propose an alternative method that estimates the ratio r first and then calculates the p-value based on the mixture of χ^2_1 and χ^2_0 with the estimated r. The motivation behind this approach is that a relatively small sample size k is required to estimate r.

We now compare sample sizes required by these two approaches. For the alternative approach, let $x'_1, \cdots, x'_k$ be k random samples of $-(n-m)\log t_{GML}$ under the null hypothesis. We estimate r by

$$\hat{r} = \frac{1}{k}\sum_{i=1}^{k} I(x'_i = 0).$$

Then $E\hat{r} = r$ and $Var(\hat{r}) = r(1-r)/k$. The p-value is estimated by

$$\tilde{p} = (1-\hat{r})P(\chi^2_1 > x_0).$$

Assuming the null distribution of non-zero $-(n-m)\log t_{GML}$ is exactly χ^2_1, we have $E\tilde{p} = p$ and $Var(\tilde{p}) = rp^2/(k(1-r))$.

For $Var(\hat{p}) = Var(\tilde{p})$, we need $k = rpl/((1-r)(1-p))$. Based on our simulations with $m = 2$, $n = 100$ and a uniform design in $[0,1]$, r is usually around 0.7. It is easy to check that for $p = 0.05$ and $r = 0.7$, we have $k \approx 0.12l$. Thus about $k = 5000$ samples are needed for the alternative method if $l = 40000$.

Simulation results in Section 5 indicate approximations based on the alternative approach are accurate when sample size is large.

3.3 F-type tests

For the hypothesis (7) under model (1), the usual F test statistic will not follow an F distribution because the hat matrix $A(\lambda)$ is not idempotent. Two F-type tests were mentioned in Raz (1990) for the following hypothesis

$$\mathrm{H}_0 : f = \text{constant}, \qquad \mathrm{H}_1 : f \neq \text{constant} \tag{17}$$

in the context of general nonparametric regressions. They were used to derive the permutation test and their performances were not investigated. In this section, we first extend these two F-type statistics for our hypothesis (7). Then we compare them with the generalized cross validation (GCV) test proposed in Wahba (1990) and the SKL test proposed by Xiang and Wahba (1995).

Let $\hat{f}_0$ be the maximum likelihood estimate of the regression function under the null model. Then $\hat{\mathbf{f}}_0 = H\mathbf{y}$, where $\hat{\mathbf{f}}_0 = (\hat{f}_0(t_1), \cdots, \hat{f}_0(t_n))'$ and $H = T(T'T)^{-1}T'$. Note that H is an idempotent hat matrix and $A(\lambda)H = HA(\lambda) = H$.

Define

$$\begin{aligned} S_1 &= \textstyle\sum_{i=1}^n (\hat{f}_\lambda(t_i) - \hat{f}_0(t_i))^2, \\ S_2 &= \textstyle\sum_{i=1}^n (y_i - \hat{f}_\lambda(t_i))^2, \\ S_3 &= \textstyle\sum_{i=1}^n (y_i - \hat{f}_0(t_i))^2, \end{aligned} \tag{18}$$

where S_1 measures the difference between $\hat{f}_0$ and $\hat{f}_\lambda$, S_2 is the residual sum of squares under H_1 and S_3 is the residual sum of squares under H_0.

In terms of the hat matrices, (18) can be re-expressed as $S_1 = \mathbf{y}'(A(\lambda) - H)^2\mathbf{y}$, $S_2 = \mathbf{y}'(I - A(\lambda))^2\mathbf{y}$ and $S_3 = \mathbf{y}'(I - H)\mathbf{y}$. In terms of $\mathbf{z}$, we have

$$\begin{aligned} S_1 &= \textstyle\sum_{\nu=1}^{n-m} (\frac{\lambda_{\nu n}/n\lambda}{1+\lambda_{\nu n}/n\lambda})^2 z_\nu^2 = S_2 + S_3 - 2\sum_{\nu=1}^{n-m} \frac{z_\nu^2}{1+\lambda_{\nu n}/n\lambda}, \\ S_2 &= \textstyle\sum_{\nu=1}^{n-m} \frac{z_\nu^2}{(1+\lambda_{\nu n}/n\lambda)^2}, \\ S_3 &= \textstyle\sum_{\nu=1}^{n-m} z_\nu^2. \end{aligned} \tag{19}$$

Contrary to the parametric case, the equality $S_1 + S_2 = S_3$ usually does not hold. Similar to Raz (1990), we consider two generalizations of the standard F test statistic

$$F_1 = (n - g_1)S_1/((g_1 - m)(S_3 - S_1)) \text{ with } g_1 = tr(A^2(\lambda)), \tag{20}$$

and

$$F_2 = g_1^*(S_3 - S_2)/((n - g_1^* - m)S_2) \text{ with } g_1^* = tr((I - A(\lambda))^2). \tag{21}$$

When λ is fixed, the permutation test statistic in Raz (1990), $(n-1)S_1/S_3$, is equivalent to F_1. Cantoni and Hastie (2000) considered a different hypothesis where λ (b) was fixed under the alternative. Their F test statistic is equivalent to F_2 when σ^2 is estimated under their alternative hypothesis. They used numerical methods for linear combinations of χ^2 variables to compute p-values (Davies 1980). This method can not be used here except for the LMP test because the smoothing parameters are not fixed under the alternative hypothesis. In Section 5, we are going to investigate the performances of F_1 and F_2 tests with a data-based choice of λ.

