Biometrika (2002), **89**, 2, *pp.* 251–263

Printed in Great Britain

Degrees-of-freedom tests for smoothing splines

By EVA CANTONI

Department of Econometrics, University of Geneva, 40, Boulevard du Pont d'Arve,
1211 Geneva 4, Switzerland
eva.cantoni@metri.unige.ch

and TREVOR HASTIE

Statistics Department, Stanford University, Sequoia Hall–390 Serra Mall, Stanford,
California 94305, U.S.A.
hastie@stat.stanford.edu

Summary

When using smoothing splines to estimate a function, the user faces the problem of choosing the smoothing parameter. Several techniques are available for selecting this parameter according to certain optimality criteria. Here, we take a different point of view and we propose a technique for choosing between two alternatives, for example allowing for two different levels of degrees of freedom. The problem is addressed in the framework of a mixed-effects model, whose assumptions ensure that the resulting estimator is unbiased. A likelihood-ratio-type test statistic is proposed, and its exact distribution is derived. Tests of linearity and overall effect follow directly. We then extend this idea to additive models where it provides a more attractive alternative than multi-parameter optimisation, and where it gives exact distributional results that can be used in an analysis-of-deviance-type approach. Examples on real data and a simulation study of level and power complete the paper.

Some key words: Additive model; Degrees-of-freedom test; Smoothing parameter selection; Smoothing spline.

1. Motivation

Selecting a value for the smoothing parameter, or equivalently the effective degrees of freedom, is a well-studied problem in nonparametric regression. In Fig. 1(a) given below, for example, one could ask if the solid line describes the relationship between the two variables well enough or if a more flexible fit, like the dashed line, is needed. The aim of this paper is to provide a test statistic for choosing between two such alternatives.

We consider the model

$$y_i = f(x_i) + \varepsilon_i, \qquad (1{\cdot}1)$$

for each individual i of a sample of size n, where y_i is the outcome, x_i the explanatory variable, f the function describing the relationship between x_i and y_i, and ε_i the error term. We focus on the estimation of f by smoothing splines. It is well known that the smoothing spline fit is computed as a linear transformation of the vector $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and that the fitted values are $\hat{y}_\lambda = (I + \lambda K)^{-1} y = S_\lambda y$; see Green & Silverman (1994, pp. 18–9) and the Appendix for more details. The parameter λ controls the smoothness of the fit, which can also be defined through the effective degrees of freedom (Hastie & Tibshirani, 1990, Ch. 3; Wahba, 1990, p. 63):

$$\mathrm{DF}_\lambda = \mathrm{tr}(S_\lambda) = \sum_{i=1}^{n} \frac{1}{1 + \lambda d_i}, \tag{1·2}$$

where d_i are the eigenvalues of the matrix K; the quadratic form $\hat{y}_\lambda^{\mathrm{T}} K \hat{y}_\lambda$ measures the roughness of the fitted function. There is a strictly monotone relationship between λ and DF_λ, allowing us to work with this latter notion, which ties in gracefully with parametric linear modelling concepts.

Classical approaches for selecting λ or DF_λ have considered the optimisation of some optimality criteria, such as estimators of the mean squared error. This is the case for crossvalidation or Mallows' C_p, for example. Another common approach is to derive analytical expressions for the mean squared error, from which the optimal value of the parameter can be obtained. This optimal value usually depends on the underlying unknown function, for which a pilot estimate must be 'plugged in'. Instead, we construct a test statistic for choosing between two predefined alternatives, much as in parametric modelling. Furthermore, it is usual practice to use such tests in building additive models (Hastie & Tibshirani, 1990). This approach avoids the optimisation of a multi-dimensional criterion; an additive model with p terms has p smoothing parameters. By limiting the parameter choice for each term to a small number of alternatives defined in terms of degrees of freedom, we allow the user to make some pragmatic choices. For example there might be four ordered alternatives for a term, such as 'absent', 'linear', '4 degrees of freedom' and '8 degrees of freedom', and the techniques discussed in this paper allow us to test hypotheses for choosing among them.

There are essentially two different approaches to model (1·1); either $f(x)$ is considered to be a fixed unknown function or else is assumed to be a realisation from a particular Gaussian process. They both lead to the same smoothing spline estimate but to different inference models. We choose the latter approach, also known as a 'mixed-effects model', since it avoids issues of bias by making stronger assumptions about the stochastic model for f.

Section 2 gives the basics of this mixed-effects model. The test statistic we are interested in is derived in § 3. A real example is worked out in § 4, which is followed in § 5 by a simulation study of the properties of the likelihood-ratio statistic. Finally, § 6 considers the extension to additive models. Technical details are collected in the Appendix.

2. Mixed-effects setting for smoothing splines

Mixed-effects models have gained popularity for analysing longitudinal and other correlated-data scenarios, and they provide a useful representation for smoothing splines (Lin & Zhang, 1999; Wahba, 1990). Consider the mixed-effects model

$$Y = X\beta + Zu + \varepsilon, \tag{2·1}$$

made up of a linear fixed effect $X\beta$, a nonlinear random effect Zu and an independent error term. In (2·1), $X = [1 \;\; x] \in \mathbb{R}^{n \times 2}$, where $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ is the vector of predictor values, $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ are unknown parameters, and $Z = Z(x) \in \mathbb{R}^{n \times (n-2)}$ is a matrix representing nonlinear functions of x. The columns of Z are orthogonal to the columns of X. Successive columns of Z are increasingly rough, as measured by the quadratic penalty matrix K, and are scaled to have decreasing Euclidean norm; see the Appendix for further technical details about Z. The vector $u \sim \mathcal{N}(0, \tau^2 I_{n-2})$ is a random effect, and hence its product with Z produces a random component nonlinear in x, whose size is controlled by the variance parameter τ^2. The vector $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is the error term, independent of u.

The best linear unbiased predictors for β and u of model (2·1) satisfy the following equations (Henderson, 1950; Robinson, 1991):

$$(X^{\mathrm{T}}X)\hat{\beta} = X^{\mathrm{T}}(y - Z\hat{u}), \quad (Z^{\mathrm{T}}Z + \lambda I_{n-2})\hat{u} = Z^{\mathrm{T}}(y - X\hat{\beta}),$$

where $\lambda = \sigma^2/\tau^2$ is assumed known. These equations are obtained by maximisation of the joint density of y and u with respect to β and u under the normality assumptions; see Henderson (1973) and Robinson (1991).

It follows from the orthogonality $X^{\mathrm{T}}Z = Z^{\mathrm{T}}X = 0$, see the Appendix, that the estimators for β and u are

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \quad \hat{u} = (\lambda I_{n-2} + Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y. \tag{2·2}$$

The 'unbiased' aspect of these predictors refers here to the property that the average value of the estimator is equal to the average value of the quantity being estimated, that is $E(\hat{u}) = E(u)$. They can also be seen as posterior means in an empirical Bayes model.

We show in the Appendix that the fitted values obtained with (2·2), namely $\hat{y}_\lambda = X\hat{\beta} + Z\hat{u}$, are the same as $\hat{y}_\lambda = S_\lambda y$. This shows (Speed, 1991) that the best linear unbiased predictors obtained for this particular mixed-effects model are identical to the smoothing spline obtained by penalised least squares.

So far we have assumed τ^2 and σ^2 to be known; the marginal likelihood for Y, which is $\mathcal{N}\{X\beta, \sigma^2(I_n + \lambda^{-1}ZZ^{\mathrm{T}})\}$, also allows inference about these parameters, or the noise-to-signal ratio λ. For example, λ could be estimated by maximum likelihood; see Wecker & Ansley (1983) and Wahba (1990, pp. 63–4). This approach is called the generalised maximum likelihood criterion and is equivalent to the maximum likelihood estimation derived from model (2·1). In this paper, we derive a likelihood-ratio-type test for choosing between two alternatives. Our approach is identical to the inference that would have been obtained on the variance parameters by empirical Bayes.

3. A likelihood ratio test statistic

3·1. *Derivation of the test statistic*

Suppose that σ is known. The inference about the smoothing parameter can be carried out on the parameter τ of the marginal distribution of Y. We consider the test of the null hypothesis $H_0: \tau = \tau_0$ versus the alternative hypothesis $H_A: \tau = \tau_1 > \tau_0$, corresponding to the test of $H_0: \lambda = \lambda_0$ versus $H_A: \lambda = \lambda_1 < \lambda_0$, or equivalently $H_0: \mathrm{DF} = \mathrm{DF}_{\lambda_0}$ against $H_A: \mathrm{DF} = \mathrm{DF}_{\lambda_1} > \mathrm{DF}_{\lambda_0}$.

Denote by $l_\tau(y)$ the density associated with the marginal distribution of Y, a $\mathcal{N}\{X\beta, \sigma^2(I_n + \lambda^{-1}ZZ^{\mathrm{T}})\}$ distribution, and consider the corresponding log likelihood-ratio statistic

$$\begin{aligned}\log l_{\tau_1}(y) - \log l_{\tau_0}(y) &\propto (y - X\beta)^{\mathrm{T}}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})^{-1} - (I_n + \lambda_1^{-1}ZZ^{\mathrm{T}})^{-1}\}(y - X\beta)\\ &= y^{\mathrm{T}}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})^{-1} - (I_n + \lambda_1^{-1}ZZ^{\mathrm{T}})^{-1}\}y,\end{aligned}$$

where the last equality holds because $(I_n + \lambda^{-1}ZZ^{\mathrm{T}})^{-1} = I_n - Z(\lambda I_{n-2} + Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}$ and $Z^{\mathrm{T}}X = 0$.

We can define a test statistic by

$$T = y^{\mathrm{T}}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})^{-1} - (I_n + \lambda_1^{-1}ZZ^{\mathrm{T}})^{-1}\}y$$
$$= y^{\mathrm{T}}(S_{\lambda_1} - S_{\lambda_0})y = y^{\mathrm{T}}(\hat{y}_{\lambda_1} - \hat{y}_{\lambda_0}), \tag{3·1}$$

where $\hat{y}_\lambda = S_\lambda y$ are the fitted values. Gray (1994) develops a similar testing procedure in the setting of survival analysis, the semiparametric proportional hazard model, to test linearity and no effect. In both these particular testing situations, the penalised likelihood ratio statistic Q_l in Gray (1994) is equivalent to our statistic (3·1) when transferred to the Gaussian regression setting. Note however that our formulation gives a general testing framework where we can test more sophisticated hypotheses on degrees of freedom than simply linear and overall effect.

The distribution of T depends on σ^2; see § 3·3 below. One can either plug in a reliable estimate, or consider a ratio-type statistic such as

$$\Lambda = \frac{y^{\mathrm{T}}(S_{\lambda_1} - S_{\lambda_0})y}{y^{\mathrm{T}}(I - S_{\tilde{\lambda}})y} = \frac{y^{\mathrm{T}}(\tilde{y}_{\lambda_1} - \hat{y}_{\lambda_0})}{y^{\mathrm{T}}(y - \hat{y}_{\tilde{\lambda}})}, \tag{3·2}$$

for some value $\tilde{\lambda}$. We will discuss issues related to the choice of $\tilde{\lambda}$ in § 3·3 along with the distribution of the statistic Λ.