When λ is estimated from data, the null distributions of F_1 and F_2, are rather complicated and no approximation is available. When λ is fixed, we can use the Satterthwaite method to approximate the numerator and the denominator of the test statistics and then use F distributions to approximate the null distributions. We note that $S_3 \geq S_2$ and $S_3 > S_1$, so F_1 and F_2 are guaranteed to be nonnegative. Simple calculation shows that F_1 can be approximated by an F distribution with degrees of freedom μ_1^2/μ_2 and θ_1^2/θ_2, where $\mu_1 = tr((A(\lambda) - H)^2) = g_1 - m$ and $\mu_2 = tr((A(\lambda) - H)^4)$, $\theta_1 = tr((I - H) - (A(\lambda) - H)^2) = n - g_1$ and $\theta_2 = tr(((I - H) - (A(\lambda) - H)^2)^2) = tr((I - A^2(\lambda))^2)$. F_2 can be approximated by an F distribution with degrees of freedom ν_1^2/ν_2 and δ_1^2/δ_2, where $\nu_1 = tr((I - H) - (I - A(\lambda))^2) = n - g_1^* - m$ and $\nu_2 = tr(((I - H) - (I - A(\lambda))^2)^2)$, $\delta_1 = tr((I - A(\lambda))^2) = g_1^*$ and $\delta_2 = tr((I - A(\lambda))^4)$. λ is fixed in above approximations. The performances of the F approximations are examined in Section 5 for several values of λ. There is no practical guidance on how to select λ. Eubank and Spiegelman (1990) considered tests based on cubic smoothing splines with fixed λ. They pointed out that the powers of their test are relatively insensitive to the choice of λ.

3.3.1 The SKL test

For non-Gaussian data, Xiang and Wahba (1995) proposed the SKL test based on the symmetrized Kullback-Leibler distance between $\hat{f}_\lambda$ and $\hat{f}_0$:

$$t_{SKL} = \frac{1}{n}[E_{\hat{f}_0}(log(\frac{\hat{f}_0}{\hat{f}_\lambda})) + E_{\hat{f}_\lambda}(log(\frac{\hat{f}_\lambda}{\hat{f}_0}))].$$

For Gaussian data, it reduces to

$$t_{SKL} = \frac{1}{n\sigma^2}\|\hat{f}_\lambda - \hat{f}_0\|^2 = \frac{1}{n\sigma^2}S_1. \tag{22}$$

When σ^2 is estimated by $S_3/(n-m)$, $t_{SKL} = (n-m)S_1/nS_3$. Thus t_{SKL} is equivalent to F_1 for fixed λ. The performance of the SKL test is compared with F-type tests and other tests in Section 5 with λ estimated from data.

3.3.2 GCV test

GCV test is based on the following GCV score (Wahba 1990)

$$V(\lambda) = n\|(I - A(\lambda))\mathbf{y}\|^2/(tr(I - A(\lambda)))^2.$$

The GCV estimate of λ, $\hat{\lambda}_{GCV}$, is the minimizer of $V(\lambda)$.

Wahba (1990) defined the GCV test statistic as

$$t_{GCV} = \frac{V(\hat{\lambda}_{GCV})}{V(\infty)} = (n-m)^2 \frac{\sum_{\nu=1}^{n-m} z_\nu^2/(1+\lambda_{\nu n}/n\hat{\lambda}_{GCV})^2}{[\sum_{\nu=1}^{n-m} 1/(1+\lambda_{\nu n}/n\hat{\lambda}_{GCV})]^2} \frac{1}{\sum_{\nu=1}^{n-m} z_\nu^2}.$$

H_0 is rejected when t_{GCV} is too small. It is easily seen that t_{GCV} is equivalent to F_2 if the smoothing parameter is fixed instead of being estimated from the GCV score. Again, the performance of the GCV test is compared with F-type tests and the other tests in Section 5 with λ estimated from data.