For projection operators, as in linear regression, statistic (3·2) is equivalent to the statistic that would compare the sums of squared residuals of the fits, because in this case $y^{\mathrm{T}}(y - \hat{y}) = (y - \hat{y})^{\mathrm{T}}(y - \hat{y}) = \|(I - H)y\|^2$, where H is the hat matrix. In fact, the heuristic approach used by Hastie & Tibshirani (1990, Ch. 3) and Chambers & Hastie (1991, Ch. 7) for the comparison of the degrees of freedom of two nonparametric fits is inspired by the theory of linear models and makes use of the information contained in the residual sum of squares by means of the test statistic

$$F = \frac{\{\|(I - S_{\lambda_0})y\|^2 - \|(I - S_{\lambda_1})y\|^2\}/(v_1 - v_0)}{\|(I - S_{\lambda_1})y\|^2/(n - v_1)}, \tag{3·3}$$

which is approximated by an $F_{v_1 - v_0, n - v_1}$ distribution with $v_i = \mathrm{tr}(2S_{\lambda_i} - S_{\lambda_i}S_{\lambda_i}^{\mathrm{T}})$ for $i = 0, 1$. This assumes that the numerator and the denominator in (3·3) are approximated by independent $\chi^2_{v_1 - v_0}$ and $\chi^2_{n - v_1}$ variables respectively. A further and computationally less expensive approximation consists of taking $v_i = \mathrm{tr}(S_{\lambda_i})$, as used in S-Plus.

Let us investigate the numerous levels of approximation involved in this procedure. The approach relies on a model of the form

$$y_i = f(x_i) + \varepsilon_i, \tag{3·4}$$

where $f(x_i)$ is supposed to be fixed and ε_i is the random component, for which a $\mathcal{N}(0, \sigma^2)$ distribution is usually assumed. Therefore, the exact distribution of the numerator in (3·3) is a linear combination of noncentral χ^2_1 variables. Hence, saying that the numerator is $\chi^2_{v_1 - v_0}$ distributed accounts for two different sources of approximation. First, function estimators defined by smoothing techniques based on model (3·4) almost always suffer from bias, which is neglected when using (3·3). Secondly, the weighted χ^2_1 combination is approximated by a unique χ^2 variable. The same comments apply to the denominator in (3·3). In addition, the F-approximation does not take into account the dependence between the numerator and denominator of (3·3).

The procedure using F could be improved by refining the distributional properties (Hastie & Tibshirani, 1990, pp. 66–7), but the bias problem will still be present. In our Bayesian procedure, we finesse the bias by assuming the mixed-effects model.

3·2. *Particular cases and generalisation*

Test statistic (3·2) tests linearity by testing $H_0 : \mathrm{DF} = 2$ versus $H_A : \mathrm{DF} > 2$, giving a competitor of the linearity test of Azzalini & Bowman (1993). It can also test overall effect, equivalent to a term being dropped, by testing $H_0 : \mathrm{DF} = 1$ versus $H_A : \mathrm{DF} > 1$. The generalisation to the test of composite hypotheses of the form $H_0 : \mathrm{DF} = \mathrm{DF}_{\lambda_0}$ against $H_A : \mathrm{DF} = \mathrm{DF}_{\lambda_1} > \mathrm{DF}_{\lambda_0}$, unspecified, can be obtained by estimating DF and σ^2 using restricted maximum likelihood under model (2·1) and using the estimated DF instead of DF_{λ_1}.

The test statistics are obtained by plugging the appropriate operators for S_λ into (3·2), such as the hat matrix H for a linear fit and the averaging operator $11^{\mathrm{T}}/n$, with $1 = (1, \ldots, 1)^{\mathrm{T}}$, for the constant fit.

3·3. *Distribution of the test statistic under H_0*

The numerator of statistic (3·2) is a quadratic form in normal variables, and equals $\sigma^2 \sum_{i=3}^n c_i z_i^2$ under H_0, with $c_i = 1 - (d_i + \lambda_0^{-1})/(d_i + \lambda_1^{-1})$ and z_i being independent standard normal variables; see the Appendix and Gray (1994). The c_i's are in fact the eigenvalues of $(I_n + \lambda_0^{-1} ZZ^{\mathrm{T}})(S_{\lambda_1} - S_{\lambda_0})$. The denominator of (3·2) under H_0 is distributed according to a $\sigma^2 \sum_{i=3}^n b_i z_i^2$ distribution, where the b_i are the eigenvalues of $(I_n + \lambda_0^{-1} ZZ^{\mathrm{T}})(I_n - S_{\tilde{\lambda}})$ and are equal to $(d_i + \lambda_0^{-1})/(d_i + \tilde{\lambda}^{-1})$.

Note that linear combinations of χ^2 variables have been well studied (Johnson & Kotz, 1970, Ch. 29), and algorithms are available for computing relevant probabilities (Davies, 1980; Farebrother, 1990).

Although the distribution of its numerator and its denominator are known, the distribution of Λ itself does not belong to a known family. One could consider approximating each of the two linear combinations of χ_1^2 variables by a χ^2 variable that matches the first moment; that is we approximate the numerator with a $\chi^2_{\sum c_i}$ distribution and the denominator with a $\chi^2_{\sum b_i}$ distribution. Then, if the two variables are almost independent, the distribution of Λ is approximately $(\sum c_i / \sum b_i) F_{\sum c_i, \sum b_i}$, with

$$\sum c_i = \mathrm{tr}\{(I_n + \lambda_0^{-1} ZZ^{\mathrm{T}})(S_{\lambda_1} - S_{\lambda_0})\}, \quad \sum b_i = \mathrm{tr}\{(I_n + \lambda_0^{-1} ZZ^{\mathrm{T}})(I_n - S_{\tilde{\lambda}})\}.$$

Both these traces have computationally more appealing expressions, as shown in the Appendix. To guarantee approximate independence between the numerator and the denominator of the statistic in (3·2), one has to choose a 'small' value $\tilde{\lambda}$ for the smoothing parameter. However, it is easy to construct simple examples, involving a 'small' $\tilde{\lambda}$ in which the weights b_i of the χ^2 combination of the denominator in (3·2) are spread out, which could render inadequate the single χ^2 approximation suggested above. A better approximation can be obtaining through a two-moment correction.