4 An overall comparison

All tests except F_1 and F_2 can be written in the form $\sum_{\nu=1}^{n-m} a_\nu z_\nu^2 / \sum_{\nu=1}^{n-m} z_\nu^2$, where coefficients for the approximate LMP, GML, GCV and SKL tests are $a_\nu^{LMP} = \lambda_{\nu n}$, $a_\nu^{GML} = \Pi_{\nu=1}^{n-m}(\lambda_{\nu n} + n\hat{\lambda}_{GML})^{\frac{1}{n-m}}/(\lambda_{\nu n} + n\hat{\lambda}_{GML})$, $a_\nu^{GCV} = (n-m)^2/[(1+\lambda_{\nu n}/n\hat{\lambda}_{GCV})^2 \sum_{\nu=1}^{n-m} 1/(1+\lambda_{\nu n}/n\hat{\lambda}_{GCV})^2]$ and $a_\nu^{SKL} = (n-m)(\lambda_{\nu n}/n\hat{\lambda})^2/[n(1+\lambda_{\nu n}/n\hat{\lambda})^2]$ with $\hat{\lambda}$ being an estimate of λ.

Note that a_ν^{LMP} and a_ν^{SKL} are decreasing while a_ν^{GML} and a_ν^{GCV} are increasing. This is because the rejection regions for the approximate LMP and SKL tests are on the right hand side while the rejection regions of the GML and GCV tests are on the left hand side. Notice that $t_{GML} \le 1$ and $t_{GCV} \le 1$. For comparison, we use the equivalent test statistics $1 - t_{GML}$ and $1 - t_{GCV}$ as the GML and GCV test statistics in this section.

The differences between the approximate LMP, GML, GCV and SKL lie in the differences between weights. The weights depend on the smoothing kernel matrix Σ, the design matrix T and the smoothing parameter λ except for the approximate LMP test. We now compare these weights for a cubic smoothing spline with $n = 100$ and a uniform design in [0,1]. Note that a_ν's are not directly comparable because their scales are different and the corresponding statistics have different distributions. In the following, we scale the four null test statistics so that they have the same 95% quantiles. We first generate 40000 sets of $\mathbf{z}$ under the null hypothesis. Note that all the test statistics are transformation invariant with respect to σ^2, so it is taken as 1 in the simulation. For each set of $\mathbf{z}$, the smoothing parameter $\hat{\lambda}_{GML}$ and $\hat{\lambda}_{GCV}$ are calculated and $\hat{\lambda}$ in a_ν^{SKL} is replaced by $\hat{\lambda}_{GCV}$. Based on 40000 null test statistics of t_{appLMP}, $1-t_{GML}$, $1-t_{GCV}$ and t_{SKL}, we find that t_{appLMP}, $0.464(1-t_{GML})$, $0.199(1-t_{GCV})$ and $0.083t_{SKL}$ have approximately the same 95% quantiles. Therefore we define $w_\nu^{LMP} = a_\nu^{LMP}$, $w_\nu^{GML} = 0.464(1 - a_\nu^{GML})$, $w_\nu^{GCV} = 0.199(1 - a_\nu^{GCV})$ and $w_\nu^{SKL} = 0.083a_\nu^{SKL}$. All the four tests are equivalent to $\sum_{\nu=1}^{n-m} w_\nu z_\nu^2 / \sum_{\nu=1}^{n-m} z_\nu^2$, where w_ν denotes one of the w_ν^{LMP}, w_ν^{GML}, w_ν^{SKL} and w_ν^{GCV}. Let $\bar{w}_\nu$ be the average of the 40000 realizations of w_ν. Note that $\bar{w}_\nu^{LMP} = w_\nu^{LMP}$ since w_ν^{LMP} does not depend on the smoothing parameter λ.

In Figure 1 we show the comparison of $\bar{w}_\nu$ for $\nu = 1, \cdots, 10$. Except for the SKL test, the weights for all tests decrease very quickly. The SKL test puts almost equal weights on all z_ν's. Although not shown in Figure 1, the weights of the SKL test are larger than those of other tests when $\nu \geq 15$. We observe that on average, the approximate LMP test has the largest weight on z_1. Thus the LMP is more powerful in the direction of the first column of U, which is the first eigenvector of $Q_2{}'\Sigma Q_2$. The localness is by no means defined in terms of the distance in W_m or the L_2 distance. It is easy to find two directions, such as $\sin(2\pi t)$ and $\cos(2\pi t)$, such that powers are very different even when they have the same distances to the null space. The GCV test has the largest weights on $z_\nu, 2 \leq \nu \leq 10$. Thus it is more sensitive to changes in these directions. The GML is a compromise between the LMP and the GCV tests.

Figure 2 plots the first four columns of U. If we consider the number of modes as the frequency of a function, then the columns of U represent functions with increasing frequencies. Thus we can expect that the approximate LMP test is the most powerful when the true function has frequency 1 and the GCV test is more powerful for higher frequency functions. These observations are confirmed by our simulation results in Section 5. In theory, one can construct new tests in the form