However, one is usually interested in obtaining p-values, which can be computed exactly much more easily by noting that the p-value $\mathrm{pr}(\Lambda > v_{\mathrm{obs}})$, for the value v_{obs} of (3·2) from the dataset, can be rewritten as

$$\mathrm{pr}[y^{\mathrm{T}}\{S_{\lambda_1} - S_{\lambda_0} - v_{\mathrm{obs}}(I - S_{\tilde{\lambda}})\}y > 0] = \mathrm{pr}(R > 0). \qquad (3·5)$$

The distribution of R is again a linear combination of χ^2 variables because $R = \sum_{i=3}^n e_i z_i^2$, with

$$e_i = c_i - v_{\mathrm{obs}} b_i = 1 - \frac{d_i + \lambda_0^{-1}}{d_i + \lambda_1^{-1}} - v_{\mathrm{obs}} b_i.$$

The e_i are the eigenvalues of $(I_n + \lambda_0^{-1} ZZ^{\mathrm{T}})\{S_{\lambda_1} - S_{\lambda_0} - v_{\mathrm{obs}}(I - S_{\tilde{\lambda}})\}$ although details are not given here.

This approach is appealing because it is scale invariant and is not affected by the lack of independence between the variable of the numerator and the denominator in (3·2). The same approach has been used by other authors; see for instance Azzalini & Bowman (1993). The choice of $\tilde{\lambda}$ is irrelevant and we suggest using $\tilde{\lambda} = \lambda_1$ to avoid additional computations. In this case, we have that $e_i = 1 - (1 + v_{\text{obs}})(d_i + \lambda_0^{-1})/(d_i + \lambda_1^{-1})$.

3·4. *Computational aspects*

The complete exact procedure is of complexity $O(n^3)$, although statistics Λ and R can be computed in linear time because fitted values can be obtained in $O(n)$ steps. The complexity of the algorithm is essentially due to the extraction of eigenvalues of dense matrices. One could improve this by considering either low-rank splines (Eilers & Marx, 1992; Hastie, 1996) or numerical algorithms that intelligently extract only the largest eigenvalues (Bai et al., 2000, Ch. 4). However, the speed of modern computers is such that samples of up to moderate sample sizes can be handled in very few seconds. Thanks to expressions (A·3) and (A·4), the approximation to the distribution of (3·2) by an F distribution involves only $O(n)$ computations. If this approximation proves accurate, it substantially reduces the computational burden of the procedure.

Procedure (3·3) based on sums of squared residuals is of order $O(n^2)$ when the degrees of freedom are computed as $v_i = \text{tr}(2S_{\lambda_i} - S_{\lambda_i}S_{\lambda_i}^{\text{T}})$, even if the statistic F itself can be computed in linear time. The computational price goes down to $O(n)$ when the approximated degrees of freedom $v_i = \text{tr}(S_{\lambda_i})$ are used instead.

4. EXAMPLE

We apply the test developed in the previous section to the vineyard dataset studied by Simonoff (1996, p. 287) and as given in Chatterjee et al. (1995).

The data consist of the grape yields of a vineyard on a small island in Lake Erie. The vineyard is divided into 52 rows and the 52 observations in the dataset correspond to the sums of the yields of the harvests in 1989, 1990 and 1991. The yield is measured as a number of ‘lugs’, a lug being a basket used to carry the harvest grapes.

Suppose we want to choose the degrees of freedom from the set {6DF, 10DF, 14DF}. Along the lines of an analysis of deviance, we perform two tests. The first one corresponds to Fig. 1(a), and compares a spline fit with 6 degrees of freedom, which defines H_0, to a spline fit with 10 degrees of freedom, H_A. According to the statistic (3·2), the null hypothesis is clearly rejected, with p-value $< 0{\cdot}002$. Figure 1(b) corresponds to the fit of a spline with 10 degrees of freedom, under H_0, versus a spline with 14 degrees of freedom. In this case, the null hypothesis is not rejected, p-value $\simeq 0{\cdot}2$, meaning that we do not need as many as 14 degrees of freedom to describe the relationship between the row in the vineyard and the amount of grapes yielded. We end up by retaining a fit with 10 degrees of freedom.

5. SIMULATION

5·1. *Protocol*

Here we conduct a small simulation study to compare the F-approximation of § 3·3 to the exact result, and to the heuristic procedure in (3·3). The simulation setting is defined by the marginal null model

$$Y \sim \mathcal{N}\{X\beta, \sigma^2(I_n + \lambda_0^{-1}ZZ^{\text{T}})\}, \qquad (5{\cdot}1)$$

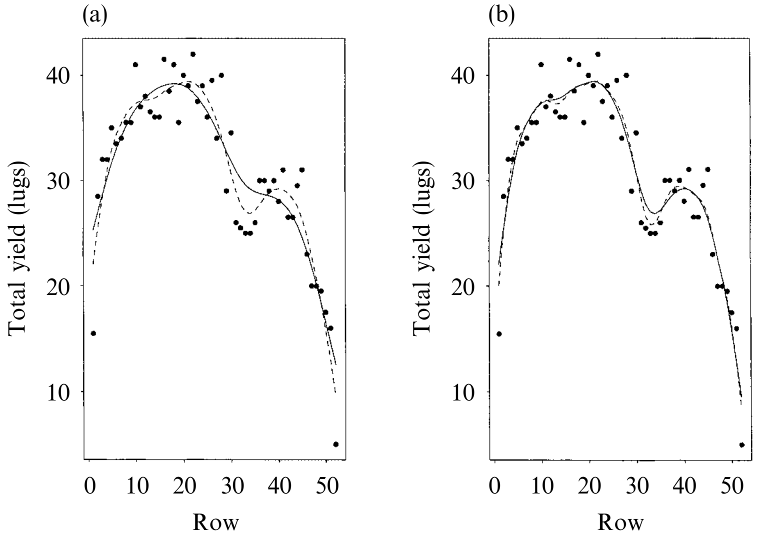

Fig. 1. Fit comparison on the vineyard dataset. (a) compares a spline fit with DF = 6 (solid line) to a spline fit with DF = 10 (dashed line); (b) compares a fit with DF = 10 (solid line) to a fit with DF = 14 (dashed line).

with $\beta = (1, 5)^{\mathrm{T}}$, $\sigma^2 = 0{\cdot}5^2$ and λ_0 corresponding to DF = 4. The alternative hypothesis considers a smoothing parameter λ_1 corresponding to DF = 7. Moreover, $X = [1 \;\; x]$, with $x \sim \mathrm{Un}(0, 1)$ generated at the beginning of the simulation. We will compare the following situations.