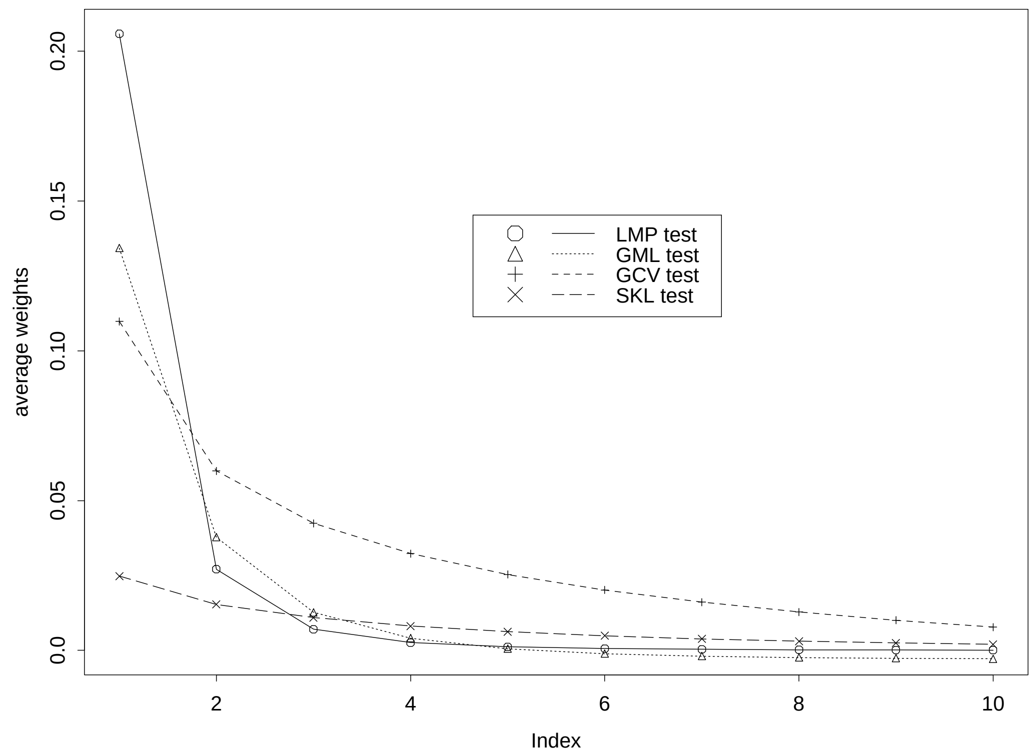


Figure 1: Weight comparisons among the approximate LMP, GML, SKL and GCV tests.

$\sum_{\nu=1}^{n-m} w_\nu z_\nu^2 / \sum_{\nu=1}^{n-m} z_\nu^2$ with weights chosen to achieve specific purposes. We have also studied weights for other smoothing spline models such as the linear spline and the periodic spline. Results remain the same.

5 Simulations

Wahba (1990) conducted a small scale simulation to compare the LMP, GML and GCV tests. Since data were generated from the stochastic Bayesian model (5), it was not clear if these results hold when data are generated from the deterministic model (1). In this section, we conduct simulations to evaluate and compare the relative powers of the LMP, GML, GCV, F_1, F_2 and SKL tests. Cubic splines (m=2) are used throughout this section.

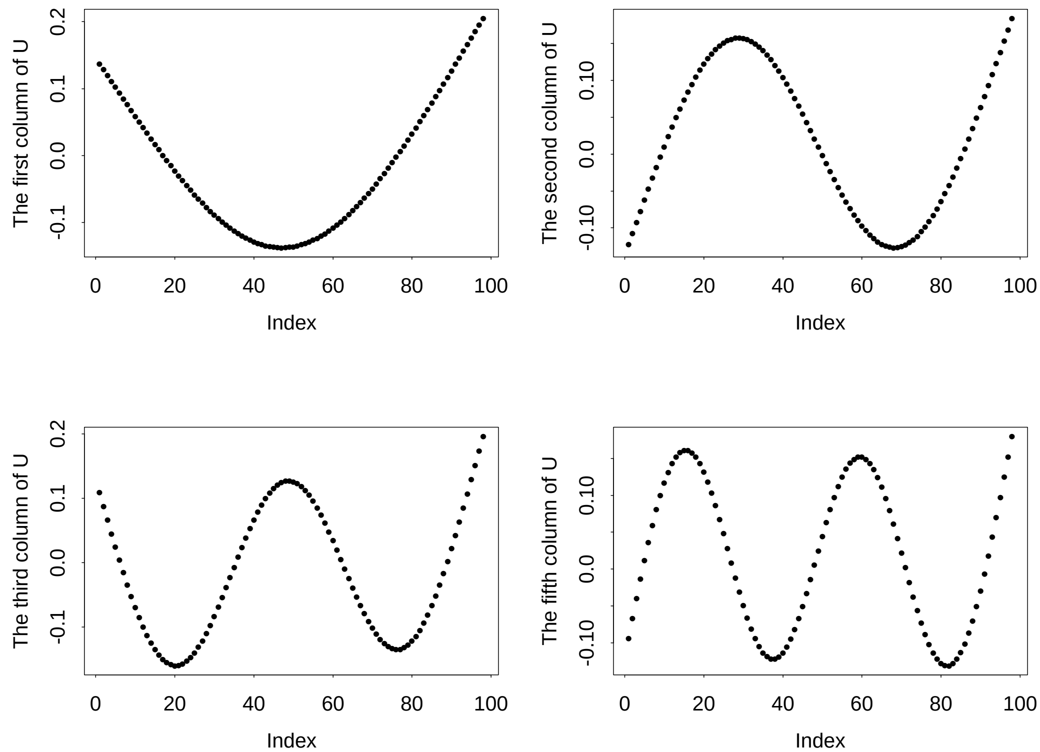


Figure 2: The first four columns of the U matrix.