Case 1: $\Lambda \sim (\sum c_i/\sum b_i)F_{\sum c_i, \sum b_i}$ with $\sum c_i$ and $\sum b_i$ according to (A·3) and (A·4) and $\tilde{\lambda}$ corresponding to DF = 20.

Case 2: $R \sim \sum e_i z_i^2$.

Case 3: $F \sim F_{\nu_1 - \nu_0, n - \nu_1}$, with $\nu_i = \mathrm{tr}(2S_{\lambda_i} - S_{\lambda_i} S_{\lambda_i}^{\mathrm{T}})$.

Case 4: $F \sim F_{\nu_1 - \nu_0, n - \nu_1}$, with $\nu_i = \mathrm{tr}(S_{\lambda_i})$.

Cases 1 and 2 use our procedure based on statistic (3·2) with approximate F and exact distribution respectively. Case 3 is the procedure based on statistic (3·3) with the matching degrees of freedom, whereas Case 4 is the same procedure but with approximated degrees of freedom.

Five thousand simulations were run with sample sizes $n = 40$ and $n = 100$ and with the precision in Davies' algorithm set to 0·0001.

5·2. *Discussion of p-values and level*

Figure 2 shows some Q–Q plots of the p-values of each technique against the uniform distribution for sample size $n = 40$. The plots for $n = 100$ were virtually identical. These plots show that the approximation to the distribution of (3·2) corresponding to Case 1 does not produce uniformly distributed p-values, but rather a distribution with shorter and lighter tails. A slightly larger deviation from the uniform distribution is observed in the upper tail, which is fortunately of minor interest in a testing procedure. The use of the test statistic F, in Cases 3 and 4, gives rise to even worse results than for Case 1. In both cases, the distribution of the p-value is far from the uniform target and is clearly skewed, particularly so in Case 4. Moreover, p-values near to 1 never appear in the simulation for these two cases. As expected, the exact distribution yields uniformly distributed p-values.

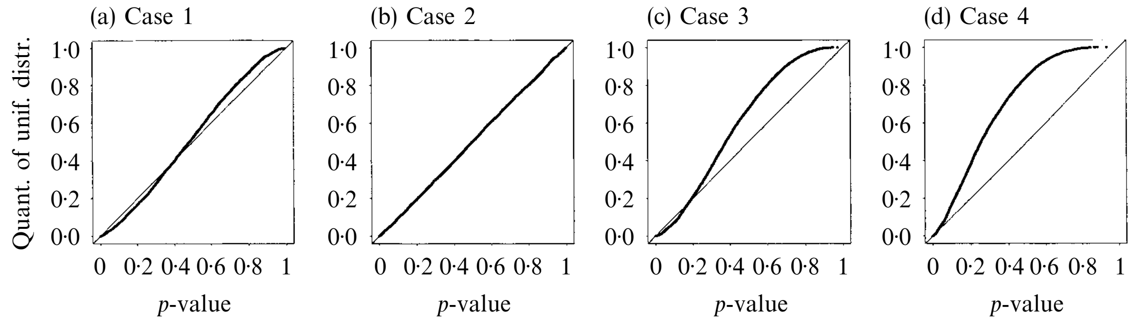

Fig. 2. Q–Q plots of the simulated p-values of the test statistic (3·2) against the quantiles of the uniform distribution when testing $H_0 : \text{DF}_0 = 4$ versus $H_A : \text{DF}_1 = 7$. Sample size is $n = 40$.

Next we examine the actual levels of the test with respect to nominal levels of 1%, 5% and 10%, estimated empirically from the simulations and shown in Table 1. The last line of Table 1 gives the standard deviations of the level estimation, which do not depend on n. As expected theoretically, the level computed by the exact distribution, Case 2, behaves well: the nominal level is always covered by approximate 95% confidence intervals. For the approximation in Case 1, the actual level ranges between 38% and 77% of the nominal level, whatever the sample size. The behaviour of the actual level is particularly bad at the extreme of the distribution; see the results at the 1% nominal level. In Case 3, the results are similar to those of Case 1, with an actual level in a range of 28–84% of the nominal level. Nevertheless, both tests under Cases 1 and 3 are conservative. This is not the case for the approach in Case 4, which is clearly not on target. This is of no surprise in view of Fig. 2 and is a consequence of the accumulation of different sources of approximation. Approximate 95% confidence intervals do not cover the true nominal level in these cases, except for Case 4 with $n = 100$ at the 1% level.

Table 1. *Actual levels of the test statistic* (3·2) *under techniques defined by Cases* 1, 2, 3 *and* 4 *when testing* $H_0 : \text{DF}_0 = 4$ *versus* $H_A : \text{DF}_1 = 7$. *The standard deviations of the level estimation, given in the last line, do not depend on n*

	Nominal 1%		Nominal 5%		Nominal 10%	
Case	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
1	0·0038	0·0038	0·0296	0·0318	0·0692	0·0774
2	0·0086	0·0108	0·0452	0·0548	0·0952	0·1040
3	0·0028	0·0036	0·0284	0·0298	0·0720	0·0840
4	0·0068	0·0076	0·0678	0·0724	0·1674	0·1728
St. dev.	0·0014	0·0014	0·0031	0·0031	0·0042	0·0042

It seems therefore that the F-approximation to statistic (3·2) can be conservative, but it has the advantage of being inexpensive. The exact result is computationally more expensive, but the gain in accurarcy is high. Use of the accurate exact distribution is therefore strongly recommended when computationally feasible. The heuristic procedure in Case 4 can even be nonconservative. The approaches in Case 1 and Case 2 both outperform the results obtained by Case 3 and Case 4.

5·3. *Power*

Following the same protocol as in § 5·2, we also conducted a power study for a sequence of alternatives corresponding to $\mathrm{DF}_{\lambda_1} = 5, \ldots, 12$ when $n = 40$. We simulated the response Y from model (5·1) with λ_0 replaced by λ_1.

The power curves, corrected to ensure a level of 5%, for Cases 1–4 are displayed in Fig. 3. These results show an overall superiority of the test statistics developed in this paper over the other approaches we considered. The power of the F-approximation of our test statistic, Case 1, is almost as high as the power of the exact test. For a similar computational cost, and considering that the error on the level was of the same magnitude, the F-approximation of Case 1 does a better job in terms of power than the approaches in Cases 3 and 4.

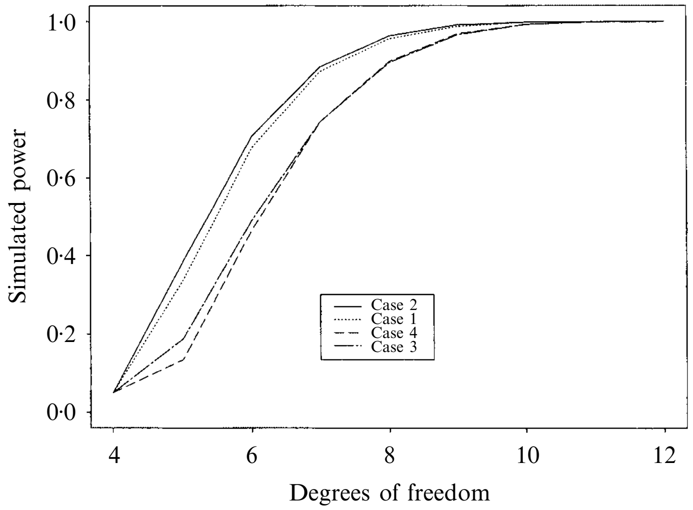

Fig. 3. Simulated powers of the test statistics for Cases 1–4, corrected to ensure a level of 5%.

6. Extension to additive models

The problem of the choice of the smoothing parameter is even more relevant in additive models, where one would prefer to avoid the optimisation of an optimality criterion over a p-dimensional space. From now on, we consider a p-dimensional additive model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \tag{6·1}$$

for $i = 1, \ldots, n$. Additive models admit a Bayesian formulation, which is a natural extension of the single predictor case, starting from the mixed-effects model

$$Y = \beta_0 1 + X\beta + \sum_{j=1}^{p} Z_j(x_j)u_j + \varepsilon,$$

with $1 = (1, \ldots, 1)^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $(X)_{ij} = x_{ij}$. As a result, the marginal distribution of Y is $\mathcal{N}\{\beta_0 1 + X\beta, \sigma^2(I_n + \sum_j \lambda_j^{-1} Z_j Z_j^{\mathrm{T}})\}$.

Additive models are usually fitted via the backfitting algorithm, and at convergence the solution can be written as $\hat{y}_\lambda = R_\lambda y = \sum_j R_{\lambda_j} y$, for a vector $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ of smoothing parameters. The degrees of freedom of the overall fit is $\mathrm{DF}_\lambda = \mathrm{tr}(R_\lambda)$, which can be decomposed into its p components $\mathrm{tr}(R_{\lambda_j})$. However, this definition of the single component degree of freedom is not attractive from the computational point of view (Hastie & Tibshirani, 1990, pp. 128–9). We will use instead the definition

$$\mathrm{DF}_{\lambda_j} = \mathrm{tr}(S_{\lambda_j}) - 1, \tag{6·2}$$

where S_{λ_j} is the smoother matrix obtained when fitting by smoothing spline the jth predictor only. The subtraction of one is because of the global constant term isolated in model (6·1).

To perform a test on the degrees of freedom of the kth component of model (6·1), we define a test statistic by analogy with (3·2). Formally, we would like to test the hypothesis $H_0 : \mathrm{DF}_k = \mathrm{DF}_{k,\lambda_0}$ $(\lambda_k = \lambda_{k,0})$ versus $H_A : \mathrm{DF}_k = \mathrm{DF}_{k,\lambda_1} > \mathrm{DF}_{k,\lambda_0}$ $(\lambda_k = \lambda_{k,1})$, while keeping the other parameters fixed as in the analysis-of-deviance approach to building additive models.

By analogy with the single predictor setting, we suggest the test statistic

$$\Lambda_{\mathrm{AM}} = \frac{y^{\mathrm{T}}(R_1 - R_0)y}{y^{\mathrm{T}}(I - R_1)y}, \tag{6·3}$$

where R_i, for $i = 0$ or 1, is the smoother matrix obtained at the convergence of the backfitting algorithm with the set of parameters including $\lambda_{k,i}$.

For the value $v_{\mathrm{AM,obs}}$ taken by Λ_{AM}, the p-value is computed by

$$\begin{aligned}\mathrm{pr}(\Lambda_{\mathrm{AM}} > v_{\mathrm{AM,obs}}) &= \mathrm{pr}[y^{\mathrm{T}}\{R_1 - R_0 - v_{\mathrm{AM,obs}}(I - R_1)\}y > 0] \\ &= \mathrm{pr}(R_{\mathrm{AM}} > 0).\end{aligned}$$

The distribution of R_{AM} under H_0 is a linear combination of χ_1^2 variables, where the weights are the eigenvalues of the matrix $(I_n + \sum_j \lambda_{j,0}^{-1} Z_j Z_j^{\mathrm{T}})\{R_1 - R_0 - v_{\mathrm{AM,obs}}(I - R_1)\}$. This follows again from general results on the distribution of quadratic forms in normal variables.

We remark that, as in the one-predictor case, statistic (6·3) can be extended to test composite hypotheses, and used to assess linearity and overall effect at no additional cost.