5.1 Power comparisons

We conduct three simulations to compare powers of these tests. 100 observations were generated from model (1) with the following three f functions:

$$f(t) = 1 + t + at^2, \tag{23}$$

$$f(t) = 1 + t + 3a(t - 0.5)^3, \tag{24}$$

$$f(t) = 1 + t + \sqrt{2}a\cos(6\pi t). \tag{25}$$

The first function is close to the first eigenvector, the second function is close to the second eigenvector and the third function is a high frequency function. The design points are $t_i = (i-1)/99, i = 1, \cdots, 100$. We use $\epsilon_i \overset{iid}{\sim} N(0, 0.2^2)$ for the first two models and $\epsilon_i \overset{iid}{\sim} N(0,1)$ for the last one. For (23), a takes one of the following five values: 0, 0.2, 0.5, 0.7 and 1. For (24), a takes one of the following four values: 0, 0.5, 1 and 1.3. For (25), a takes one of the following four values: 0, 0.3, 0.5 and 1. For all three models, we are testing the hypothesis (7) with $m = 2$.

We repeat each setting 1000 times. The null distributions are generated by the Monte Carlo method described in Section 4 with simulation sample size 40000. The smoothing parameters in the test statistics are estimated for each simulation sample. The proportion of rejections with significance level 0.05 for each test is obtained by counting the percentage of rejection in the 1000 repetitions. Results are shown in Tables 1, 2 and 3.

	a=0	a=0.2	a=0.5	a=0.7	a=1
LMP test	0.055	0.103	0.483	0.759	0.978
GML test	0.051	0.093	0.454	0.734	0.969
GCV test	0.051	0.083	0.325	0.582	0.9
F_1 test	0.052	0.102	0.441	0.702	0.931
F_2 test	0.049	0.1	0.436	0.7	0.934
SKL test	0.049	0.047	0.104	0.245	0.554

Table 1: Proportion of rejections in 1,000 replications under model (23).

	a=0	a=0.5	a=1.0	a=1.3
LMP test	0.049	0.070	0.096	0.136
GML test	0.048	0.165	0.538	0.826
GCV test	0.048	0.161	0.553	0.840
F_1 test	0.048	0.118	0.425	0.720
F_2 test	0.044	0.155	0.507	0.806
SKL test	0.041	0.080	0.261	0.525

Table 2: Proportion of rejections in 1,000 replications under model (24).

Generally speaking, all tests hold their levels properly. As expected from discussions in Section 4, the approximate LMP test is the best under model (23) but the worst under models (24) and (25). This confirms that the approximate LMP test is the most powerful only in the direction of the first eigenvector. The GML test performs well for low frequency functions (models (23) and (24)). The lack of power under model (25) when $a = 0.3$ and $a = 0.5$ is caused by a combination of bad choices of smoothing parameters (GML method tends to oversmooth in these cases) and small weights of the GML test for higher frequency functions. The GCV

	a=0	a=0.3	a=0.5	a=1.0
LMP test	0.053	0.054	0.058	0.047
GML test	0.054	0.094	0.229	0.985
GCV test	0.055	0.362	0.872	1.000
F_1 test	0.052	0.076	0.483	1.000
F_2 test	0.053	0.161	0.699	1.000
SKL test	0.045	0.327	0.849	1.000

Table 3: Proportion of rejections in 1,000 replications under model (25).

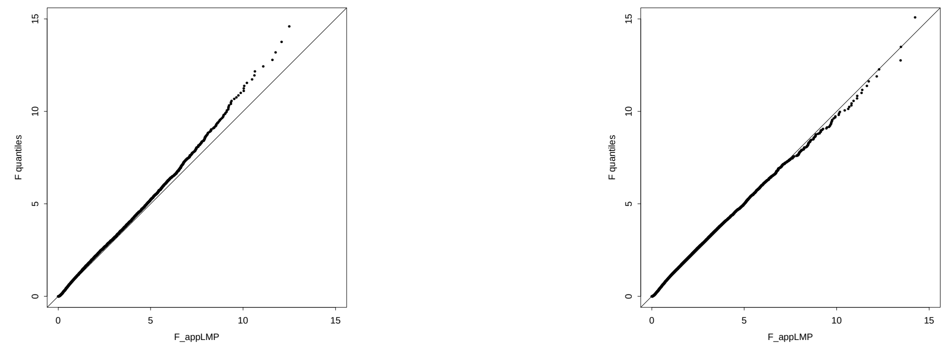
method has similar performance as the GML test under model (24) and is the best under model (25), again as expected. The F_1 and F_2 tests perform similarly as the GML test. The SKL test lacks power to detect lower frequency functions. None of these tests performs consistently well for all simulation settings. The best method to use in practice depends on the shape of the true function. To detect departure in the form of the first eigenvector of U, the approximate LMP method is recommended. To detect low-frequency departure, the GML method is recommended. To detect departure of higher frequencies, the GCV method is recommended.