We illustrate the use of this procedure on a diabetes dataset (Sockett et al., 1987; Hastie & Tibshirani, 1990, p. 304), which comes from a study aiming at describing the factors that affect the patterns of insulin-dependent diabetes mellitus in children. The relationship between the concentration of C-peptide and the predictors Age and Base.Deficit, a measure of acidity, is under study for $n = 43$ children. We consider the model

$$\log(\text{C-peptide}) = \beta_0 + f_1(\text{Age}) + f_2(\text{Base.Deficit}). \tag{6·4}$$

Figure 4 shows the curves fitted by smoothing splines with $\mathrm{DF}_{\mathrm{Age}} = 2$, solid line in Fig. 4(a), and $\mathrm{DF}_{\mathrm{Base.Defect}} = 3$, in Fig. 4(b). Let us focus on the predictor Age. Is this fit flexible enough to describe the true underlying relationship? We can consider allowing more degrees of freedom for this variable, say 4 or 6 degrees of freedom, see Fig. 4(a), keeping the degrees of freedom of Base.Deficit equal to 3. We use statistic (6·3) to compare these alternatives. The test of the null hypothesis $H_0 : \mathrm{DF}_{\mathrm{Age}} = 2$ versus the alternative $H_A : \mathrm{DF}_{\mathrm{Age}} = 4$ yields a p-value of 2×10^{-6} indicating clearly that 2 degrees of freedom are not enough. The test of $H_0 : \mathrm{DF}_{\mathrm{Age}} = 4$ and $H_A : \mathrm{DF}_{\mathrm{Age}} = 6$, conditional on 3 degrees of freedom for Base.Deficit, yields a p-value of 0·56, which suggests that 6 degrees of freedom are probably unnecessarily high for describing this relationship.

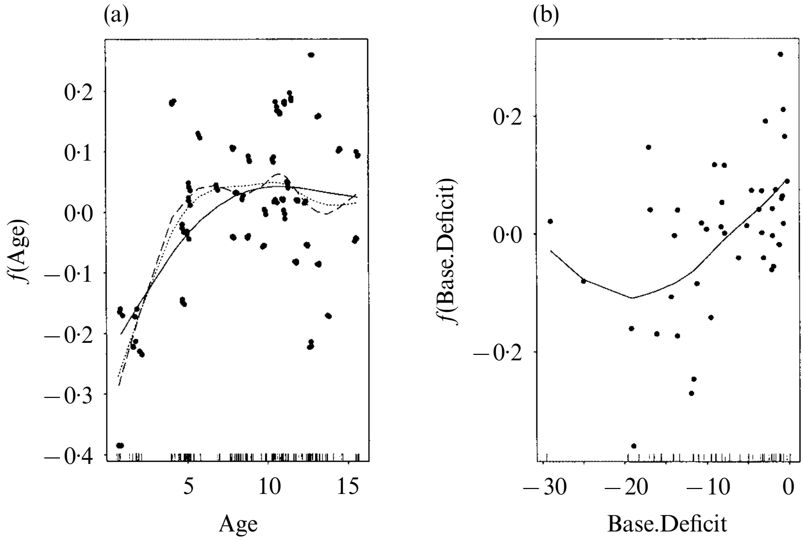

Fig. 4. Additive fit of the diabetes dataset. (a): Age with $\text{DF}_{\text{Age}} = 2$ (solid), $\text{DF}_{\text{Age}} = 4$ (dotted) and $\text{DF}_{\text{Age}} = 6$ (dashed). (b): Base.Deficit with $\text{DF}_{\text{Base.Deficit}} = 3$.

Acknowledgement

Eva Cantoni thanks the Swiss National Science Foundation for its support. Trevor Hastie was partially supported by the National Science Foundation, and the National Institutes of Health. This work has been carried out during the post-doctoral year of Eva Cantoni at Stanford University.

Appendix

Technicalities

Details for the mixed-effects formulation. Smoothing splines are the solution of the penalised criterion

$$J(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 \, dt. \tag{A·1}$$

If we assume that the solution is a spline, and for a parameterisation in terms of $f = (f(x_1), \ldots, f(x_n))^{\mathrm{T}}$, the penalty in (A·1) can also be written as $f^{\mathrm{T}}Kf$; see Green & Silverman (1994, p. 13) for details.

We define the eigenvalue decomposition of $K = UDU^{\mathrm{T}}$ with $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_n$, and partition it as follows:

$$[U_1 \; U_2] \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{bmatrix} U_1^{\mathrm{T}} \\ U_2^{\mathrm{T}} \end{bmatrix}, \tag{A·2}$$

with $D_1 = \text{diag}(0, 0)$ and $D_2 = \text{diag}(d_3, \ldots, d_n)$, where d_i, for $i = 3, \ldots, n$, are the nonzero eigenvalues of K. The columns of the matrix U_1 span the linear space and are orthogonal to the columns of U_2. This implies that U_2 will be orthogonal to any linear function of x; in particular, we will have $U_2^{\mathrm{T}}X = 0$.

Using decomposition (A·2), we have that

$$S_\lambda = (I + \lambda K)^{-1} = U_1U_1^{\mathrm{T}} + U_2(I_{n-2} + \lambda D_2)^{-1}U_2^{\mathrm{T}}.$$

The matrix $Z = Z(x)$ in model (2·1) is defined by the relationship $U_2 D_2 U_2^{\mathrm{T}} = (ZZ^{\mathrm{T}})^{-}$; see also Wahba (1990, pp. 16–20) and Speed (1991). This implies that $Z = U_2 D_2^{-1/2}$, and that $Z^{\mathrm{T}} X = 0$.

Moreover, the best linear unbiased predictor fitted values obtained with (2·2) are

$$\begin{aligned}\hat{y} &= X\hat{\beta} + Z\hat{u} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y + Z(\lambda I_{n-2} + Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y\\ &= \{U_1 U_1^{\mathrm{T}} + U_2 D_2^{-1/2}(D_2^{-1} + \lambda I_{n-2})^{-1} D_2^{-1/2} U_2^{\mathrm{T}}\}y = \{U_1 U_1^{\mathrm{T}} + U_2(I_{n-2} + \lambda D_2)^{-1} U_2^{\mathrm{T}}\}y,\end{aligned}$$

which shows the equivalence with the decomposition of S_λ.