Simulations are also conducted for other functions and spline models. Results are similar.

5.2 Approximations to null distributions

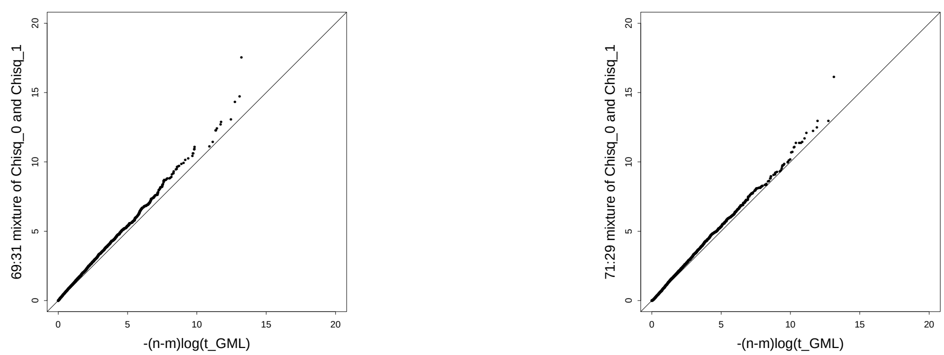
Regarding the Monte Carlo null distributions as the truth, we now investigate accuracy of various approximations proposed in this paper. We consider two sample sizes, 100 and 200, with design points evenly spaced in $[0, 1]$. Monte Carlo null distributions are generated as described in Section 4 with simulation size 40000. Figure 3 shows QQplots of the Monte Carlo null distributions of F_{appLMP} defined in (14) against their approximate F distributions with degrees of freedom d_1 and $n - m$. The approximations are good except in the tail of the distribution with sample size 100.

In Figure 4, we compare the approximation based on $r\chi_0^2 + (1 - r)\chi_1^2$ to the Monte Carlo null distribution of $-(n - m) \log t_{GML}$, where t_{GML} is defined in (16). The estimates of r, $\hat{r}$, are obtained by simulating 5000 of $-(n - m) \log t_{GML}$ under the null hypothesis and counting the proportion of zeros out of the 5000. Again, the approximation is good for sample size 200.



(a) $n = 100$ (b) $n = 200$

Figure 3: QQplot of quantiles of the F distribution with degrees of freedom d_1 and $n - m$ against quantiles of the Monte Carlo null distribution of F_{appLMP} in (14).



(a) $n = 100$, $\hat{r} = 0.69$ (b) $n = 200$, $\hat{r} = 0.71$

Figure 4: QQplot of quantiles of the $r\chi_0^2 + (1 - r)\chi_1^2$ distribution against quantiles of the Monte Carlo null distribution of $-(n - m) \log t_{GML}$.

To further assess the accuracy of these approximations, we repeat the power calculations for the two tests under model (23) in Section 5.1 using approximate null distributions. Table 4 shows the results. We see again that both approximations work reasonably well for sample size 200. For sample size 100, the levels are off. Therefore, the Monte Carlo method is recommended when the sample size is small.

We also examined the F approximations to the null distributions of the F_1 and

	size	null	a=0	a=0.2	a=0.5	a=0.7	a=1
LMP	100	simu	0.055	0.103	0.483	0.759	0.978
		approx	0.044	0.098	0.454	0.713	0.943
	200	simu	0.047	0.392	0.991	1	1
		approx	0.046	0.38	0.99	1	1
GML	100	simu	0.051	0.093	0.454	0.734	0.969
		approx	0.041	0.087	0.395	0.673	0.94
	200	simu	0.044	0.156	0.704	0.946	1.000
		approx	0.040	0.146	0.697	0.938	1.000

Table 4: Comparison of power calculations based on simulated null distributions (denoted as "simul") and approximated distributions (denoted as "approx"). For each test, two sample sizes, $n = 100$ and $n = 200$, are used.

F_2 statistics. Note that λ is fixed in these approximations. In Figure 5, for $\lambda = 0.01, 0.001, 0.0001, 0.00001$ and $n = 100$, we show the QQplots of the F distributions with degrees of freedom ν_1^2/ν_2 and δ_1^2/δ_2 against the Monte Carlo null distribution of F_2. They show good approximations for different λ. The F approximation to the null distribution of the F_1 statistic is also examined. The plots are very similar to those in Figure 5.

5.3 Robustness of the tests

The approximate LMP, GML and SKL tests are derived based on the assumption that observations follow iid normal distributions. To check the robustness of the tests to the iid normal assumption, we generate observations from

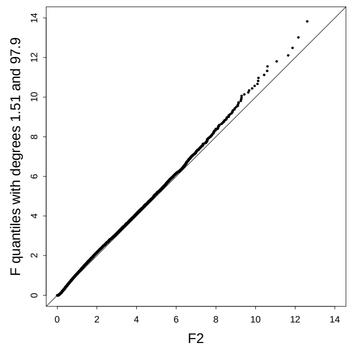
$$y_i = 1 + t_i + at_i \exp(-2t_i) + \epsilon_i, \qquad i = 1, \cdots, 100, \tag{26}$$

where $t_i = (i-1)/99$. We consider four choices of a, 0, 0.1, 0.3 and 0.5, and five choices of random errors: $\epsilon_i \overset{iid}{\sim} N(0, 0.2^2)$, $\epsilon_i \overset{iid}{\sim} t_3$ (t distribution with 3 degrees of freedom), $\epsilon_i \overset{iid}{\sim}$ 50:50 mixture of $N(-2, 2^2)$ and $N(2, 1)$, $\epsilon_i \sim$ AR(1) with autoregression coefficient 0.25, and $\boldsymbol{\epsilon} \sim N(0, W)$ where W is a diagonal matrix with diagonal elements evenly spaced between 0.04 and 0.22. Data are scaled to match variances. For each setting, we repeat the simulation 1000 times. Null distributions are calculated by Monte Carlo method.

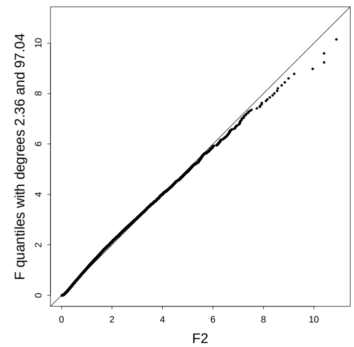
Table 5 shows that all tests hold the levels except when random errors are correlated. The powers are similar when random errors are from iid distributions. When

		LMP	GML	GCV	F_1	F_2	SKL
a=0	$N(0, 0.2^2)$	0.049	0.047	0.048	0.047	0.046	0.052
	t_3	0.052	0.058	0.043	0.053	0.051	0.037
	Mixture	0.047	0.049	0.048	0.047	0.050	0.042
	AR(1)	0.128	0.176	0.540	0.089	0.147	0.574
	$N(0, W)$	0.055	0.055	0.051	0.056	0.056	0.037
a=0.1	$N(0, 0.2^2)$	0.063	0.064	0.061	0.075	0.071	0.045
	t_3	0.090	0.084	0.069	0.085	0.080	0.042
	Mixture	0.073	0.083	0.085	0.082	0.087	0.065
	AR(1)	0.110	0.173	0.498	0.090	0.149	0.545
	$N(0, W)$	0.041	0.047	0.062	0.062	0.046	0.063
a=0.3	$N(0, 0.2^2)$	0.134	0.128	0.102	0.120	0.115	0.049
	t_3	0.148	0.144	0.094	0.138	0.134	0.042
	Mixture	0.138	0.135	0.108	0.132	0.131	0.066
	AR(1)	0.168	0.219	0.550	0.092	0.175	0.595
	$N(0, W)$	0.069	0.067	0.064	0.062	0.063	0.054
a=0.5	$N(0, 0.2^2)$	0.417	0.398	0.289	0.403	0.393	0.095
	t_3	0.453	0.434	0.335	0.421	0.409	0.114
	Mixture	0.414	0.403	0.300	0.385	0.378	0.107
	AR(1)	0.212	0.262	0.553	0.130	0.185	0.587
	$N(0, W)$	0.104	0.103	0.090	0.110	0.106	0.069

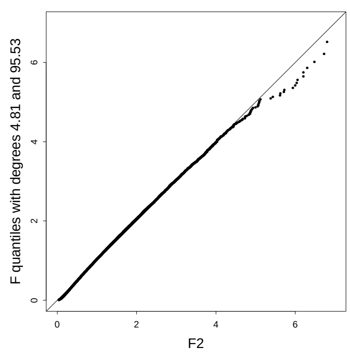
Table 5: Proportion of rejections in 1,000 replications. Random errors are generated from independent $N(0, 0.2^2)$, t_3 distribution, mixture of normal distributions, AR(1) process and $N(0, W)$.