Distribution of T. We have that

$$\begin{aligned}T &= y^{\mathrm{T}}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})^{-1} - (I_n + \lambda_1^{-1}ZZ^{\mathrm{T}})^{-1}\}y\\ &= (y - X\beta)^{\mathrm{T}} Z\{(\lambda_1 I_{n-2} + Z^{\mathrm{T}}Z)^{-1} - (\lambda_0 I_{n-2} + Z^{\mathrm{T}}Z)^{-1}\}Z^{\mathrm{T}}(y - X\beta)\\ &= y^{*\mathrm{T}}\{D_2^{-1}(\lambda_1 I_{n-2} + D_2^{-1})^{-1} - D_2^{-1}(\lambda_0 I_{n-2} + D_2^{-1})^{-1}\}y^*\\ &= \sigma^2 z^{\mathrm{T}}(I_{n-2} + \lambda_0^{-1} D_2^{-1})\{D_2^{-1}(\lambda_1 I_{n-2} + D_2^{-1})^{-1} - D_2^{-1}(\lambda_0 I_{n-2} + D_2^{-1})^{-1}\}z\\ &= \sigma^2 \sum_{i=3}^{n} c_i z_i^2,\end{aligned}$$

where

$$y^* = U_2^{\mathrm{T}}(y - X\beta) \sim \mathcal{N}\{0, \sigma^2(I_{n-2} + \lambda_0^{-1} D_2^{-1})\},$$

and $z = \sigma^{-1}(I + \lambda_0^{-1} D_2^{-1})^{-1/2} y^*$ follows a standard normal distribution.

F-approximation. Consider first $\sum c_i = \mathrm{tr}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})(S_{\lambda_1} - S_{\lambda_0})\}$. We have

$$\begin{aligned}\sum c_i &= \mathrm{tr}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})(S_{\lambda_1} - S_{\lambda_0})\}\\ &= \mathrm{tr}[U_2(I_{n-2} + \lambda_0^{-1} D_2^{-1})\{(I_{n-2} + \lambda_1 D_2)^{-1} - (I_{n-2} + \lambda_0 D_2)^{-1}\}U_2^{\mathrm{T}}]\\ &= \frac{\lambda_0 - \lambda_1}{\lambda_0}\,\mathrm{tr}\{U_2(I_{n-2} + \lambda_1 D_2)^{-1} U_2^{\mathrm{T}}\} = \frac{\lambda_0 - \lambda_1}{\lambda_0}\{\mathrm{tr}(S_{\lambda_1}) - 2\},\end{aligned} \tag{A·3}$$

where we used the fact that $U_1^{\mathrm{T}} U_2 = 0$ and $\mathrm{tr}(U_1 U_1^{\mathrm{T}}) = 2$. Similarly,

$$\mathrm{tr}(I_n - S_{\tilde{\lambda}}) = n - \mathrm{tr}(S_{\tilde{\lambda}}), \quad \mathrm{tr}\{\lambda_0^{-1} ZZ^{\mathrm{T}}(I_n - S_{\tilde{\lambda}})\} = \frac{\tilde{\lambda}}{\lambda_0}\{\mathrm{tr}(S_{\tilde{\lambda}}) - 2\}$$

give

$$\sum b_i = \mathrm{tr}\{(I_n + \lambda_0^{-1}ZZ^{\mathrm{T}})(I_n - S_{\tilde{\lambda}})\} = n + \left(\frac{\tilde{\lambda}}{\lambda_0} - 1\right)\mathrm{tr}(S_{\tilde{\lambda}}) - 2\,\frac{\tilde{\lambda}}{\lambda_0}. \tag{A·4}$$

References

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

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A. & Van der Vorst, H. (2000). *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide.* Philadelphia: SIAM.

Chambers, J. M. & Hastie, T. J. (Ed.) (1991). *Statistical Models in S.* Belmont, CA: Wadsworth.

Chatterjee, S., Handcock, M. S. & Simonoff, J. S. (1995). *A Casebook for a First Course in Statistics and Data Analysis.* New York: Wiley.

Davies, R. B. (1980). [Algorithm AS 155] The distribution of a linear combination of χ^2 random variables. *Appl. Statist.* **29**, 323–33.

Eilers, P. H. C. & Marx, B. D. (1992). Generalized linear models with P-splines. In *Advances in GLIM and Statistical Modelling. Proceedings of the GLIM92 Conference*, Ed. L. Fahrmeir et al., pp. 72–7. New York: Springer.

Farebrother, R. W. (1990). [Algorithm AS 256] The distribution of a quadratic form in normal variables. *Appl. Statist.* **39**, 294–309.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50**, 640–52.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Hastie, T. (1996). Pseudosplines. *J. R. Statist. Soc.* B **58**, 379–96.

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

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* **21**, 309–10.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay. L. Lush*, pp. 10–41. Champaign, IL: Am. Soc. Anim. Sci.–Am. Dairy Sci. Assoc.–Poultry Sci. Assoc.

Johnson, N. L. & Kotz, S. (1970). *Continuous Univariate Distributions*, **2**. Boston: Houghton-Mifflin.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models. *J. R. Statist. Soc.* B **61**, 381–400.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with Discussion). *Statist. Sci.* **6**, 15–51.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Sockett, E. B., Daneman, D., Clarson, C. & Ehrich, R. M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes millitus in children. *Diabetologia* **30**, 453–9.

Speed, T. (1991). Comment on 'That BLUP is a good thing: The estimation of random effects'. *Statist. Sci.* **6**, 42–4.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wecker, W. E. & Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Assoc.* **78**, 81–9.

[*Received April* 2001. *Revised August* 2001]