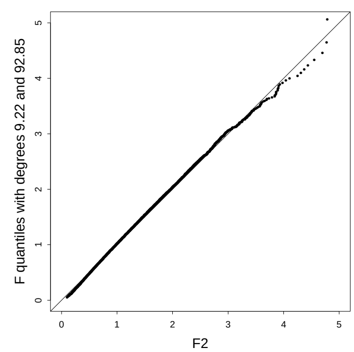
(a) $\lambda = 0.01$



(b) $\lambda = 0.001$



(c) $\lambda = 0.0001$



(d) $\lambda = 0.00001$

Figure 5: QQplots of quantiles of the F approximations against the quantiles of the Monte Carlo null distributions of F_2.

random errors are from $N(0, W)$, the powers are significantly lower than those in the iid cases. The comparative results of different tests remain the same. We conclude that the tests are quite robust to the violation of the normality assumption, but not to the independence assumption. For independent data with heterogeneous variances, the tests can approximately maintain their levels but lack power. We also conducted simulations under settings as in Section 5.1. Results remain the same.

6 Discussion

The connection between smoothing spline models and the Bayesian models (the mixed effects models) transfers the hypothesis on parametric regression to a much simpler hypothesis on a variance component. The approximate LMP, GML and GCV tests derived from the Bayesian model (or the mixed effect model) work well under the deterministic models. The good properties of the tests make them desirable for more complicated models. The hypothesis (7) can be written more generally as

$$\mathrm{H}_0 : f \in \mathcal{M}_0, \qquad \mathrm{H}_1 : f \in \mathcal{M}_1 \text{ and } f \notin \mathcal{M}_0,$$

where $\mathcal{M}_0$ is the model space under the null hypothesis, and $\mathcal{M}_1$ is a bigger model space which contains a substantially large family of plausible functions. $\mathcal{M}_0$ could be a linear or nonlinear parametric model, or a simple nonparametric model. For example, to test a nonlinear regression model $\mathcal{M}_0$, one can use nonlinear partial splines (Wahba 1990) or nonlinear nonparametric regression models (Ke and Wang 2002) as $\mathcal{M}_1$. To test an additive model $\mathcal{M}_0$ (Hastie and Tibshirani 1990), one can use SS ANOVA models (Wahba 1990) as $\mathcal{M}_1$ and test interaction components equal zero. This approach can also be employed to test the functional form of fixed effects in a mixed effects model. For example, to test linear mixed effects models, one may use the semi-parametric mixed effects models in Wang (1998b) as $\mathcal{M}_1$. To test nonlinear mixed effects models, one may use the semi-parametric nonlinear mixed effects models in Ke and Wang (2001) as $\mathcal{M}_1$. Another direction is to extend current test methods for non-Gaussian data, which will allow us to test the generalized linear models (McCullagh and Nelder 1989), the generalized additive models (Hastie and Tibshirani 1990) and the generalized linear mixed effects models (Breslow and Clayton 1993). All current methods are sensitive to the independence assumption. Thus new methods need to be developed for correlated data. Some research has been done in these directions. Guo (2001) generalized the GML test to the mixed effect SS ANOVA models. Zhang and Lin (2002) generalized the score test, which is equivalent to the approximate LMP test, to the semiparametric additive mixed models with non-Gaussian data. The SKL test was initially developed for smoothing spline models with non-Gaussian data by Xiang and Wahba (1995). These generalizations all showed good performances. We are currently working on extensions of the approximate LMP, GML and GCV tests for SS ANOVA model with non-Gaussian data. Preliminary results are encouraging. Tests for more complicated models as described above will be pursued in the future.

References

Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships, *Journal of the Royal Statistical Society B* **55**: 549–557.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statisticial Association* **88**: 9–25.

Cantoni, E. and Hastie, T. (2000). Degrees of freedom tests for smoothing splines, To appear in Biometrika.

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, *The Annals of Statistics* **16**: 113–119.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Crainiceanu, C. M. and Ruppert, D. (2002). Asymptotic distribution of likelihood ratio tests in linear mixed models, Unpublished.

Crainiceanu, C. M., Ruppert, D. and Vogelsang, T. J. (2002). Probability that the mle of a variance component is zero with applications to likelihood ratio tests, submitted to the Journal of the American Statistical Association.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables, *Applied Statistics* **29**: 323–333.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a liear model via nonparametric regression techniques, *Journal of the american statistical association* **85**: 387–392.

Guo, W. (2001). Inference in smoothing spline anova, To appear in the Journal of The Royal Statistical Society, Ser B.

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

Ke, C. and Wang, Y. (2001). Semi-parametric nonlinear mixed effects models and their applications (with discussion), *Journal of the American Statistical Association* **96**: 1272–1298.

Ke, C. and Wang, Y. (2002). Nonparametric nonlinear regression models, Unpublished.

Kimeldorf, G. and Wahba, G. (1971). Some results on tcheycheffian spline functions, *J. Math. Anal. Appl.* **33**: 82–94.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

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.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association* **82**: 605–610.

Wahba, G. (1990). *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.

Wang, Y. (1998a). Smoothing spline models with correlated random errors, *Journal of the American Statistical Association* **93**: 341–348.

Wang, Y. (1998b). Mixed-effects smoothing spline ANOVA, *Journal of the Royal Statistical Society B* **60**: 159–174.

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. and Lin, X. (2002). Hypothesis testing in semiparametric additive mixed models, To appear in Biostatistics.